Analysis

Edwin Park

2023

Contents

1 Real Numbers

> "The great discovery of the late nineteenth century was that numbers can be understood abstractly via axioms, without necessarily needing a concrete model; of course a mathematician can use any of these models when it is convenient, to aid his or her intuition and understanding, but they can also be just as easily discarded when they begin to get in the way."
>
> Tao [**?**, 19]

1.1 Axiomatisation of the Reals

The existence of a model (a structure satisfying the axioms) will be given in the next section. The standard axiomatisation of $\mathbb{R}$ is as follows:

1. $\mathbb{R}$ is a field with respect to $+$ and $*$.
2. $\mathbb{R}$ is totally ordered with respect to $\leq$. [1]
3. The two operations preserve order; $0 \leq x \wedge 0 \leq y \Rightarrow 0 \leq x * y$ and $x \leq y \Rightarrow x + z \leq y + z$.
4. The relation $\leq$ is complete; any non-empty subset of $\mathbb{R}$ bounded above has a least upper bound.

We note that Axiom 4 is nonfirstorderisable. There is an alternative, more concise axiomatisation due to Tarski.

[1] For $\leq$ to be totally ordered means that:
(a) $\forall x,\ x \leq x$ (Reflexivity)
(b) $\forall x, y,\ x \leq y \wedge y \leq x \Rightarrow x = y$ (Antisymmetry)
(c) $\forall x, y, z,\ x \leq y \wedge y \leq z \Rightarrow x \leq$ (Transitivity)
(d) $\forall x, y,\ x \leq y \vee y \leq x$ (Totality)

1.2 Cauchy Sequences

1.2.1 Definitions

While $\mathbb{Q}$ is dense (by its property that $\forall x, y\ \exists z\ x < y \Rightarrow x < z < y$), it is not "complete" in the sense that certain definable concepts (e.g. $\sqrt{2}$) are not in $\mathbb{Q}$. From the following definitions, equivalence classes of Cauchy sequences will form a model for the axioms from 1.1.

Definition 1.2.1.1. A *sequence* $(a_q)_{q=m}^{\infty}$ is a mapping from $\{n \in \mathbb{Z} \mid n \geq m\}$ to $\mathbb{Q}$. (Should the mapping be computable, definable, or not necessarily either?) (What is the definition of mapping?)

Definition 1.2.1.2. A *Cauchy sequence* is a sequence satisfying the property that for any rational number $\varepsilon > 0$, there is a finite $N \in \mathbb{N}$ such that $\forall i, j > N,\ d(a_i, a_j) < \varepsilon$. (Is it always possible to determine whether a (computable) sequence is Cauchy?)

Definition 1.2.1.3. Two Cauchy sequences (a_i) and (b_i) are equivalent (written $(a_i) \sim (b_i)$) if for any rational number $\varepsilon > 0$, there is a finite $N \in \mathbb{N}$ such that $\forall i > N,\ d(a_i, b_i) < \varepsilon$. (Is equivalence computable?)

Remark 1.2.1.1. The notion of equivalence defined above forms an equivalence relation.

Proof. Reflexivity and symmetrty are obvious. Given $(a_i) \sim (b_i)$ and $(b_i) \sim (c_i)$, we can use N_1 and N_2 to write:

$$\forall i > \max(N_1, N_2),\ d(a_i, c_i) \leq d(a_i, b_i) + d(b_i, c_i) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Note that we have used the triangle inequality axiom for metric spaces. □

1.2.2 Arithmetic Operations on $\mathbb{R}$

We now define the arithmetic operations on the equivalence classes.

Definition 1.2.2.1. $[(a_n)] + [(b_n)] := [(a_n + b_n)]$, where $(a_n + b_n)$ is the sequence where the n^{th} term is $a_n + b_n$.

To make sure this definition of addition is well-defined, we need to make sure that the addition of two Cauchy equivalence classes is a unique Cauchy equivalence class.

> "You see, there's a catch when you define things using equivalence relations. If you ever wish to define some operation on equivalence classes, you can't just willy-nilly define it for one member of the class and expect it to make sense for all members of the class."
>
> Kemp [?, 6]

Firstly, we show that the addition of two Cauchy equivalence classes is a Cauchy equivalence class. Noting that:

$$\begin{aligned} d(a_i + b_i, a_j + b_j) &= |(a_i + b_i) - (a_j + b_j)| \\ &= |(a_i - a_j) + (b_i + b_j)| \\ &\leq |(a_i - a_j)| + |(b_i + b_j)| \\ &= d(a_i, a_j) + d(b_i + b_j), \end{aligned}$$

(is this true for general metrics?) we can write:

$$\forall i > \max(N_1, N_2),\ d(a_i + b_i, a_j + b_j) \leq d(a_i, a_j) + d(b_i + b_j) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Also, additions of equivalent sequences are equivalent; given $(a_n) = (a'_n)$, we will show that $(a_n + b_n) = (a'_n + b_n)$.

$$\begin{aligned} \forall i > N,\ d(a_i + b_i, a'_i + b_i) &= |(a_i + b_i) - (a'_i + b_i)| \\ &= |a_i - a'_i| < \varepsilon. \end{aligned}$$

Thus, $(a_n + b_n) \sim (a'_n + b_n) \sim (a'_n + b'_n)$. □
Multiplication is defined similarly, but with caution with 0 and reciprocation.

1.2.3 $\leq$ on $\mathbb{R}$

Definition 1.2.3.1. $x \leq y$ iff $y - x$ is 0 or positive.

1.2.4 Least Upper-bound Axiom

From the structure we have built, it is not too difficult to show that it models the first three axioms from 1.1. We now show that Axiom 4 is satisfied by our construction.

Definition 1.2.4.1.

Theorem 1.2.4.1. Any non-empty set $S \subseteq \mathbb{R}$ with an upper bound (a real number M satisfying $\forall x \in S,\ x \leq M$) has exactly one minimal upper bound u such that $u \leq M$ for all upper bounds M of S.

Proof. Let $u_0 = M$ for some upper bound M of S and $\ell_0 = x$ for some $x \in S$. If $a_n = \frac{\ell_n + s_n}{2}$ is an upper bound for S, assign $\ell_{n+1} = \ell_n$ and $u_{n+1} = a_n$. Otherwise, assign $\ell_{n+1} = a_n$ and $u_{n+1} = u_n$. This assignment preserves the upper-bound property of u_n, and so by induction, $\forall n \in \mathbb{N},\ u_n$ is an upper bound for S. Similarly, $\forall n \in \mathbb{N},\ \exists x \in S,\ \ell_n \leq x$.

We want to show that (u_n) and (l_n) are Cauchy sequences, but a caveat is that they are sequences of real numbers, while our definition only allowed sequences of rationals. We may easily extend our definition of a sequence in 1.2.1.1 as a mapping to $\mathbb{R}$ instead of $\mathbb{Q}$. This is not circular as we have already built $\mathbb{R}$ independently from $\mathbb{Q}$ (i.e. it is not circular as we are not re-defining the reals, but effectively defining a new construct which happens to have the same name "sequence" as our previous construct). A crucial nontrivial property is that equivalence classes of Cauchy sequences of reals are isomorphic to the reals. That is to say, the equivalence class of a real Cauchy sequence is equivalent to the equivalence class of some rational Cauchy sequence. (Equivalently, Cauchy sequences of reals converge to a real number.)

The proof is as follows. First we note that rationals can get arbitrarily close to any real number. Now, given a real sequence (r_n), we construct a rational sequence (q_n) such that $|u_n - q_n| < \frac{1}{n}$. Then, by construction, $[(r_n)] = [(q_n)]$. All we need to show is that (q_n) is Cauchy. To do so, we note, given $n, m > \max(N_1, N_2)$ for appropriate N_1, N_2:

$$\begin{aligned} |q_n - q_m| &= |(q_n - u_n) + (u_n - u_m) + (u_m - q_m)| \\ &\leq |(q_n - u_n)| + |(u_n - u_m)| + |(u_m - q_m)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

For N_1 we need a natural number satisfying $\frac{1}{N_1} < \frac{\varepsilon}{3}$, which is guaranteed by the Archimedean property (proof omitted).

$[(u_n)]$ is an upper bound for S. If not, then for some $s \in S$, $[(u_n)] < s \Rightarrow [(u_n)] + \varepsilon < s$ for some small enough ε. But as (u_n) is non-increasing, we can find a n such that $u_n < [(u_n)] + \varepsilon$. [2] But then, $u_n < [(u_n)] + \varepsilon < s$, which is a contradiction as u_n is always an upper bound for S.

Using a similar argument, no real number smaller than $[(\ell_n)]$ is an upper bound for S. Suppose otherwise. Let s be an upper bound of S such that $s < [(\ell_n)] \Rightarrow s < [(\ell_n)] - \varepsilon$ for some small enough ε. But as (ℓ_n) is non-decreasing, we can find a n such that $\ell_n > [(\ell_n)] - \varepsilon$. But then,

$$\ell_n > [(\ell_n)] - \varepsilon > s \Rightarrow \ell_n \text{ is greater than all elements in } S,$$

which is a contradiction by construction of ℓ_n. □

[2] Choose n such that $d(u_n, [(u_n)]) < \varepsilon$. By the non-increasing property we have $u_n \geq [(u_n)] \Rightarrow |u_n - [(u_n)]| = (u_n - [(u_n)])$. Then,

$$\begin{aligned} u_n - d(u_n, [(u_n)]) &= u_n - (u_n - [(u_n)]) \\ &= [(u_n)] \\ \Rightarrow u_n - d(u_n, [(u_n)]) + \varepsilon &= [(u_n)] + \varepsilon \\ \Rightarrow u_n &< [(u_n)] + \varepsilon \end{aligned}$$

Remark 1.2.4.1. A potential point of confusion is that the property that u_n is an upper bound extends to $\lim\limits_{n\to\infty} u_n$[3], while the property $\forall n \in \mathbb{N},\ \exists x \in S,\ \ell_n \leq x$ of ℓ_n does not extend to $\lim\limits_{n\to\infty} \ell_n$. As a counterexample consider the set $[0,1)$.

[3] Where (a_n) is a Cauchy sequence, $\lim\limits_{n\to\infty} a_n$ is defined to be the real number corresponding to $[(a_n)]$.

2 Metric Spaces

2.1 Definitions

Definition 2.1.0.1. A metric space is a tuple (X, d) of a set X and a function $d : X \times X \to [0, \infty)$ satisfying:

1. $d(x, x) = 0$.
2. $x \neq y \Rightarrow d(x, y) > 0$.
3. $d(x, y) = d(y, x)$.
4. $d(x, z) \leq d(x, y) + d(y, z)$.

2.1.1 Open and Closed Sets

Definition 2.1.1.1. The open ball in a metric space (X, d) is defined as $B(x_0, r) := \{x \in X \mid d(x, x_0) < r\}$. Munkres writes this as the r-neighbourhood at x_0, or $U(x_0, r)$.

Definition 2.1.1.2. An interior point of a set E is a point such that there is an open ball (of positive radius) centred at it which is completely contained in E. If there is such a ball completely disjoint from E then it is an exterior point, and a boundary point if neither. The set of boundary points is denoted by ∂E.

Definition 2.1.1.3. A set is closed if it contains all its boundary points, open if it contains none, neither if neither and clopen if it has no boundary.

Remark 2.1.1.1. Open balls are open. Consider $B = U(x, r)$ and $x' \in B$. Let $r' = r - d(x', x)$, and $B' = U(x', r')$. Then, for any $y \in B$, $d(y, x) \leq d(y, x') + d(x', x) < r' + d(x', x) = r \Rightarrow y \in B$.

Definition 2.1.1.4. A cluster point, or a point of accumulation, of a non-empty set S is a point such that every punctured ball centred on it contains points in S.

Remark 2.1.1.2. In order to deal with annoying exam questions whose proofs may depends definitions, we demonstrate here the equivalence of different definitions of closed and open sets.

1. A set is open iff it contains none of its boundary points, and closed if it contains all.
2. A set if open iff for every point in it, there is some open ball centred there contained in the set.
3. A set is closed/open iff its complement is open/closed.
4. A set is closed iff it contains all its cluster points.

Proof. From its definition, we can see that a set can never contain its exterior points. So, 1 implies that a set is open iff it only contains interior points. But then, 1 $\Rightarrow$ 2 follows from the definition of interior points. Also, 2 $\Rightarrow$ 1 as 2 implies every point in an open set is an interior point.

To show 3, we note that a set and its complement share the same boundary; $\partial S = \partial S'$. This follows from the definition; the boundary points of S' are the points such that every open ball centred on it contains both points in S and points not in S; this is precisely the boundary for S. Thus, a set contains all its boundary points iff its complement contains none, and so 1 $\iff$ 3.

Finally, cluster points by definition cannot be exterior. So if a set is closed, it must contain all cluster points (as closed sets contain all non-exterior points). Conversely, we note that every boundary point is a cluster point, as per their definitions. Thus, a set containing all its cluster points contains all its boundary points, and we are done. □

Proposition 2.1.1.1. Closed sets are closed under (finite applications of) the set union operation.

Proof. Suppose otherwise; that for some closed sets A and B, $A \cup B$ is not closed. Then there is some boundary point x of $A \cup B$, such that $x \notin A \cup B$. However, x cannot be an exterior point for both A and B, as every open ball centred at it contains points either always in A or always in B. WLOG, x is a boundary point for A, which contradicts the assumption that A is closed. □

Proposition 2.1.1.2. Closed sets are closed under (possibly infinite applications of) the set union operation.

Proof. Suppose otherwise; that for some set of closed sets A, $\bigcap A$ is not closed. Then there is some boundary point x of $\bigcap A$, such that $x \notin B$, $\forall B \in \bigcap A$. However, x cannot be an exterior point for any $B \in \bigcap A$, as every open ball centred at it contains points common to B, $\forall B \in \bigcap A$. Thus, $x \in B$, $\forall B \in \bigcap A$ (as closed sets contain all non-boundary points), a contradiction. □

Remark 2.1.1.3. For analogous results for open sets, use De Morgan's laws; $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$.

Definition 2.1.1.5. A set is bounded if there is a ball containing it.

Theorem 2.1.1.1 (Bolzano-Weierstrass)**.** Every bounded set $S \in \mathbb{R}^n$ with infinitely many elements has a cluster point.

Proof. Take a hypercube bounding S of length L, and slice it into 2^n smaller hypercubes, each with side length $\frac{L}{2}$. The union of these pieces contains infinitely many points of S, so at least one of these slices must contain infinitely many points in S. Choose one and repeat; the sequence of centre-points for the chosen hypercubes is a Cauchy sequence for a cluster point of S. To see why, create a punctured ball on the point represented by the Cauchy sequence. Now, we note that the hypercubes corresponding to the entries in the Cauchy sequence get arbitrarily small, and their centres arbitrarily close to the limit, and so a certain hypercube in the sequence is contained within the punctured ball. Also, all hypercubes in the sequence contain infinitely many points in S, and we are done. □

Remark 2.1.1.4. The theorem does not hold for general metric spaces.

Corollary 2.1.1.1 (Bolzano-Weierstrass for subsequences)**.** Every bounded sequence (in $\mathbb{R}^n$) has a convergent subsequence.

Proof. Take the Cauchy sequence from before, and replace each entry with any point in the corresponding hypercube, whose index in the sequence is greater than the last entry's. We now have some subsequence of the original, which is Cauchy. □

2.1.2 Continuity

Definition 2.1.2.1. Where (X, d_x) and (Y, d_y) are metric spaces, a function $f : X \to Y$ is continuous at $x \in X$ iff $f(x)$ is equal to the mapping of f to any Cauchy sequence of x (under the metric d_y) i.e. $f(x) = (f(x_0), f(x_1), ...)$.

Theorem 2.1.2.1. This is equivalent to the usual epsilon-delta definition: f is continuous at x if

$$\forall \varepsilon > 0 \, \exists \, \delta > 0, \; d_x(x_i, x) < \delta \Rightarrow d_y(f(x_i), f(x)) < \varepsilon,$$

or

$$\forall \varepsilon > 0 \, \exists \, \delta > 0, \; f(U_X(x, \delta)) \subseteq U_Y(f(x), \varepsilon).$$

Proof. As x_i gets arbitrarily close to x (under d_x), $f(x_i)$ gets arbitrarily close to to $f(x)$ (under d_y). For a given $\varepsilon > 0$, find the index i s.t $\forall i, \; d_y(f(x_i), f(x)) < \varepsilon$ (which we are guaranteed as the sequence is Cauchy). Then, choose $d_x(x_i, x)$ as our δ – then, the index j such that $\forall j, \; d_x(x_j, x) < \delta$ satisfies $j > i$. The case of $\delta = 0$ can be avoided by choosing a Cauchy sequence of x whose entries are never equal to x. If no such Cauchy sequence can be found, choose a δ such that the antecedent is always false; then f is vacuously continuous at x. □

Remark 2.1.2.1. Continuitiy is preserved under composition, and continuity preserves connectedness (proof omitted).

Theorem 2.1.2.2. f is continuous $\iff$ f^- preserves openness/closedness of subsets of the codomain.

Proof.
$L \Rightarrow R:$
Preservation of openess: Let V be an open subset of the codomain. Choose x in $f^-(V)$. By openness, $\exists \, \varepsilon, \; U_Y(f(x), \varepsilon) \subseteq V$. But by continuity, this implies that $\exists \, \delta, \; f(U_X(x, \delta)) \subseteq U_Y(f(x), \varepsilon)$. But then, by definition, $U_X(x, \delta) \subseteq f^-(U_Y(f(x), \varepsilon)) \subseteq f^-(V)$.

Preservation of closedness: Let V be a closed subset of the codomain. Consider any Cauchy (x_i) with entries in $f^-(V)$ which converges to x. Then, by closedness and continuity,
$(f(x_i)) \in V \Rightarrow f(x) \in V \Rightarrow x \in f^-(V)$.

$R \Rightarrow L:$
Preservation of openess: Where $f : X \to Y$, choose $\varepsilon > 0$ and consider $f(x)$ for any $x \in X$. Then, $U_Y(f(x), \varepsilon)$ is open in Y (neighbourhoods are always open). Thus, $f^-(U_Y(f(x), \varepsilon))$ is open, and so we can find a neighbourhood centred at x contained in it:
$\exists \, \delta, \; U_X(x, \delta) \subseteq f^-(U_Y(f(x), \varepsilon)) \Rightarrow f(U_X(x, \delta)) \subseteq U_Y(f(x), \varepsilon)$ and we are done.

Unlike a direct method like before, this time we will use the complementary relationship between open and closed sets. Consider an open set U in Y. Then $Y \setminus U$ is closed in Y (because $U \setminus Y$ is open). Then, by assumption, $f^-(Y \setminus U)$ is closed, but $f^-(Y \setminus U) = X \setminus f^-(U)$ (from some set algebra) so $f^-(U)$ is open, and we are done. □

Remark 2.1.2.2. Maps preserving openness are called open maps. f is an open map $\not\Leftrightarrow$ f is continuous. The respective counterexamples are:

$$L \not\Rightarrow R: \quad f : \mathbb{R} \to \mathbb{Z}, \; f(x) = \lfloor x \rfloor$$

$$R \not\Rightarrow L: \quad f : \mathbb{R} \to \mathbb{R}, \; f(x) = 0.$$

2.1.3 Connectedness

Before defining connectedness, it will be convenient to discuss openness and closedness of subspaces (subspace implying that the measure of the metric space is kept) of metric spaces. In particular, we note how $[0, 1)$ is neither closed or open in $(\mathbb{R}, d)$ (where d is Euclidean) but open in the subspace $X = ([0, 2], d)$; $[0, 1) = U_X(0, 1)$. We may characterise open/closed sets in subspaces as follows:

Theorem 2.1.3.1. Consider a subspace X of the metric space (M, d). $A \subseteq X$ is open/closed in X iff there is an open/closed $\mathcal{O} \in M$ such that $A = X \cap \mathcal{O}$.

Proof. We consider the open case first.
$L \Rightarrow R$: Construct an open set $\mathcal{O}$ in M as:

$$\mathcal{O} = \bigcup_{a \in A} U_M(a, \varepsilon_a)$$

where ε_a is chosen so that $U_X(a, \varepsilon_a) \in A$. Then,

$$\begin{aligned} X \cap \mathcal{O} &= X \cap \bigcup_{a \in A} U_M(a, \varepsilon_a) \\ &= \bigcup_{a \in A} (X \cap U_M(a, \varepsilon_a)) \\ &= \bigcup_{a \in A} (U_X(a, \varepsilon_a)) && (1) \\ &= A && (2) \end{aligned}$$

$R \Rightarrow L$: We want to show that, given a set $\mathcal{O}$ open in M, $A = X \cap \mathcal{O}$ is open in X. Choose any $a \in A$. We know that $a \in \mathcal{O}$, so by openness of $\mathcal{O}$ we can choose a ε such that $U_M(a, \varepsilon) \subseteq \mathcal{O}$. But this gives $U_X(a, \varepsilon) = X \cap U_M(a, \varepsilon) \subseteq X \cap \mathcal{O} = A$.

The case for closed sets follow easily. Let B be a subset of X.

$$\begin{aligned} B \text{ is closed in } X &\iff X \setminus B \text{ is open in } X \\ &\iff X \setminus B = X \cap \mathcal{O}, \text{ for some open } \mathcal{O} \subseteq M \\ &\iff X \cap (X \setminus B)' = X \cap (X \cap \mathcal{O})' \\ &\iff X \cap (X \cap B')' = X \cap (X' \cup \mathcal{O}') \\ &\iff X \cap (X' \cup B) = X \cap \mathcal{O}' \\ &\iff X \cup B = X \cap \mathcal{C} \text{ for some closed } \mathcal{C} \subseteq M \\ &\iff B = X \cap \mathcal{C}. \end{aligned}$$

□

Definition 2.1.3.1. A set S is disconnected iff there exists two non-empty disjoint sets covering it, but not individually. This is equivalent to saying that S has a non-trivial proper clopen set in S.

Proof.
$L \Rightarrow R$: Suppose we are working in the metric space (X, d), and that the open sets A and B demonstrate the disconnectedness of $S \subseteq X$, i.e. $S \subseteq A \cup B$, $A \cap B = \varnothing$, $A \cap S \neq \varnothing \neq B \cap S$. Now, $A \cap S$ is open in S as A is open in X. It follows that $S \setminus (A \cap S) = B \cap S$ (1) is closed in S, but by symmetry of A and B, $B \cap S$ is also open and $A \cap S$ is also closed.

[1] An open ball in X is equivalent to the same open ball in M, but consisting only of the points belonging to X; $U_X(x, r) = X \cap U_M(x, r)$.

[2] Let $B = \bigcup_{a \in A}(U_X(a, \varepsilon_a))$. $A \subseteq B$ as B runs over all points in A. $B \subseteq A$ as all neighbourhoods in the union are contained in A.

$R \Rightarrow L$: Suppose that a set S in a metric space (X, d) has no non-trivial proper subsets which are clopen in S. Assume for the sake of contradiction that S is disconnected. Then by the $L \Rightarrow R$ proof above we can construct non-trivial proper clopen subsets in S, a contradiction. □

Theorem 2.1.3.2. Continuity preserves connectedness.

Proof. Where $f : X \to Y$ is continuous and a non-empty $S \subseteq X$ is connected, let $\mathcal{O}_1$ and $\mathcal{O}_2$ satisfy the conditions to show that a $f(S)$ is disconnected, for the sake of contradiction. We will show that $f^{-}(\mathcal{O}_1)$ and $f^{-}(\mathcal{O}_2)$ satisfy the conditions to show that S is disconnected. Both contain points in S as $\mathcal{O}_1$ and $\mathcal{O}_2$ contain points in $f(S)$. This also shows that they are non-empty. They are disjoint as otherwise, if x was common to both, $f(x) \in \mathcal{O}_1 \cap \mathcal{O}_2$, a contradiction. Finally, they collectively cover S as $f(S) \subseteq \mathcal{O}_1 \cup \mathcal{O}_2 \Rightarrow S \subseteq f^{-}(\mathcal{O}_1 \cup \mathcal{O}_2) = f^{-}(\mathcal{O}_1) \cup f^{-}(\mathcal{O}_2)$. □

Definition 2.1.3.2. An *interval* is a set in the extended reals $\mathbb{R}^*$ denoted by $[a, b] := \{x \mid a \leq x \leq b\}$, with either "$\leq$" replaceable by a "$<$", along with a change from a square bracket to a paranthesis.

Theorem 2.1.3.3. X is an interval $\iff a, b \in X \wedge a < b \Rightarrow \forall x \in (a, b),\ x \in X$

Proof.
$L \Rightarrow R$: Follows trivially by definition of an interval.

$R \Rightarrow L$: Let $u = \sup(X)$ and $l = \inf(X)$. We want to show that X contains all elements "in between" u and l. For any x such that $l < x < u$, by definition of supremum and infimum, we can find $a, b \in X$ such that $a < x < b$, and thus $x \in X$. □

Theorem 2.1.3.4. Where $S \subseteq \mathbb{R}$, S is connected $\iff$ S is an interval.

Proof.
$L \Rightarrow R$: Consider any two points a, b in S and suppose that $x \in (a, b)$, $x \notin S$. But then $(-\infty, x)$ and (x, ∞) demonstrate the disconnectedness of S, contradicting the hypothesis.

$R \Rightarrow L$: We first show the results for closed, bounded intervals. For the sake of contradiction, let $U, V \subseteq [a, b]$ be clopen subsets (in $[a, b]$) demonstrating disconnectedness of the closed interval. Assume wlog that $b \in V$ and let u be the supremum of U. $u \in U$ by closedness of U. Noting that $(b - \varepsilon, b] \in V$ by openness, it follows that $u < b$ by disjointness, but then, $[u, u + \epsilon) \in U$ as U is open in $[a, b]$, contradicting that u is a supremum.

For general intervals, the results follows from the fact that connectedness is preserved under certain unions, and all intervals can be generated from a union of closed bounded intervals. See "4-connected.pdf". □

Corollary 2.1.3.1 (Intermediate Value Theorem)**.** Where $f : S \to \mathbb{R}$ is continuous, S is connected, and $a < b$, $a, b \in f(S) \Rightarrow [a, b] \subseteq f(S)$.

Remark 2.1.3.1. Connected open sets are polygonally connected.

Proof. Let x be a point in the open connected set S, and X be the smallest subset of S containing x closed under polygonal connection; that is, y is polygonally connected to $x \iff y \in X$. We note that X must be open, as if it were to have a boundary point b, by openness of S there is a neighbourhood of b in S, then in turn, this neighbourhood would be contained in X (as every point in a neighbourhood is polygonally connected to its centre), contradicting the assumption that b is a boundary of X. Similarly, suppose that $S \setminus X = S \cap X'$ has a boundary point b. This means that every open ball centred at b contains points in $(S \setminus X)' = S' \cup X$. In particular, every neighbourhood of b must contain points in X, as otherwise this contradicts S being open. But then, as this open ball contains a point in X, its centre b must also be polygonally connected to x, a contradiction, as $S \setminus X$ is supposed to be points in S unreachable from x. Thus, $S \setminus X$ is open.

Now, notice that $S \setminus X$ and S are open, disjoint, and collectively cover S. As S is connected, at least one of them cannot contain points in S – this must be $S \setminus X$ as X contains x; thus, $S \setminus X = \varnothing$. □

2.1.4 Compactness

Definition 2.1.4.1. A compact set is a set which is closed and bounded. This definition is equivalent to saying that all sequences in a compact set S have a subsequence converging to a point in S.

Proof. A closed set contains all cluster points, so any sequence in it must converge to it. Conversely, the set must be closed as it contains all its cluster points. Also, it must be bounded as otherwise, one can construct an unbounded sequence, which has no convergent subsequences. □

Remark 2.1.4.1. For general topological spaces this definition is somewhat problematic, so a definition using open covers may instead be used.

Theorem 2.1.4.1 (Heine-Borel)**.**
$K \subseteq \mathbb{R}$ is compact $\iff$ (C covers $K \Rightarrow$ a finite subset of C covers K). A set C covers B iff $B \subseteq \bigcup C$ and $\forall A \in C$, A is open.

Proof.
$L \Rightarrow R$: Given $K \subseteq \mathbb{R}$ is compact, assume bwoc that coverings of K have no finite subcover. Using the hypercube argument from before, at least one of the smaller hypercubes must require an infinite number of elements from a given covering C. Let L be the limit point of the Cauchy sequence of the centres of the hypercubes. C must have some open set B containing L, but this contradicts the fact that each of our hypercubes required an infinite amount of elements of C to be covered. (We can find an arbitrarily small hypercube in our sequence that fits inside B.)

$R \Rightarrow L$: Given the RHS, assume bwoc that K is not closed i.e. there exists some point $c \notin K$ whose Cauchy sequence is in K. The set of neighbourhoods of all $x \in K$ with their radii chosen not to include some neighbourhood of c (say, with half the distance from x to c) covers K, but has no finite subcovering – given a finite subcovering, each of its elements is disjoint from a certain neighbourhood of c, although that neighbourhood certainly contains elements in K. Thus, K must be closed.

Also, given the RHS, consider the covering $C = \{B(x, 1) \mid x \in K\}$. A finite subcovering of this consists of a finite number of balls of radius 1, so K must be bounded. □

Lemma 2.1.4.1. Continuous maps preserve compactness.

Proof. Let f be a continuous map and X a closed set.

$$\begin{aligned}(f(x_0), f(x_1), \dots) &= f((x_0, x_1, \dots)) && \text{(by continuity of } f\text{)}\\ &= f(x),\ x \in X && \text{(by closedness of } X\text{)}\end{aligned}$$

Thus, any Cauchy sequence in $f(X)$ is contained in $f(X)$ – closedness is preserved by continuous maps.

Now, assume X is bounded (but not necessarily closed). For $f(X)$ to be unbounded means that there exists a sequence in $f(X)$ which has no convergent subsequence. For such a sequence $(f(x_0), f(x_1), \dots)$, consider the corresponding sequence in X, $(x_0, x_1, \dots)$. While this sequence may not be Cauchy it is bounded, so there is some convergent subsequence $(x_{\phi(n)})$. But then, $(f(x_{\phi(0)}), f(x_{\phi(1)}), \dots) = f((x_{\phi(n)}))$ by continuity of f, so $(f(x_i))$ has a subsequence that converges – a contradiction. □

Theorem 2.1.4.2. Compactness ensures attainment of maxima. (f attains a maximum on $c \in \mathrm{dom} f$ if $f(c) = \sup_{\mathrm{dom} f} f$.)

Proof. Let $X = \mathrm{dom}_f$. $\sup_X f = (f(x_0), f(x_1), \dots)$ for some sequence $(x_0, x_1, \dots)$ in X. By boundedness, some subsequence $(x_{\phi(n)})$ converges, and is guaranteed to be in X by closedness. In particular, we have $f((x_{\phi(n)})) = (f(x_{\phi(0)}), f(x_{\phi(1)}), \dots) = \sup_X f$ and we are done. □

Definition 2.1.4.2. A function from metric space (X, d_x) to (Y, d_y) is uniformly continuous if for every $\varepsilon > 0$ there is a global parameter δ such that $d_x(x_0, x_1) < \delta \Rightarrow d_y(f(x_0) - f(x_1)) < \varepsilon$.

Theorem 2.1.4.3 (Heine-Cantor)**.** Continuity implies uniformness on compact sets.

Proof. For a continuous $f : X \to Y$ where X is compact, fix $\varepsilon > 0$ and assign each $x \in X$ with a δ_x such that $d_x(x, a) < \delta_x \Rightarrow d_y(f(x), f(a)) < \frac{\varepsilon}{2}$, which we are guaranteed by continuity. Now, we notice that $d_x(a, x), d_x(x, b) < \min(\delta_a, \delta_b) \Rightarrow d_y(f(a), f(b)) \leq d_y(f(a), f(x)) + d_y(f(x), f(b)) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}$. So setting $\delta = \min_{x \in X}(\delta_x)$ gives uniform continuity, but the problem is that as there are potentially an infinite number of δ_x to consider, which may set δ to 0 (this is not allowed by definition, where $\delta > 0$.)

The trick here is to use the Heine-Borel theorem to reduce the number of δ_x needed to a finite amount. The set

$$C = \left\{ B\left(x, \frac{\delta_x}{2}\right) \mid x \in X \right\}$$

covers X so there is a finite subset $C' \subseteq C$ also covering X.

Now take δ to be the minimal radius of the balls in C' (this will be the global parameter we need). For any $x \in X$, we can find some ball in C' containing it, say $B(x_i, \frac{\delta_{x_i}}{2})$. Given $d_x(x, y) < \delta$, y must also be contained in this ball;

$$d_x(x_i, y) \leq d_x(x_i, x) + d_x(x, y) \leq \frac{\delta_{x_i}}{2} + \delta \leq \delta_{x_i}.$$

Then,

$$\begin{aligned} d_y(f(x), f(y)) &\leq d_y(f(x), f(x_i)) + d_y(f(x_i), f(y)) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

Theorem 2.1.4.4 (Rolle's Theorem)**.** Where $f : [a, b] \to \mathbb{R}$ is continuous on its domain, differentiable on (a, b), and satisfies $f(a) = f(b)$, $\exists\, x \in ((a, b),\ f'(x) = 0$.

Proof. The domain is compact f attains its maxima somewhere in it. Suppose that they at least one of them is contained in (a, b) – say (wlog) a maximum on $x \in (a, b)$. By definition, we have $f(x \pm h) \leq f(x)$. Thus,

$$\begin{aligned} f'(x) &= \lim_{h \to 0^+} \frac{f(x + h) - f(x)}{h} \\ &\leq 0. \end{aligned}$$

The inequality follows as the term in the limit is always non-positive; if every term is non-positive/non-negative in a Cauchy sequence, the corresponding real number will be non-positive/non-negative. (Why?) But also,

$$\begin{aligned} f'(x) &= \lim_{h \to 0^-} \frac{f(x + h) - f(x)}{h} \\ &\geq 0. \end{aligned}$$

As f is differentiable on (a, b), $f'(x)$ exists and must be 0.

If both extrema are located on a and b, then by definition $\forall x \in [a, b],\ f(x) = f(a)$ and it follows that f' is 0 everywhere on (a, b). □

Corollary 2.1.4.1 (Mean-value Theorem)**.** Where $f : [a, b] \to \mathbb{R}$ is continuous on its domain and differentiable on (a, b), $\exists\, x \in (a, b),\ f'(x) = \frac{f(b) - f(a)}{b - a}$.

Proof. Consider $g(x) = f(x) - mx$, where $m = \frac{f(b) - f(a)}{b - a}$. Then g satisfies the conditions for Rolle's theorem, and we know that:

$$\exists\, x \in [a, b],\ g'(x) = 0$$

$$\Rightarrow \frac{d}{dx}(f(x) - mx) = 0$$

$$\Rightarrow f'(x) = m$$

□

2.1.5 Limits

Definition 2.1.5.1. Where $f : X \to Y$, $\lim\limits_{x \to x_0} f(x) := [f(x_i)]$ where $[(x_i)] = x_0$. This is equivalent to the epsilon-delta definition.

Proof. Choose $f(x_i)$ in the sequence such that it is within ε of its limit. Then, $d_x(x_i, x_0)$ is the desired δ. □

Proposition 2.1.5.1. Given $f : \mathbb{R}^m \to \mathbb{R}^n$, $f = (f_1, f_2, \ldots, f_n)$,
$\lim\limits_{\vec{x} \to \vec{a}} f(\vec{x}) = \vec{L} \iff \forall i,\ \lim\limits_{\vec{x} \to \vec{a}} f_i(\vec{x}) = L_i$.

Proof.

$$\begin{aligned}
\|f(\vec{x}) - \vec{L}\| \to 0 &\iff \sqrt{\sum_{q=1}^{n} (f_i(\vec{x}) - L_i)^2} \to 0 \\
&\iff \sum_{q=1}^{n} (f_i(\vec{x}) - L_i)^2 \to 0 \\
&\iff \forall i,\ (f_i(\vec{x}) - L_i)^2 \to 0 && \text{(each term is positive)} \\
&\iff \forall i,\ f_i(\vec{x}) \to L_i
\end{aligned}$$

□

Corollary 2.1.5.1. It follows that a vector-valued function is continuous iff all of its components are.

2.2 Differentiation

2.2.1 Definitions

Recall our definition of discrete derivatives in which we took differences in consecutive values of a sequence f. Extending this definition to functions from reals to reals, we find the problem that we can't find consecutive values by completeness. However, we can calculate the derivative by means of a Cauchy sequence, or a limit;

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

Note the introduced denominator. (Why is it needed? Notice that such a denominator is also needed in discrete cases where differences are taken in jumps of 2 or higher.)

Before generalising this to functions from $\mathbb{R}^m$ to $\mathbb{R}^n$, let us take notes on how it is done in a discrete context. Consider the function $f : \mathbb{N}^2 \to \mathbb{N}^3$, $f(m, n) = (n - m, m^2, mn)$:

$f(x,y)$	$f(1,*)$	$f(2,*)$	$f(3,*)$	$f(4,*)$	$f(5,*)$
$f(*,1)$	$(0,1,1)$	$(1,1,2)$	$(2,1,3)$	$(3,1,4)$	$(4,1,5)$
$f(*,2)$	$(-1,4,2)$	$(0,4,4)$	$(1,4,6)$	$(2,4,8)$	$(3,4,10)$
$f(*,3)$	$(-2,9,3)$	$(-1,9,6)$	$(0,9,9)$	$(1,9,12)$	$(2,9,15)$
$f(*,4)$	$(-3,16,4)$	$(-2,16,8)$	$(-1,16,12)$	$(0,16,16)$	$(1,16,20)$
$f(*,5)$	$(-4,25,5)$	$(-3,25,10)$	$(-2,25,15)$	$(-1,25,20)$	$(0,25,25)$

Our function is no longer a sequence, and so a natural extension of the discrete derivative is not immediately clear. We can see that there are two ways we can proceed; we can make a sequence out of this table by choosing a line (either downwards, rightwards, or diagonally) and calculating differences as if values along that line formed a sequence, or we can try to model how the function changes as its input increases by a distance of 1 in the Euclidean/taxicab metric. The former way is the spirit of taking a directional derivative, while the latter is $\tilde{.}$ Rearranging the discrete derivative equation $((\Delta f(n)) * k = f(n+k) - f(n))$ may make generalisation clearer. Extending that equation to our current example, we get something like:

$$(\Delta f(m,n)) \begin{bmatrix} a \\ b \end{bmatrix} = f(m+a, n+b) - f(m,n).$$

That is, $\Delta f(m,n)$ is a 3×2 matrix. From this equation alone however, we are not sure whether $\Delta f(m,n)$ is well-defined (that is, it only has one value) – its value may as well change for different values of a and b. Let us try calculating $\Delta f(2,2)$ for our example. Then, letting the entries of $\Delta f(2,2)$ be represented by a_i, we may calculate the difference of $f(2,2)$ with $f(2,3)$ and $f(3,2)$ (points that distance 1 away in the Euclidean metric):

$$\begin{aligned} \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \\ a_5 & a_6 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} &= f\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) - f\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}\right) \\ &= \begin{bmatrix} 1 \\ 4 \\ 6 \end{bmatrix} - \begin{bmatrix} 0 \\ 4 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \end{aligned}$$

This gives $a_1 = 1$, $a_3 = 0$, $a_5 = 2$, but leaves no information for the other entries. They can be obtained by using the offset (0,1):

$$\begin{aligned} \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \\ a_5 & a_6 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} &= f\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) - f\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}\right) \\ &= \begin{bmatrix} -1 \\ 5 \\ 2 \end{bmatrix} \end{aligned}$$

Giving $a_2 = -1$, $a_4 = 5$, $a_6 = 2$. So our discrete derivative would be:

$$\Delta f(2,2) = \begin{bmatrix} 1 & -1 \\ 0 & 5 \\ 2 & 2 \end{bmatrix}.$$

From this example, it is easy to see that discrete derivatives are always well-defined, when differences between coordinates which are only 1 unit ahead are considered.

Now, to generalise to the reals, we have to consider the limits of values as the offset approaches $\vec{0}$. A naive generalisation from the discrete equation gives the derivative of f at $\vec{x}$ as the unique matrix $D_f(\vec{x})$ satisfying

$$\lim_{\vec{a} \to \vec{0}} f(\vec{x} + \vec{a}) - f(\vec{x}) - D_f(\vec{x})\vec{a} = 0.$$

But the immediate problem is that any matrix satisfies this equation as $\vec{a} \to \vec{0}$. To generalise to the reals, we note that we want our derivative matrix to represent the change in the function given a unit offset in the input. The change in the function is given by the Cauchy sequence:

$$\left\{ \frac{f(\vec{x} + \vec{a}) - f(\vec{x})}{||\vec{a}||} \right\}_{\vec{a}}$$

Of course, this value depends on the choice of the Cauchy sequence for $\vec{a}$ (even if they're all equal to $\vec{0}$). Now, $D_f(\vec{x})\vec{a}$ should approximate the change in the function as the input of the function is offset by $\vec{a}$.

Normalising this to represent an offset by a unit distance gives $\frac{D_f(\vec{x})\vec{a}}{\|\vec{a}\|}$. We want (why?)

$$\left\{\frac{D_f(\vec{x})\vec{a}}{\|\vec{a}\|}\right\}_{\vec{a}} = \left\{\frac{f(\vec{x}+\vec{a}) - f(\vec{x})}{\|\vec{a}\|}\right\}_{\vec{a}}$$

, where the sequence of $\vec{a}$ is the same for both sequences. Rephrasing this gives:

$$\lim_{\vec{a}\to 0} \frac{f(\vec{x}+\vec{a}) - f(\vec{x}) - D_f(\vec{x})\vec{a}}{\|\vec{a}\|} = 0.$$

Again, it may as well be that $D_f(\vec{x})$ depends on the sequence of $\vec{a}$. However, the standard definition for derivatives demands (?) that $D_f(\vec{x})$ should be the same for any such sequence.

To demonstrate the uniqueness of the derivative matrix, suppose that D_1 and D_2 satisfy the conditions to be a derivative matrix. Then:

$$\begin{aligned}\lim_{\vec{a}\to 0}\left(\frac{f(\vec{x}+\vec{a}) - f(\vec{x}) - D_1\vec{a}}{\|\vec{a}\|} - \frac{f(\vec{x}+\vec{a}) - f(\vec{x}) - D_2\vec{a}}{\|\vec{a}\|}\right) &= 0\\ \Rightarrow \lim_{\vec{a}\to 0} \frac{D_2\vec{a} - D_1\vec{a}}{\|\vec{a}\|} &= 0\\ \Rightarrow \lim_{\vec{a}\to 0}(D_2 - D_1)\hat{a} &= 0\end{aligned}$$

$(D_2 - D_1)$ must be the zero matrix, as $(D_2 - D_1)\hat{a} = 0$ for all unit vectors $\hat{a}$.

We note that the existence of a derivative matrix for the reals hinges on the property that the change in the function depends only on how the function changes when one goes along the basiss vectors; that is, we are effectively defining the matrix in the same way for discrete derivatives, by considering the function's behavious along the axes, but then rejecting the derivative if it fails to predict behaviour in any other direction.

Thus, we see that a differentiable function has defined partial derivatives but not necessarily the converse.

(meta: how do lienar functions fit into this? why not multiplicative?)

2.2.2 Theorems

Theorem 2.2.2.1. A C^1 function is differentiable.

Proof. The key idea is that we can successively apply the single-variable mean-value theorem, to approximate one component at a time. Fixing all components but the first, we can treat f as a single-variable function with derivative f_{x_1}, and the mean-value theorem asserts the existence of some $b_1 \in (0, a_1)$ such that:

$$f_{x_1}\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} + \begin{bmatrix}b_1\\0\\\vdots\\0\end{bmatrix}\right) = \frac{f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} + \begin{bmatrix}a_1\\0\\\vdots\\0\end{bmatrix}\right) - f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}\right)}{a_1}$$

$$\Rightarrow f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} + \begin{bmatrix}a_1\\0\\\vdots\\0\end{bmatrix}\right) = f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}\right) + a_1 f_{x_1}\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} + \begin{bmatrix}b_1\\0\\\vdots\\0\end{bmatrix}\right)$$

Similarly, fixing all variables but the second and applying the mean-value theorem again gives:

$$f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}+\begin{bmatrix}a_1\\0\\\vdots\\0\end{bmatrix}+\begin{bmatrix}0\\a_2\\\vdots\\0\end{bmatrix}\right)=f\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}\right)+a_1f_{x_1}\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}+\begin{bmatrix}b_1\\0\\\vdots\\0\end{bmatrix}\right)+a_2f_{x_2}\left(\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}+\begin{bmatrix}a_1\\0\\\vdots\\0\end{bmatrix}+\begin{bmatrix}0\\b_2\\\vdots\\0\end{bmatrix}\right)$$

Continuing in this manner gives:

$$f(\vec{x}+\vec{a})=f(\vec{x})+\sum_{i=1}^{n}a_if_{x_i}(\vec{c_i}),$$

where

$$\vec{c_i}=\vec{x}+\sum_{j=1}^{i-1}a_j\hat{e}_j+b_i\hat{e}_i.$$

Now, let our derivative matrix $D=\begin{bmatrix}f_{x_1}(\vec{x}) & f_{x_2}(\vec{x}) & \cdots & f_{x_n}(\vec{x})\end{bmatrix}$. Then:

$$\begin{aligned}\lim_{\vec{a}\to\vec{0}}\frac{f(\vec{x}+\vec{a})-f(\vec{x})-D\vec{a}}{\|\vec{a}\|}&=\lim_{\vec{a}\to\vec{0}}\frac{\sum_{i=1}^{n}a_if_{x_i}(\vec{c_i})-\sum_{i=1}^{n}a_if_{x_i}(\vec{x})}{\|\vec{a}\|}\\&=\lim_{\vec{a}\to\vec{0}}\sum_{i=1}^{n}\frac{a_i}{\|\vec{a}\|}\left(f_{x_i}(\vec{c_i})-f_{x_i}(\vec{x})\right)\\&=0\end{aligned}$$

The last equality follows by noting that

$$\begin{aligned}\lim_{\vec{a}\to\vec{0}}\vec{c_i}&=\lim_{\vec{a}\to\vec{0}}\left(\vec{x}+\sum_{j=1}^{i-1}a_j\hat{e}_j+b_i\hat{e}_i\right)\\&=\vec{x},\end{aligned}$$

as $b_j\in(0,a_j)\Rightarrow(a_j\to0\Rightarrow b_j\to0)$. It follows that $\lim_{\vec{a}\to\vec{0}}(f_{x_i}(\vec{c_i})-f_{x_i}(\vec{x}))=0$. Also $\frac{a_i}{\|\vec{a}\|}\in[0,1]$. Thus, each term in the sum of the limit above tends to 0, and so the entire limit is equal to 0 (assuming finite dimensions). □

Theorem 2.2.2.2 (Clairaut's Theorem)**.** Where f is C^2, $f_{xy}=f_{yx}$.

Proof. First, we provide some intuition, starting with discrete case.

$$\begin{tikzcd}\begin{array}{|c|c|c|}\hline 1&3&4\\\hline 7&5&6\\\hline 2&1&5\\\hline\end{array} \arrow[r, "\partial_x"] \arrow[d, "\partial_y"] & \begin{array}{|c|c|}\hline 2&1\\\hline -2&1\\\hline -1&4\\\hline\end{array} \arrow[d, "\partial_y"] \\ \begin{array}{|c|c|c|}\hline 6&2&2\\\hline -5&-4&1\\\hline\end{array} \arrow[r, "\partial_x"] & \begin{array}{|c|c|}\hline -4&0\\\hline 1&3\\\hline\end{array}\end{tikzcd}$$

Clairaut's theorem always holds in the discrete case. To see why this is, consider the case for a general 2-by-2 grid:

$$\begin{tikzcd}
\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array} \arrow[r, "\partial_x"] \arrow[d, "\partial_y"] & \begin{array}{|c|}\hline b-a \\ \hline d-c \\ \hline\end{array} \arrow[d, "\partial_y"] \\
\begin{array}{|c|c|}\hline c-a & d-b \\ \hline\end{array} \arrow[r, "\partial_x"] & \boxed{a+d-b-c}
\end{tikzcd}$$

The theorem follows from the fact that $(d-b)-(c-a)=(d-c)-(b-a)$.

We use a similar logic to prove the case for reals, using the fact that

$$(f(a+h,b+k)-f(a+h,b))-(f(a,b+k)-f(a,b)) = (f(a+h,b+k)-f(a,b+k))-(f(a+h,b+k)-f(a,b)).$$

We work with the LHS first. We note that the LHS is a difference between the function where only the first variable has changed. Thus, we may reduce to an expression with ∂_x using the mean-value theorem:

$$\begin{aligned}(f(a+h,b+k)-f(a+h,b))-(f(a,b+k)-f(a,b)) &= h\partial_x[f(x,y+k)-f(x,y)]_{(a+h',b)}\\ &= h\partial_x[kf_y(x,y+k')]_{(a+h',b)}\\ &= hkf_{yx}(a+h',b+k')\end{aligned}$$

And by similar logic on the RHS:

$$(f(a+h,b+k)-f(a,b+k))-(f(a+h,b+k)-f(a,b)) = khf_{xy}(a+h'',b+k'')$$

As LHS=RHS, we get:

$$\begin{aligned}hkf_{yx}(a+h',b+k') &= khf_{xy}(a+h'',b+k'')\\ \Rightarrow f_{yx}(a+h',b+k') &= f_{xy}(a+h'',b+k'') && ((h,k)\neq(0,0))\\ \Rightarrow \lim_{(h,k)\to(0,0)} f_{yx}(a+h',b+k') &= \lim_{(h,k)\to(0,0)} f_{xy}(a+h'',b+k'')\\ \Rightarrow f_{yx}(a,b) &= f_{xy}(a,b).\end{aligned}$$

The last implication follows as f is C^2; the second partial derivatives are continuous. □

Theorem 2.2.2.3 (Chain Rule)**.** Consider $f:\mathbb{R}^a\to\mathbb{R}^b$ and $g:\mathbb{R}^b\to\mathbb{R}^c$, with $\vec{x}\in\text{dom}g$. Then:

$$D_{g\circ f}(\vec{x}) = D_g(f(\vec{x}))D_f(\vec{x}).$$

Proof. We want to say something about $(g\circ f)(\vec{x}+\vec{a})$ with information about $(g\circ f)(\vec{x})$. From the definition of the derivative we can write:

$$f(\vec{x}+\vec{a}) = f(\vec{x}) + D_f(\vec{x})\vec{a} + o(a)$$

with $o(\vec{a})$ denoting the little-o notation; $o(\vec{a})$ satisfies

$$\lim_{\vec{a}\to 0}\frac{o(\vec{a})}{|\vec{a}|} = 0.$$

Now,

$$\begin{aligned}(g\circ f)(\vec{x}+\vec{a}) &= g(f(\vec{x}+\vec{a}))\\ &= g(f(\vec{x})+D_f(\vec{x})\vec{a}+o(\vec{a}))\\ &= g(f(\vec{x})) + D_g(f(\vec{x}))(D_f(\vec{x})\vec{a}+o(\vec{a})) + o(D_f(\vec{x})\vec{a}+o(\vec{a}))\\ &= g(f(\vec{x})) + D_g(f(\vec{x}))D_f(\vec{x})\vec{a} + D_g(f(\vec{x}))o(\vec{a}) + o(D_f(\vec{x})\vec{a}+o(\vec{a}))\end{aligned}$$

Now it suffcies to show that $D_g(f(\vec{x}))o(\vec{a}) + o(D_f(\vec{x})\vec{a} + o(\vec{a})) = o(\vec{a})$. As $D_g(f(\vec{x}))$ is independent of $\vec{a}$, it follows that $D_g(f(\vec{x}))o(\vec{a}) = o(\vec{a})$. For brevity, let $\vec{w} = D_f(\vec{x})\vec{a} + o(\vec{a})$. We now just want to show that $o(\vec{w}) = o(\vec{a})$, i.e.

$$\vec{a} \to 0 \Rightarrow \frac{o(\vec{w})}{\|\vec{a}\|} \to 0$$

By definition, we know that $\vec{w} \to 0 \Rightarrow \frac{o(\vec{w})}{\|\vec{w}\|} \to 0$. But note that if $\vec{a} \to 0$, then $\vec{w} = D_f(\vec{x})\vec{a} + o(\vec{a}) \to 0$, as $D_f(\vec{x})$ is independent of $\vec{a}$. So we have

$$\vec{a} \to 0 \Rightarrow \frac{o(\vec{w})}{\|\vec{w}\|} \to 0.$$

To show that $\frac{o(\vec{w})}{\|\vec{a}\|} \to 0$, we will use the fact that $\frac{o(\vec{w})}{\|\vec{a}\|} = \frac{o(\vec{w})}{\|\vec{w}\|}\frac{\|\vec{w}\|}{\|\vec{a}\|}$, and show that $\frac{\|\vec{w}\|}{\|\vec{a}\|}$ is finite.

$$\begin{aligned} \|D_f(\vec{x})\vec{a} + o(\vec{a})\| &\leq \|D_f(\vec{x})\vec{a}\| + \|o(\vec{a})\| \\ &\leq M\,\|\vec{a}\| + \frac{\|o(\vec{a})\|}{\|\vec{a}\|}\,\|\vec{a}\|\ ^{(4)} \end{aligned}$$

So,

$$\lim_{\vec{a}\to\vec{0}} \frac{\|\vec{w}\|}{\|\vec{a}\|} = \lim_{\vec{a}\to\vec{0}} M + \frac{\|o(\vec{a})\|}{\|\vec{a}\|} = M.$$

Thus,

$$\vec{a} \to 0 \Rightarrow \frac{o(\vec{w})}{\|\vec{w}\|} \to \vec{0} \wedge \frac{\|\vec{w}\|}{\|\vec{a}\|} \to M \Rightarrow \frac{o(\vec{w})}{\|\vec{w}\|}\frac{\|\vec{w}\|}{\|\vec{a}\|} = \frac{o(\vec{w})}{\|\vec{a}\|} \to 0.$$

□

2.2.3 Inverse Function Theorem

Apparently, this theorem is important. (reflect later)

Definition 2.2.3.1. Where f maps a metric space to itself, it is a *contraction* if $d(f(x), f(y)) \leq cd(x, y)$ with $0 < c \leq 1$, and a strict contraction if $0 < c < 1$.

Theorem 2.2.3.1 (Banach Fixed-Point Theorem)**.** Strict contractions have at most one fixed point, and exactly one if X is non-empty and complete.

Proof. If f were to have two fixed points x and y, $d(f(x), f(y)) \leq cd(x, y) \Rightarrow d(x, y) \leq cd(x, y)$, a contradiction.

Now suppose that the metric space X is non-empty and complete. Choose a $x_0 \in X$ and consider the sequence $x_n = [f]^n(x_0)$. That this sequence is Cauchy can be established by noting that $d(x_{n+1}, x_n) \leq cd(x_n, x_{n-1}) \Rightarrow d(x_{n+1}, x_n) \leq c^n d(x_1, x_0)$. Then, one can use the triangle inequality to expand $d(x_m, x_n)$ and use the geometric series formula. We propose that the limit of the sequence, x^* (which exists by completeness), is the fixed point of f.

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} f(x_{n-1}) = f(\lim_{n\to\infty} x_{n-1}) = f(x^*).$$

The penultimate equality is justified by continuity of f. (Why is f continuous?) □

Theorem 2.2.3.2 (Inverse Function Theorem)**.** Where f with an open domain is C^1 with an invertible derivative on x_0, there is an open subset U of the domain containing x_0 on which f is bijective. Also, f^- (on this restriction) satisfies

$$D_{f^-}(f(x_0)) = (D_{f^-}(x_0))^-.$$

[4] Where T is a matrix, $\|T\vec{v}\| = \|\vec{v}\|\,\|T\hat{v}\|$. But $f : \{\vec{x} \mid \|\vec{x}\| = 1\} \to \mathbb{R},\ f(\vec{x}) = \|T\hat{v}\|$ has a compact domain, so attains a maximum, say a value of M. Then, $\|T\vec{v}\| \leq M\,\|\vec{v}\|$.

The main part of the proof, as presented by Tao, is not exactly intuitive. It hinges on a corollary of the fixed-point theorem that states that if g is strictly contractive on $U(0,r)$, $f = g + I$ is injective on that neighbourhood and $U(0,(1-c)r) \subset f(U(0,r)) \subset U(0,(1+c)r)$. The proof is as follows.

If $g(x) = g(y)$, then as g is a contraction, this implies that $x = y$.

To show that the image of $U(0,r)$ under f includes $U(0,(1-c)r)$, we "cleverly" abuse the fixed-point theorem. see wikipedia.

Proof. The second part is trivial. Assume that f has an inverse. Then,

$$D_{f^- \circ f}(x) = I \Rightarrow D_{f^-}(f(x))D_f(x) = I \Rightarrow D_{f^-}(f(x)) = (D_f(x))^-.$$

I cbf writing this shit just read the stuff on wikipedia the c1 argument is better done in Tao, tho. □

3 Applications

3.1 Taylor Series

3.1.1 Discrete Analogy

We have previously discussed Taylor series for discrete sequences. We revisit the discrete case but in a slightly different perspective, to allow a clearer generalisation to the continuous case.

Noting that $f(x+1) = f(x) + \Delta f(x)$, $f(x+2) = f(x) + \Delta f(x) + \Delta f(x+1)$, and so on, we can write:

$$\begin{aligned} f(x+a) &= f(x) + \Delta f(x) + \Delta f(x+1) + \cdots + \Delta f(x+a-1) \\ &= f(x) + \sum_{q=0}^{a-1} \Delta f(x+q) \end{aligned}$$

But now, we note that in turn, we can write

$$\begin{aligned} \Delta f(x+q) &= \Delta f(x) + \Delta^2 f(x) + \Delta^2 f(x+1) + \cdots + \Delta^2 f(x+q-1) \\ &= \Delta f(x) + \sum_{q_2=0}^{q-1} \Delta^2 f(x+q_2). \end{aligned}$$

Substituting this into () we get:

$$\begin{aligned} f(x+a) &= f(x) + \sum_{q=0}^{a-1} \Delta f(x+q) \\ &= f(x) + \sum_{q=0}^{a-1} \left(\Delta f(x) + \sum_{q_2=0}^{q-1} \Delta^2 f(x+q_2) \right) \\ &= f(x) + \Delta f(x) a + \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} \Delta^2 f(x+q_2) \end{aligned}$$

Repeating the process once more:

$$\begin{aligned} f(x+a) &= f(x) + \Delta f(x) a + \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} \Delta^2 f(x+q_2) \\ &= f(x) + \Delta f(x) a + \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} \left(\Delta^2 f(x) + \sum_{q_3=0}^{q_2-1} \Delta^3 f(x+q_3) \right) \\ &= f(x) + \Delta f(x) a + \Delta^2 f(x) \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} (1) + \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} \sum_{q_3=0}^{q_2-1} \Delta^3 f(x+q_3) \\ &= f(x) + \Delta f(x) a + \Delta^2 f(x) \frac{a^2}{2} + \sum_{q=0}^{a-1} \sum_{q_2=0}^{q-1} \sum_{q_3=0}^{q_2-1} \Delta^3 f(x+q_3) \end{aligned}$$

Noting that

$$\begin{aligned} \sum_{q_1=0}^{a-1} \sum_{q_2=0}^{q_1-1} \cdots \sum_{q_{k-2}=0}^{q_{k-3}-1} \sum_{q_{k-1}=0}^{q_{k-2}-1} \sum_{q_k=0}^{q_{k-1}-1} (1) &= \sum_{q_1=0}^{a-1} \sum_{q_2=0}^{q_1-1} \cdots \sum_{q_{k-2}=0}^{q_{k-3}-1} \sum_{q_{k-1}=0}^{q_{k-2}-1} (q_{k-1} - 0) \\ &= \sum_{q_1=0}^{a-1} \sum_{q_2=0}^{q_1-1} \cdots \sum_{q_{k-2}=0}^{q_{k-3}-1} (\frac{q_{k-2}^2}{2} - 0) \\ &\vdots \end{aligned}$$

$$= \frac{a^{\underline{k}}}{k!},$$

we see that the above formulation is equal to the Newton series we derived before, as eventually $\delta^i f$ becomes 0, and so the iterated sum at the end eventually disappears:

$$f(x+a) = \sum_{q=0}^{k} \frac{k^{\underline{q}}}{q!}[\Delta^q f](x).$$

However when discussing error terms for the continuous Taylor series, it is more convenient to work with this form:

$$f(x+a) = \underbrace{f(x) + \Delta f(x)a + \Delta^2 f(x)\frac{a^{\underline{2}}}{2}}_{\text{Taylor Polynomial}} + \underbrace{\sum_{q=0}^{a-1}\sum_{q_2=0}^{q-1}\sum_{q_3=0}^{q_2-1} \Delta^3 f(x+q_3)}_{\text{Error}}$$

We can give an upper bound for the error term. For the error term for the $k-1^{\text{st}}$ Taylor polynomial, let

$$M = \max_{n\in[0,a]} \left\{\Delta^k f(x+q)\right\}.$$

Then:

$$\begin{aligned}
\sum_{q_1=0}^{a-1}\sum_{q_2=0}^{q_1-1}\cdots\sum_{q_k=0}^{q_{k-1}-1} \Delta^k f(x+q_k) &\leq \sum_{q_1=0}^{a-1}\sum_{q_2=0}^{q_1-1}\cdots\sum_{q_k=0}^{q_{k-1}-1} M \\
&= M\sum_{q_1=0}^{a-1}\sum_{q_2=0}^{q_1-1}\cdots\sum_{q_k=0}^{q_{k-1}-1} (1) \\
&= M\frac{a^{\underline{k}}}{k!}.
\end{aligned}$$

Now, extending this to the continuous case is almost trivial.

3.1.2 Continuous Taylor series

First, we write:

$$f(x+a) = f(x) + \int_0^a f'(x+q)\,dq$$

But now, we note that in turn, we can write

$$f'(x+q) = f'(x) + \int_0^q f''(x+q_2)\,dq_2$$

Substituting this into () we get:

$$\begin{aligned}
f(x+a) &= f(x) + \int_0^a f'(x+q)\,dq \\
&= f(x) + \int_0^a \left(f'(x) + \int_0^q f'(x+q_2)\,dq_2\right)\,dq \\
&= f(x) + \int_0^a f'(x)\,dq + \int_0^a\int_0^q f''(x+q_2)\,dq_2\,dq \\
&= f(x) + f'(x)a + \int_0^a\int_0^q \left(f'(x) + \int_0^q f''(x+q_2)\,dq_2\right)\,dq_2\,dq
\end{aligned}$$

Repeating the process once more:

$$f(x+a) = f(x) + f'(x)a + \int_0^a\int_0^q f''(x+q_2)\,dq_2\,dq$$

$$
\begin{aligned}
&= f(x) + f'(x)a + \int_0^a \int_0^q \left(f''(x) + \int_0^{q_2} f'''(x+q_3)\, dq_3 \right) dq_2\, dq \\
&= f(x) + f'(x)a + \int_0^a \int_0^q f''(x)\, dq_2\, dq + \int_0^a \int_0^q \int_0^{q_2} f'''(x+q_3)\, dq_3\, dq_2\, dq \\
&= f(x) + f'(x)a + f''(x) \int_0^a \int_0^q dq_2\, dq + \int_0^a \int_0^q \int_0^{q_2} f'''(x+q_3)\, dq_3\, dq_2\, dq \\
&= \underbrace{f(x) + f'(x)a + f''(x)\frac{a^2}{2}}_{\text{Taylor Polynomial}} + \underbrace{\int_0^a \int_0^q \int_0^{q_2} f'''(x+q_3)\, dq_3\, dq_2\, dq}_{\text{Error}}
\end{aligned}
$$

We can generalise the Taylor polynomial term by noting that

$$\int_0^a \int_0^{q_1 - 1} \cdots \int_0^{q_{k-1}-1} dq_k \cdots dq_2\, dq_1 = \frac{a^k}{k!}.$$

Thus the k^{th} degree Taylor polynomial of $f(x+a)$ centred at x is given by:

$$\sum_{q=0}^{k} \frac{a^q}{q!} f^{(q)}(x).$$

As before, we can give an upper bound for the error term for the $k-1^{\text{st}}$ Taylor polynomial. Let

$$M = \max_{q \in [0,a]} \left\{ f^{(k)}(x+q) \right\}.$$

Then:

$$
\begin{aligned}
\int_0^a \int_0^{q_1 - 1} \cdots \int_0^{q_{k-1}-1} f^{(k)}(x+q_k)\, dq_k \cdots dq_2\, dq_1 &\leq \int_0^a \int_0^{q_1 - 1} \cdots \int_0^{q_{k-1}-1} M\, dq_k \cdots dq_2\, dq_1 \\
&= M \int_0^a \int_0^{q_1 - 1} \cdots \int_0^{q_{k-1}-1} dq_k \cdots dq_2\, dq_1 \\
&= M \frac{a^k}{k!}
\end{aligned}
$$

By a similar argument, $I\frac{a^k}{k!}$ is a lower bound for the error where

$$I = \min_{q \in [0,a]} \left\{ f^{(k)}(x+q) \right\}.$$

Thus, this gives us $I\frac{a^k}{k!} \leq \text{Error} \leq M\frac{a^k}{k!}$, and so $\exists\, C \in [I, M]$ such that Error $= C\frac{a^k}{k!}$. By the intermediate value theorem, $\exists\, q \in [0,a]$ such that Error $= f(x+q)\frac{a^k}{k!}$.

3.1.3 Multivariate Taylor Series

To extend our Taylor series to multivaraite continuous functions, it is most convenient to interpret the multivariate function as a univariate function using directional derivatives, rather than using the bottom-up approach used previously.

$$
\begin{aligned}
f(\vec{x} + \vec{a}) &= f(\vec{x} + |\vec{a}|\hat{a}) \\
&= f(\vec{x}) + \int_0^{|\vec{a}|} D_{\hat{a}} f(\vec{x} + q\hat{a})\, dq \\
&= f(\vec{x}) + \int_0^{|\vec{a}|} D_{\hat{a}} f(\vec{x}) + \int_0^q D_{\hat{a}}^2 f(\vec{x} + q_2\hat{a})\, dq_2\, dq^{(5)} \\
&= f(\vec{x}) + D_{\hat{a}} f(\vec{x})|\vec{a}| + \int_0^{|\vec{a}|} \int_0^q D_{\hat{a}}^2 f(\vec{x} + q_2\hat{a})\, dq_2\, dq
\end{aligned}
$$

$$\vdots$$

$$= \underbrace{f(\vec{x}) + D_{\hat{a}}f(\vec{x})|\vec{a}| + D_{\hat{a}}^2 f(\vec{x})\frac{|\vec{a}|^2}{2}}_{\text{Taylor Polynomial}} + \underbrace{\int_0^{|\vec{a}|}\int_0^{q}\int_0^{q_2} D_{\hat{a}}^3 f(\vec{x}+q_3\hat{a})\,dq_3\,dq_2\,dq}_{\text{Error}}$$

Thus, the Taylor polynomial is given by:

$$\sum_{q=0}^{k}\frac{|\vec{a}|^q}{q!}D_{\hat{a}}^q f(\vec{x}).$$

Using our formula for directional derivatives,

$$\begin{aligned} D_{\vec{a}}f &= D_f\vec{a} \\ &= \nabla f\cdot\vec{a} \\ &= [\partial_{x_1}f,\ldots,\partial_{x_n}f]\begin{bmatrix} a_1 \\ \vdots \\ a_n\end{bmatrix} \\ &= \sum_{i=1}^{n} a_i\partial_{x_i}f \end{aligned}$$

and noting that

$$\begin{aligned} |\vec{a}|D_{\hat{a}} &= |\vec{a}|D_f\hat{a} \\ &= D_f\vec{a} \\ &= D_{\vec{a}}f, \end{aligned}$$

we may express the Taylor polynomial in terms of partials:

$$\begin{aligned} \sum_{q=0}^{k}\frac{|\vec{a}|^q}{q!}D_{\hat{a}}^q f(\vec{x}) &= \sum_{q=0}^{k}\frac{1}{q!}D_{\vec{a}}^q f(\vec{x}) \\ &= \sum_{q=0}^{k}\frac{1}{q!}\left[[D_{\vec{a}}]^q f\right](\vec{x}) \\ &= \sum_{q=0}^{k}\frac{1}{q!}\left[\left[\sum_{i=1}^{n}a_i\partial_{x_i}\right]^q f\right](\vec{x}) \end{aligned}$$

As before, we can give an upper bound for the error term for the $k-1^{\text{st}}$ Taylor polynomial. Let

$$M = \max_{q\in[0,|\vec{a}|]}\left\{D_{\hat{a}}^k f(\vec{x}+q\hat{a})\right\}.$$

Then the upper bound for the error is given by

$$\begin{aligned} \frac{|\vec{a}|^k}{k!}M &= \frac{|\vec{a}|^k}{k!}D_{\hat{a}}^k f(\vec{x}+q'\hat{a}) \\ &= \frac{1}{k!}D_{\vec{a}}^k f(\vec{x}+q'\hat{a}) \\ &= \frac{1}{k!}\left[\left[\sum_{i=1}^{n}a_i\partial_{x_i}\right]^k f\right](\vec{x}+q'\hat{a}) \end{aligned}$$

[5] Note the potentially confusing notation here. $D_{\hat{a}}^q f(\vec{x})$ means to first apply the operator $[D_{\hat{a}}]^q$ to f, then use the resulting operator on $\vec{x}$.

3.2 Classification of points in 2 dimensions

If at a point $\vec{x}$, $D_f(\vec{x})$ is the zero matrix, and the the second directional derivative is:

- positive for all directions, then that point is a local minimum.
- negative for all directions, then that point is a local maximum.
- positive for some directions and negative for others, then that point is a saddle.

So, we want to know when $\left[D_{\vec{a}}f\right]^2(\vec{x})$ is positive or negative. The problem becomes simple when the second directional derivative is re-expressed:

$$\begin{aligned}
\left[D_{\vec{a}}\right]^2 f &= D_{\vec{a}}(D_{\vec{a}}f)\\
&= D_{\vec{a}}(D_f\vec{a})\\
&= D_{\vec{a}}(\begin{bmatrix} f_{x_1} & f_{x_2} & \cdots & f_{x_n}\end{bmatrix}\vec{a})\\
&= D_{\vec{a}}(\begin{bmatrix} a_1 f_{x_1} + a_2 f_{x_2} + \cdots + a_n f_{x_n}\end{bmatrix})\\
&= D(\begin{bmatrix} a_1 f_{x_1} + a_2 f_{x_2} + \cdots + a_n f_{x_n}\end{bmatrix})\vec{a}\\
&= \begin{bmatrix} a_1 f_{x_1x_1} + \cdots + a_n f_{x_nx_1} & a_1 f_{x_1x_2} + \cdots + a_n f_{x_nx_2} & \cdots & a_1 f_{x_1x_n} + \cdots + a_n f_{x_nx_n}\end{bmatrix}\vec{a}\\
&= \begin{bmatrix} a_1 & a_2 & \cdots & a_n\end{bmatrix}\begin{bmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n}\\ f_{x_2x_1} & f_{x_2x_2} & & f_{x_2x_n}\\ \vdots & & \ddots & \vdots\\ f_{x_nx_1} & f_{x_nx_2} & \cdots & f_{x_nx_n}\end{bmatrix}\vec{a}\\
&= \vec{a}^T H_f \vec{a}
\end{aligned}$$

Thus, the problem reduces to knowing when $\vec{a}^T H_f \vec{a}$ is positive or negative when applied to a point $\vec{x}$. Note however, that H_f is a symmetric matrix, and so there exists a matrix P such that $P^{-}DP = H_f$, where D is diagonal. See 4.2.1 for a proof. (Note that the eigenvalues are functions, as we have left H_f as an operator.)

Theorem 3.2.0.1. For a symmetric matrix H_f with eigenvalues $\lambda_i > 0$, $i \in \{1, ..., n\}$,

$$\forall \vec{u} \neq \vec{0}, \quad \vec{u}^T H_f \vec{u} > 0 \iff \forall i \in \{1, ..., n\}, \quad \lambda_i > 0,$$

and the same statement with $<$ replaced by $\leq$, $>$ or $\geq$ also holds. Also,

$$\exists\, \vec{u}, \vec{w} \text{ s.t } \vec{u}^T H_f \vec{u} > 0 \wedge \vec{w}^T H_f \vec{w} < 0 \iff \exists\, i, j \text{ s.t } \lambda_i > 0 \wedge \lambda_j < 0.$$

Proof. Let $\vec{v} = P\vec{u}$, where P is the appropriate change of basis matrix. $P^T = P^{-}$ by orthogonal diagonalisation.

$$\begin{aligned}
\vec{u}^T H_f \vec{u} &= \vec{u}^T P^T D P \vec{u}\\
&= \vec{v}^T D \vec{v}\\
&= \begin{bmatrix} v_1 & \cdots & v_n\end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}\begin{bmatrix} v_1\\ \vdots\\ v_n\end{bmatrix}\\
&= \sum_{i=1}^{n} \lambda_i v_i^2
\end{aligned}$$

So $\lambda_i > 0 \Rightarrow \forall \vec{u} \neq \vec{0}$, $\vec{u}^T H_f \vec{u} > 0$. At least one v_i is guaranteed to be non-zero as P is a basis.

Conversely, for an eigenvalue λ_i, consider its unit eigenvector $\hat{v}$. $\hat{v}^T H_f \hat{v} > 0$ by hypothesis but $\hat{v}^T H_f \hat{v} = \hat{v}^T \lambda_i \hat{v} = \lambda_i$, so $\lambda_i > 0$. The other two statements follow similarly. □

Definition 3.2.0.1. A symmetric matrix is [positive/positive semi-/negative/negative semi-] definite if all eigenvalues are [$>$ / $\geq$ / $<$ / $\leq$] 0, respectively. It is indefinite if it is neither postivie semi- or negative semi-definite.

From our theorem above, we see that a critical point is a minimum/maximum iff H_f positive/negative semi-definite on it.

An effective method to check positive/negative definiteness of a Hermitian matrix is through Sylvester's Criterion, to which we give an unintuitive proof below.

Theorem 3.2.0.2 (Sylvester's Criterion)**.** A matrix M is positive definite iff all its leading principal minors are positive.

Proof. Let δ_k denote the k^{th} leading principal minor. We prove the theorem by induction. The base case is trivial. Now assuming that the theorem holds for matrices of size $n \times n$, let A be a $n+1 \times n+1$ matrix with all leading principal minors positive.

To show that A cannot have 2 or more negative eigenvalues, we consider the unintuitively clever construction $\vec{w} = v_n\vec{u} - u_n\vec{v}$, where $\vec{u}$ and $\vec{v}$ are eigenvectors corresponding to the two negative eigenvalues. By construction, $w_n = 0$ so $\vec{w}^T A\vec{w}$ is equivalent to evaluating the quadratic form of $A^{(n)}$ (the $n \times n$ upper-left submatrix) at $(w_1, \ldots, w_n)$, which is greater than 0 by hypothesis – $A^{(n)}$ is positive definite. But,

$$\begin{aligned}(v_n\vec{u} - u_n\vec{v})^T A(v_n\vec{u} - u_n\vec{v}) &= (v_n\vec{u}^T - u_n\vec{v}^T)A(v_n\vec{u} - u_n\vec{v}) \\ &= (v_n\vec{u}^T - u_n\vec{v}^T)(Av_n\vec{u} - Au_n\vec{v}) \\ &= v_n\vec{u}^T(Av_n\vec{u} - Au_n\vec{v}) - u_n\vec{v}^T(Av_n\vec{u} - Au_n\vec{v}) \\ &= v_n^2\vec{u}^T A\vec{u} + u_n^2\vec{v}^T A\vec{v} \\ &< 0.\end{aligned}$$

Also, there cannot be a single negative eigenvalue by the positivity of δ_{n+1}. This also implies that no eigenvalues are 0, and thus all eigenvalues are positive. □

Note that the determinant of a matrix is the product of its eigenvalues. The proof is as follows. The solutions to the characteristic equation $\det(A - \lambda I)$ is given by the eigenvalues λ_i. So

$$\det(A - \lambda I) = a\prod_{i=1}^{n}(\lambda - \lambda_i) \qquad (a \in \mathbb{R})$$

From the Laplace algorithm for determinants, it can be inductively seen that the leading coefficient of the characteristic polynomial is $(-1)^n$. And so, for the leading coefficients to match on both sides, we require $a = (-1)^n$. Now, consider the constant term of the characteristic polynomial, which is given when $\lambda = 0$. This gives

$$\det(A) = (-1)^n\prod_{i=1}^{n}(\lambda)$$

And so for 2×2 matrices, $\det(A) > 0 \wedge a_{1,1} > 0 \Leftrightarrow \lambda_i > 0$, $\det(A) > 0 \wedge a_{1,1} < 0 \Leftrightarrow \lambda_i < 0$ and $\det(A) < 0 \Leftrightarrow \lambda_1 > 0 \wedge \lambda_2 < 0$.

3.2.1 Lagrange Multipliers

Suppose we wanted to identify the extrema of f , not on its domain but on $\vec{x}$ such that $g(\vec{x}) = 0$. Then, intuitively it follows that $\vec{x}_0$ is an extrema iff $D_{\vec{a}}f(\vec{x}_0) = 0$ for all $\vec{a}$ such that one is able to travel in the direction of $\vec{a}$ and still satisfy g – that is, for some small enough ε, $g(\vec{x} + t\vec{a}) = 0$, $\forall t \in [0, \epsilon]$. An equivalent characterisation is that $\vec{a}$ satisfies $D_{\vec{a}}g(\vec{x}_0) = 0$.

To avoid headaches, I skip any proof of Lagrange multipliers.

3.3 Vector Calculus

Definition 3.3.0.1. A scalar/vector field is a function from vectors to scalars/vectors.

Definition 3.3.0.2.

$$\nabla := \begin{bmatrix} D_{x_1} \\ D_{x_2} \\ \vdots \\ D_{x_n} \end{bmatrix}.$$

Definition 3.3.0.3. The tangent, normal, and binormal vectors $\vec{T}$, $\vec{N}$ and $\vec{B}$ of a curve $\vec{r}(t)$ are defined as:

$$\vec{T} := \frac{dr}{dt}\frac{dt}{ds}$$

Note that $s(t+\varepsilon) - s(t) = \|\vec{r}(t+\varepsilon) - \vec{r}(t)\|$

Alright, being rigorous here probably makes things worse. We proceed with intuition. Consider a curve $\vec{r}(t)$. It is easy to see that the direction of $\vec{r}\,'(t)$ signifies the direction of where the curve is heading, while the magnitude is the speed at which the arclength is being covered. Let $s(t)$ be the arclength as a function of time. Then,

$$s'(t) := \|\vec{r}\,'(t)\|,$$

so

$$s(t) = \int_{t_0}^{t} \|\vec{r}\,'(\tau)\| \, d\tau.$$

From now on, we also parametrise our curve with arclength. This is conveninent when defining concepts such as curvature and torsion, which otherwise depend on the speed at which one travels along the curve. When the curve is parametrised by arclength, we remove this ambiguity by fixing the speed.

To perhaps make the distinction clearer we let $\vec{\varsigma}$ denote the same curve parametrised by arclength; that is, $\vec{\varsigma} \circ s = \vec{r}$ (or if the notation is confusing, $\vec{\varsigma}(s(t)) = \vec{r}(t)$). Also, $\vec{T_s}$ and $\vec{T_t}$ will be the tangent vector parametrised by arclength and time respectively, with $\vec{T_s} \circ s = \vec{T_t}$, with similar conventions for the normal and binormal vectors. Also, X' denotes the derivative of X with respect to its parametrisation.

We define the tangent vector $\vec{T_t}$ to be the unit vector pointing in the direction of $\vec{r}\,'$; that is,

$$\begin{aligned} \vec{T_t} &:= \frac{\vec{r}\,'}{\|\vec{r}\,'\|} \\ &= \frac{(\vec{\varsigma} \circ s)'}{s'} \\ &= \frac{(\vec{\varsigma}\,' \circ s)s'}{s'} && \text{(Chain Rule)} \\ &= \vec{\varsigma}\,' \circ s \\ &=: \vec{T_s} \circ s. \end{aligned}$$

Note that we need $\vec{r}\,' \neq \vec{0}$.

We quantify how much the direction of the curve is changing by analysing the change in the tangent vector i.e change in the direction of the curve. Thus, the direction of $\vec{T_s}\,'$ signifies the direction of where the direction of the curve is heading, while the magnitude quantifies how fast the direction is changing. (Note that by using arclength, we quantify the change in direction per change in unit arclength, instead of change in direction per change in distance travelled in 1 unit of time.) This magnitude is assigned the "curvature", symbolised by κ. The normal vector $\vec{N_s}$ is defined to be the unit vector pointing in the direction of $\vec{T_s}\,'$; that is,

$$\kappa := \left\|\vec{T_s}\,'\right\|;$$

$$\begin{aligned}\vec{N}_s &:= \frac{\vec{T}_s{}'}{\left\|\vec{T}_s{}'\right\|} \\ \Rightarrow \vec{T}_s{}' &= \kappa \vec{N}_s,\end{aligned}$$

and

$$\begin{aligned}(\vec{T}_s \circ s)' &= (\vec{T}_s{}' \circ s)s' \\ \Rightarrow T_t{}' &= (\kappa \vec{N}_s \circ s)s' \\ &= s'\kappa \vec{N}_s \circ s \\ &= s'\kappa \vec{N}_t.\end{aligned}$$

We note that a vector valued function is always orthogonal to its derivative, given it has a constant norm. A simple proof is by using the product rule for dot products.

We now want to define the torsion of the curve, τ. If one were to keep the tangent vector constant, the curve would travel in a single direction, with 0 curvature; if one were to keep the tangent and normal vectors constant, the curve would travel in a single plane, with 0 torsion.

We define the binormal vector $\vec{B} := \vec{T} \times \vec{N}$ to characterise the plane on which the curve would live in if the tangent and normal vectors at that instant were to be kept constant. The change in this vector, $\vec{B}_s{}'$, quantifies (...) the change in the normal vector of the plane (per unit change in arclength). It is hard to intuitively reduce this vector into a single word we are familiar with (maybe there is idk). However, it turns out that this direction is equivalent to $\vec{N}$.

$$\begin{aligned}\vec{B}_s{}' &= (\vec{T}_s \times \vec{N}_s)' \\ &= \vec{T}_s{}' \times \vec{N}_s + \vec{T}_s \times \vec{N}_s' \\ &= \vec{T}_s \times \vec{N}_s'\end{aligned}$$

Noting that $\vec{B}_s{}'$ is orthogonal to $\vec{B}$ (as it is unit (because $\vec{T}$ and $\vec{N}$ are unit)) and $\vec{T}$ (from above), we conclude that it must be pointing in the same direction (up to a factor of -1) as $\vec{N}$. (it seems like theres not factor of -1 though?) By convention, the torsion is defined to be the negative of the coefficient of $\vec{N}_s$:

$$\begin{aligned}\vec{B}_s{}' &= -\tau \vec{N}_s; \\ \vec{B}_t{}' &= -s'\tau \vec{N}_t;\end{aligned}$$

Theorem 3.3.0.1 (Serret-Frenet Formula)**.** When parametrised by arclength, we have:

$$\begin{bmatrix}\vec{T}\\ \vec{N}\\ \vec{B}\end{bmatrix}' = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\vec{T}\\ \vec{N}\\ \vec{B}\end{bmatrix}.$$

Proof. We have already shown the top and bottom rows. We show the middle row by proving the matrix must be skew-symmetric. Let

$$Q = \begin{bmatrix}\vec{T}\\ \vec{N}\\ \vec{B}\end{bmatrix}.$$

We note that Q forms an orthonormal basis, so $Q^{-} = Q^T$. The matrix in question is $Q'Q^{-} = Q'Q^T$. The trick is apply the product rule to $QQ^T = I$:

$$\begin{aligned}(QQ^T)' &= Q'Q^T + Q(Q^T)' \\ &= Q'Q^T + Q(Q')^T\end{aligned}$$

$$
\begin{aligned}
&= Q'Q^T + (Q'Q^T)^T \\
\Rightarrow Q'Q^T &= -(Q'Q^T)^T
\end{aligned}
$$

□

Corollary 3.3.0.1. Using the chain rule, we get:

$$\begin{bmatrix} \vec{T_t} \\ \vec{N_t} \\ \vec{B_t} \end{bmatrix}' = s' \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \vec{T_t} \\ \vec{N_t} \\ \vec{B_t} \end{bmatrix}.$$

The form in the corollary leads to explicit expressions for curvature and torsion in terms of $\vec{r}$. A direct approach from trying to calculate κ as $\left\|\vec{T_s}'\right\|$ does not work (for some reason). Instead, we may guess that computing κ and τ involves taking the second and third derivatives of $\vec{r}$, and compute those expressions first. We try to calculate τ. κ will follow easily. We have

$$
\begin{aligned}
\vec{r}' &= \|\vec{r}'\| T_t \\
&= s'T_t,
\end{aligned}
$$

and so

$$
\begin{aligned}
\vec{r}'' &= s''T_t + s'T_t' \\
&= s''T_t + (s')^2\kappa\vec{N_t},
\end{aligned}
$$

and

$$
\begin{aligned}
\vec{r}''' &= s'''T_t + s''T_t' + 2s's''\kappa\vec{N_t} + (s')^2\kappa\vec{N_t}' \\
&= s'''T_t + s''s'\kappa\vec{N_t} + 2s's''\kappa\vec{N_t} + (s')^3\kappa(-\kappa\vec{T_t} + \tau\vec{B_t}) \\
&= (s''' - (s')^3\kappa^2)T_t + 3s's''\kappa\vec{N_t} + (s')^3\kappa\tau\vec{B_t}
\end{aligned}
$$

We can isolate the τ term in the coefficient of $\vec{B_t}$ by taking a dot product with $\vec{B_t}$. Noting that our expressions for $\vec{r}'$ and $\vec{r}''$ include $\vec{T}$ and $\vec{N}$, and $\vec{B} = \vec{T} \times \vec{N}$, we obtain an expression for the binormal vector:

$$
\begin{aligned}
\vec{r}' \times \vec{r}'' &= (s')^3\kappa\vec{B_t} \\
\Rightarrow \|\vec{r}' \times \vec{r}''\| &= (s')^3\kappa \\
\Rightarrow (\vec{r}' \times \vec{r}'') \cdot \vec{r}''' &= ((s')^3\kappa)^2\tau \\
\Rightarrow \tau &= \frac{(\vec{r}' \times \vec{r}'') \cdot \vec{r}'''}{\|\vec{r}' \times \vec{r}''\|^2}.
\end{aligned}
$$

4 Linear Algebra

4.1 Linear Operators

4.2 Eigen

Diagonal matrices are particularly easy to work with, having the effect of scaling axes. In particular, their powers are simply powers of each element. Abusing the change of basis concept from before, we can simplify certain matrices by showing they are equivalent to diagonal matrices under a certain basis.

For example, asdasd. However, not all matrices are diagonal under some basis; consider the rotation matrix asdasdasd.

In general, for a matrix A to be diagonalisable, there needs to exist a change of basis matrix $P = P_{S\to B} = \begin{bmatrix}\vec{p}_1 & \vec{p}_2 & \dots & \vec{p}_n\end{bmatrix}$ such that $P^{-}AP$ is diagonal. This means that

$$\begin{aligned} AP &= P\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{bmatrix}\\ \Rightarrow A\vec{p_i} &= \lambda_i p_i, \end{aligned}$$

that is, p_i are eigenvectors of A. Notice that the invertibility requirement on P implies that each vector in P be independent, that is, each eigenvector of A be independent.

4.2.1 Spectral Theorem

Definition 4.2.1.1. A matrix A is Hermitian iff $A = A^{\mathsf{H}}$, where A^{H} denotes the conjugate transpose of A. An equivalent formulation that generalises to operators is that it satisfies $\langle A\vec{u}, \vec{w}\rangle = \langle \vec{u}, A\vec{w}\rangle$ – that is, it is equivalent to its adjoint.

Note; hermitian operator is an operator from a finite vector space V to itself.

Theorem 4.2.1.1 (Hermitian Spectral Theorem)**.** The eigenvectors of a Hermitian matrix can form an orthonormal basis.

Proof. The proof uses induction on the size n. First we note that a Hermitian operator is guaranteed to have at least one eigenvector over complex vector spaces, by the fundamental theorem of algebra. The key step in induction is noting that Hermiticity guarantees preservation of subspace-invariance under orthogonal complementation.

Choose an eigenvector $\vec{v}$. Its span is invariant under $\mathcal{A}$ (the operator corresponding to A), so $\operatorname{span}(\vec{v})^{\perp}$ is also. But this means that $\mathcal{A}$ is a Hermitian operator on $\operatorname{span}(\vec{v})^{\perp}$, so we may choose another eigenvector in $\operatorname{span}(\vec{v})^{\perp}$, and continue this process until we exhaust dimensions. Clearly all eigenvectors chosen are orthogonal.

It remains to prove that Hermiticity guarantees preservation of subspace-invariance under orthogonal complementation. Suppose that $\langle \vec{v}, \vec{u}\rangle = 0$, where $\vec{v}$ is an eigenvector of A. Then,

$$\begin{aligned}\langle A\vec{u}, A\vec{v}\rangle &= \langle A^2\vec{u}, \vec{v}\rangle \\ &= \lambda^2\langle \vec{u}, \vec{v}\rangle \\ &= 0.\end{aligned}$$

□

Corollary 4.2.1.1. The above result holds also for symmetric maps under real inner product spaces.

Proof. The proof above need not be modified, as we are guaranteed the existence of eigenvectors as all eigenvalues of symmetric matrices are real. The proof follows from considering two expression equivalent to $\vec{v}^{\mathsf{H}}A\vec{v}$, where $\vec{v}$ is an eigenvector of A.

$$\begin{aligned}\vec{v}^{\mathsf{H}}A\vec{v} &= \vec{v}^{\mathsf{H}}\lambda\vec{v} \\ &= \lambda\vec{v}^{\mathsf{H}}\vec{v} \\ &= \lambda\left\|\vec{v}\right\|.\end{aligned}$$

But also,

$$\begin{aligned}\vec{v}^{\mathsf{H}}A\vec{v} &= \vec{v}^{\mathsf{H}}A^{\mathsf{H}}\vec{v} \\ &= (A\vec{v})^{\mathsf{H}}\vec{v} \\ &= (\lambda\vec{v})^{\mathsf{H}}\vec{v} \\ &= \overline{\lambda}\vec{v}^{\mathsf{H}}\vec{v}.\end{aligned}$$

Thus, $\overline{\lambda} = \lambda$ and we are done. □

5 Omitted Details

2.1.3

$$\begin{aligned}
S \setminus (A \cap S) &= S \cap (A \cap S)' \qquad (1)\\
&= S \cap (A' \cup S')\\
&= S \cap A'\\
&= S \cap (A \cup B) \cap A'\\
&= S \cap B \cap A'\\
&= S \cap B \qquad (B \subseteq A')
\end{aligned}$$