Directed Reading Program, Spring 2022

Calculus of Variations

October 4, 2022
# Contents

1 What is Calculus of Variations? ............................... 5
   1.1 What are variations? ........................................ 6
   1.2 Optimization and variational problems ................... 7

2 Preliminaries .................................................. 11
   2.1 Formulating variational problems ......................... 11
      2.1.1 Shortest distance between two points ................. 11
      2.1.2 Graph with least surface of revolution ............... 12
      2.1.3 The catenary ........................................... 12
      2.1.4 The brachistochrone .................................... 13
      2.1.5 Dido’s problem ......................................... 14
      2.1.6 Hamilton’s principle .................................... 14
      2.2 Inverse function theorem ($\mathbb{R}^n$ version) ........... 18
      2.3 Implicit function theorem ($\mathbb{R}^n$ version) .......... 22

3 Variational Optimization ...................................... 29
   3.1 The Euler-Lagrange equation ............................... 29
      3.1.1 Formulation ............................................. 29
      3.1.2 Solution to the planar geodesics (Problem A) .......... 32
      3.1.3 Solution to the least surface of revolution (Problem B) ........................................... 33
      3.1.4 Solution to the brachistochrone (Problem D) .......... 33
      3.2 Discontinuities in $\| \cdot \|_\infty$ ............................ 36
      3.3 Integral Constraints ...................................... 37
      3.3.1 Formulation ............................................. 38
      3.3.2 Solution to Dido’s problem (Problem E) ............... 39
      3.3.3 Solution to the catenary (Problem C) ................... 40
      3.4 The Second Variation ...................................... 41
      3.4.1 Review of finite dimensional second derivative test ............ 42
      3.4.2 Formulation ............................................. 42
4 Higher Dimensional Variational Problems

4.1 The Euler-Lagrange equation for trajectories in higher dimensional linear spaces
4.1.1 Formulation
4.1.2 Planar pendulum
4.1.3 The Hamiltonian

4.2 The Euler-Lagrange equation for higher dimensional domains
4.2.1 Formulation
4.2.2 Minimal surface
4.2.3 A soap film
4.2.4 Stable flows
4.2.5 Schrödinger’s equation (unfinished)

5 Analysis on Metric Spaces

5.1 Metric spaces
5.1.1 Basic definitions
5.1.2 Topologies of metric spaces
5.1.3 Completion of metric spaces

5.2 Normed linear spaces
5.2.1 Vector spaces
5.2.2 Normed linear spaces
5.2.3 Linear operators

5.3 Sobolev spaces
5.3.1 Weak derivatives
5.3.2 Properties of weak derivatives and Sobolev spaces
5.3.3 Weak compactness in Sobolev spaces
What is Calculus of Variations?

In 1754, at the age of eighteen, Lagrange read the article “Une méthode pour trouver des lignes courbes jouissant de propriétés de maximum ou de minimum” by the great Euler, concerning the isometric problem, which is the determination of the shape of the closed plane curve having a given length and enclosing the maximum area. Inspired by this lesson, he obtained his first original mathematical result, and dared to communicate it by letter to
Euler, already at the time a leading figure in science. His letter remained unanswered.

Lagrange, however, was undeterred, and continued to reflect on Euler’s article. In 1755 he wrote a second letter to Euler in which he described the new method that he had developed, that is, his own manner for dealing with the problem examined by Euler. That method would be named by Euler himself in one of his letters: “calculus of variations”. This time, Euler replied to Lagrange, in terms of praise in the epigraph above. This letter was enough of a recommendation to secure Lagrange a position as a teacher at the Royal School of Artillery in Turin.

Figure 1.1: Leonhard Euler (1707-1783) and his paper *A method for finding curved lines enjoying properties of maximum or minimum, or solution of isoperimetric problems in the broadest accepted sense* (1743). This work is concerned with the calculus of variations. Euler’s main contribution to this subject was to change it from a discussion of essentially special cases to a discussion of very general classes of problems.

### 1.1 What are variations?

Variations are small changes in functions and functionals.

**Example 1.1 (Derivatives).** Deriving derivatives uses variations. Let $O$ be an open set of $\mathbb{R}^n$ and $F : O \to \mathbb{R}^m$ be a continuous function. We say $F$ is differentiable at a point $x \in O$ with *derivative* (sometimes called the *total derivative*) $L_x$, if $DF_x : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map (so $DF_x$ must have the form of an $n \times m$ matrix) so that

$$
\lim_{h \to 0} \frac{\|F(x+h) - F(x) - DF_x(h)\|}{\|h\|} = 0
$$

(1.1)

**Exercise 1.1.** Show that when $m = 1, n = 1$, then

$$DF_x = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}$$

which is simply $F'(x)$.
Lagrange’s accomplishment was that he was able to find anew the results of Euler while freeing himself from geometric intuition (displacing the graph of the function), and replacing it with a machinery of operations of calculus.

In fact, for a continuously differentiable function $F : O \subset \mathbb{R}^n \to \mathbb{R}^m$, we know how to write down $DF_x$ explicitly:

$$
DF = \begin{pmatrix}
\frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_n} \\
\frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial F_m}{\partial x_1} & \frac{\partial F_m}{\partial x_2} & \cdots & \frac{\partial F_m}{\partial x_n}
\end{pmatrix}
$$

The matrix in (1.2) is called the Jacobian matrix.

**Exercise 1.2.** Prove (1.2). [Hint: For $i = 1, 2, \ldots, n$,

$$
F_i(x + h) - F_i(x) = \sum_{j=1}^{n} \left[ F_i(x + \vec{a}_j) - F(x + \vec{a}_{j-1}) \right]
$$

$$
= \sum_{j=1}^{n} h_j \int_{0}^{1} \frac{\partial F_i}{\partial x_j} \left|_{x + \vec{a}_{j-1} + th_j \vec{e}_j} \right. \ dt
$$

where $h_j$ is the $j$th coordinate of $h$, $\vec{a}_j = (h_1, h_2, \ldots, h_{j}, 0, 0, \ldots) \in \mathbb{R}^n$, and $\vec{e}_j$ has $j$th coordinate equal to 1 and zero otherwise. Then prove that

$$
\lim_{h \to 0} \left\| \sum_{j=1}^{n} h_j \int_{0}^{1} \frac{\partial F_i}{\partial x_j} \left|_{x + \vec{a}_{j-1} + th_j \vec{e}_j} \right. \ dt - \sum_{j=1}^{n} \frac{\partial F_i}{\partial x_j} \left|_x \right. h_j \right\| = 0
$$

**1.2 Optimization and variational problems**

In real-life practice, we often want to maximize or minimize a real-valued function, which lies within the realm of the mathematical optimization. The principle of many optimization problems is to find the points where the
derivative is a 0 matrix; i.e., extreme values can only occur at critical points, meaning the points where all partial derivatives vanish. Later on, we will also discuss how to optimize function with respect to some constrains, and our approach also relies on the derivatives.

**Exercise 1.3** (Unconstrained optimization problem). Optimize the regression line \( y = kx + b \) so that it fits the data points \((1, 2), (2, 3), (4, 2)\) best (so the error function is minimized). The error function is given by \( \epsilon(k, b) = \sum_{i=1}^{n}(Y_i - (kX_i + B))^2 \) where \( \{(X_i, Y_i)\}_{i=1}^{n} \) is the data set.

**Exercise 1.4** (Constrained optimization problem). Maximize utility \( u(x, y) = xy \) subject to the constraint \( c(x, y) = x + 4y = 240 \). Here the price of per unit \( x \) is 1, the price of \( y \) is 4 and the budget available to buy \( x \) and \( y \) is 240.

In calculus of variations, we often use variations to optimize functionals. Roughly speaking, a functional is a real or complex-valued function from a space \( X \), and \( X \) is often a space of functions.

**Example 1.2** (The evaluation functional). Given a set \( X \), consider the space \( \mathbb{R}^{X} \) of all real-valued functions on \( X \). For each \( x \in X \), we can define a functional on \( \mathbb{R}^{X} \)

\[
ev_x(f) := f(x)
\]

This functional is linear, meaning that for all \( c \in \mathbb{R} \) and \( f, g \in \mathbb{R}^{X} \), \( \ev_x(cf + g) = c\ev_x(f) + \ev_x(g) \).

**Example 1.3** (Differentiation). Let \( O \) be an open set in \( \mathbb{R}^n \). Consider the space \( C^1(O) \) be the space of all continuously differentiable real-valued functions on \( O \). For each \( x \in O \) and \( i = 1, 2, \ldots, n \), the following differential operator at \( x \)

\[
D(f) = \left. \frac{\partial}{\partial x_i} f \right|_x
\]

is a linear functional.

**Example 1.4** (Integration as a functional). Consider the space \( C[0, 1] \) of all continuous functions over the interval \([0, 1]\). Since all continuous functions over a closed interval is integrable, we can define a functional on \( C[0, 1] \)

\[
I(f) := \int_{0}^{1} f(x) \, dx
\]

This functional is linear.

**Exercise 1.5.** Find the minimum and the minimizer of the functional defined on \( C[0, 1] \)

\[
L^1(f) := \int_{0}^{1} |f(x)| \, dx
\]

Prove that the minimizer is unique. This functional is not linear.

The following “trivial” optimization problem can be formulated in terms of functionals.

**Problem A** (Shortest distance between two points). What is the shortest distance between two points \( p_0, p_1 \) on the plane \( \mathbb{R}^2 \)?
Suppose we have a path $\gamma(t) = (x(t), y(t))$ where $x, y : [0, 1] \to \mathbb{R}$ are piecewise differentiable functions. Then the “cost” (the quantity to be optimized) of the path in this problem is the arclength of the curve, so we define the cost functional $J$ on the space of all piecewise smooth curves by

$$J(\gamma) = \int_0^1 \sqrt{\dot{x}(t)^2 + \dot{y}(t)^2} \, dt \quad (1.3)$$

Moreover, we have a constraint in this problem, that is, the path must begin at $p_0$ and end at $p_1$; i.e., $(x(0), y(0)) = p_0$ and $(x(1), y(1)) = p_1$. Therefore Problem A can be formulated as to minimize the functional $J$ with respect to the constraint $\gamma(0) = p_0, \gamma(1) = p_1$. This kind of problems are called the variational problems, because the methods of finding their solutions usually involve variations, by perturbing the input function by a small amount.

Here are some classic variational problems that we will discuss and solve:

**Problem B** (Graph with least surface of revolution). Revolve the curve $y = f(x)$ between $(a, c)$ and $(b, c)$ about the y-axis. Which function $f$ yields the least surface area?

**Problem C** (The catenary). Suspend a string of length $\gamma$ between two points. What shape will the string take?

**Problem D** (The brachistochrone (βραχιστοκ χρονος)). What is the shape of a frictionless slide from one place to a lower one that yields the fastest transit time?

Figure 1.3: Johann Bernoulli posed the problem of the brachistochrone to the readers of Acta Eruditorum in June, 1696.
Preliminaries

2.1 Formulating variational problems

In this section, we will formulate several classic variational problems as optimization problems of functionals and we list the constraints whenever applicable.

2.1.1 Shortest distance between two points

Problem A (Shortest distance between two points). *What is the shortest distance between two points* \( p_0, p_1 \) *on the plane* \( \mathbb{R}^2 \)?

By rotating the \( \mathbb{R}^2 \)-plane, without loss of generality we may place one point at the origin \((0, 0)\), the other at \((a, 0)\). Consider any continuously differentiable function \( y = f(x), 0 \leq x \leq a \), with \( f(0) = 0 = f(a) \). Then the cost functional is defined on the space \( C^1[0,a] \) of continuous differentiable function over \([0,a]\) and defined by

\[
J(f) = \int_0^a \sqrt{1 + f'(x)^2} \, dx
\]  

Note \( J(f) \) reaches its minimum when \( f'(x) = 0 \), so \( f \) is a constant function. Consider the constraint that \( f(0) = f(a) = 0 \), so \( f(x) = 0 \) is the required minimizer.
2.1.2 Graph with least surface of revolution

**Problem B** (Graph with least surface of revolution). *Revolve the curve $y = f(x)$ between $(a, c)$ and $(b, d)$ about the $y$-axis. Which function $f$ yields the least surface area?*

![Figure 2.2: A curve $y = f(x)$ is revolved about $y$-axis to produce a surface of least area](image)

The cost functional is defined on the space $C^1[a, b]$ of continuous differentiable function over $[a, b]$ and defined by

$$J(f) = \int_a^b x\sqrt{1 + f'(x)^2} \, dx$$  \hspace{1cm} (2.2)$$

with constraint that $f(a) = c$ and $f(b) = d$.

### 2.1.3 The catenary

**Problem C** (The catenary). *Suspend a string of length $\gamma$ between two points. What shape will the string take?*

Let’s fix the two ends of a string at $(-a, b)$ and $(a, b)$ respectively. The position of a point on the string can be characterized by its distance (the arc length on the string) from one end; call this variable $s$. Then the potential energy $W(s)$ of any point $s$ on the string satisfies that

$$\frac{dW}{ds} = \rho g \cdot y$$

where $\rho$ is the constant density of the string per unit length, and $g$ is the gravity acceleration constant. The shape of the hanging string can be describe by a continuous differentiable function $f$ over $[-a, a]$, as the string fits the curve $y = f(x)$; then $\frac{ds}{dx} = \sqrt{1 + f'(x)^2}$. The cost that the catenary minimizes is the potential energy of the string, which is

$$J(f) = \int_{\text{the string}} \rho g f(x) \cdot \sqrt{1 + f'(x)^2} \, dx, \quad J : C^1[-a, a] \to \mathbb{R}$$  \hspace{1cm} (2.3)$$

12
Additionally the constraints are
\[
\begin{cases}
  f(-a) = f(a) = b & \text{fixed ends at } (-a, b), (a, b) \\
  \int_{-a}^{a} \sqrt{1 + f'(x)^2} \, dx = \gamma & \text{fixed string length } \gamma
\end{cases}
\] (2.4)

Variational problems that have constraints as a constant value for the line integral of the variable function, like the second equation of (2.4), are called *isoperimetric problems*. Isoperimetric problems are the first variational problems that Euler and Lagrange set off to solve.

### 2.1.4 The brachistochrone

**Problem D** (The brachistochrone). *What is the shape of a frictionless slide from one place to a lower one that yields the fastest transit time?*

The kinetic energy \( \frac{mv^2}{2} \) of sliding body of mass \( m \) at the moment of velocity \( v \) must equal the potential energy lost from the initial position. Suppose a slide is built from \((0, 0)\) to \((b, d)\) and the slide fits the curve \( y = f(x) \). Then

\[
\frac{mv^2}{2} = mg \cdot f(x), \quad v = \sqrt{2g \cdot f(x)}
\]

Let \( s(x) \) be the distance (the arc length on the slide) of the object from the starting position \( (0, 0) \), when its horizontal coordinate is \( x \). Then \( s(x) = \int_{0}^{x} \sqrt{1 + f'(t)^2} \, dt \), \( \frac{ds}{dx} = \sqrt{1 + f'(x)^2} \). Hence the cost to minimize is the total time needed to slide down to the terminal position \((b, d)\):

\[
J(f) = \text{time} = \int \frac{d(\text{distance})}{\text{velocity}} = \int_{0}^{b} \frac{dx}{\sqrt{1 + f'(x)^2}} = \int_{0}^{b} \sqrt{1 + f'(x)^2} \, dx \quad J : C^{1}[0, b] \to \mathbb{R} \] (2.5)

The constraints are given by the two ends of the slide: \( f(0) = 0 \) and \( f(b) = d \).
Figure 2.4: A slide is built from $(0, 0)$ to $(b, d)$; we want to find the slide shape that yield the least travel time for a sliding body.

### 2.1.5 Dido’s problem

According to myth, after her husband was killed by her brother Pygmalion, Queen Dido of Carthage fled Tyre with her retinue (and a sizable portion of the Tyre treasury) to Mediterranean Africa. There she purchased from a naive king all the land that could be enclosed by the hide of an ox. After slicing the hide into thin strips and tying end to end, she enclosed a sizable parcel that became the city-state of Carthage, *circa* 853 BC. Virgil in his *Aeneid* translates the myth 300 years back in time to Fall of Troy in order to interweave her story with that of Aeneas.

**Problem E** (Dido’s problem). Granted a portion of a coastline of Africa as border, what is the largest country that can be enclosed by a given remaining perimeter?

Suppose the coastline is the line segment $[a, b]$ on the $x$-axis and the interior of the country is bounded by the curve $y = f(x)$ with $f(a) = f(b) = 0$.

Then the cost (or we might want to call it gain here) to maximize is the area enclosed by $y = f(x)$ and $x$-axis:

$$J(f) = \int_a^b f(x) \, dx, \quad J : C^1[a, b] \to \mathbb{R}$$

(2.6)

The constraint is given by the fixed remaining perimeter beside the coastline that

$$\int_a^b \sqrt{1 + f'(x)^2} \, dx = \gamma$$

(2.7)

### 2.1.6 Hamilton’s principle

Hamilton’s principle is one of the most important insights of science. We will think of it as the fundamental axiom of mechanical systems. The principle predicts the nonrelativistic motion of mechanical systems from galaxies to billiard balls. We formulate Hamilton’s principle in a slightly imprecise manner, but later on the program we will discuss it again rigorously.
Figure 2.5: Dido Purchases Land for the Foundation of Carthage, Matthäus Merian the Elder, 1630. Suppose the coastline is the line segment \([a, b]\) on the \(x\)-axis and the land is on the upper plane

**A brief overview of normed linear spaces**

For a vector space \(V\), the pair \((V, \| \| : V \times V \to \mathbb{R}_{\geq 0})\) is called a **normed vector space** if

1. *(Triangle inequality)* \(\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|\) for all \(v_1, v_2 \in V\).

2. *(Positive homogeneity)* \(c\|v\| = \|cv\|\) for all \(v \in V\) and scalars \(c \in \mathbb{R}\).

3. *(Positive definiteness)* If \(\|v\| = 0\), then \(v = 0\).

An elementary example of a normed linear space is \(\mathbb{R}^n\) with the Euclidean norm \(\| \|_2\):

\[
\|(a_1, a_2, \ldots, a_n)\|_2 = \sqrt{\sum_{k=1}^{n} a_k^2}
\]

For our purpose, we can think of \(V\) as the space \(C^1[a, b]\) of continuously differentiable functions over the compact interval \([a, b]\) with the norm \(\| \|_{C^1}\):

\[
\|f\|_{C^1, \infty} = \sup_{x \in [a, b]} |f(x)| + \sup_{x \in [a, b]} |f'(x)|, \quad f \in C^1[a, b]
\] (2.8)

Other examples of norms and normed vector spaces will be introduced later. For normed linear spaces \((V, \| \|_V), (W, \| \|_W)\), a mapping \(A : V \to W\) is called a **bounded linear operator** if

1. *(Linearity)* \(A(cv_1 + v_2) = cAv_1 + Av_2\) for all \(v_1, v_2 \in V\) and \(c \in \mathbb{R}\).

2. *(Boundedness)* There exists some \(M\) so that \(\frac{\|Av\|_W}{\|v\|_V} < M\) for all \(v \in V\) with \(v \neq 0\).

**Exercise 2.1.** Show that the mapping \(I, J\) defined on \((C^1[a, b], \| \|_{C^1})\) by

\[
I(f) = \int_{a}^{b} f(x) \, dx, \quad J(f) = \int_{a}^{b} [f(x) + f'(x)] \, dx
\]

are bounded linear operators.
Exercise 2.2. Prove that every linear transformation $M : \mathbb{R}^n \to \mathbb{R}^m$ is bounded. [Hint: One way is to use Cauchy-Schwartz inequality and some linear algebra, for all $x \in \mathbb{R}^n, \|Mx\|^2 = \langle Mx, Mx \rangle = \langle MTMx, x \rangle \leq \|MTMx\|\|x\| \leq \sigma^2\|x\|^2$ where $\sigma$ is the first singular value of $M$; or a less accurate estimate of the upper bound of $\|Mv\|/\|v\|$ by simple calculations is that $\|Mx\|^2 = \sum_{i=1}^n \left( \sum_{j=1}^m a_{ij}^2 \right) \left( \sum_{j=1}^n x_j^2 \right) = \sum_{i,j=1}^{n,m} a_{ij}^2 \|x\|^2$.]

Exercise 2.3. For any normed linear spaces $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$, prove that the collection $\mathcal{B}(V,W)$ of all bounded linear operators from $V$ to $W$ together with the norm $\|\cdot\|_{\mathcal{B}}$ defined by

$$\|A\|_{\mathcal{B}} = \sup_{v \in V, v \neq 0} \frac{\|Av\|_W}{\|v\|_V}, \quad A \in \mathcal{B}(V,W)$$

forms a normed linear space. Furthermore, prove that if $A$ is invertible, then $\|A^{-1}\|_{\mathcal{B}}^{-1} = \inf_{v \in V} \frac{\|Av\|_W}{\|v\|_V}$

A functional $J : V \to W$ is called differentiable or Fréchet differentiable at $f \in V$ if there exists a bounded linear operator $A : V \to W$ so that

$$\lim_{\|h\|_V \to 0} \frac{\|V(f + h) - V(f) - Ah\|_W}{\|h\|_V} = 0$$

(2.10)

$A$ is called the Fréchet derivative of $J$ at $f$, and is often denoted by $D(J)f$. Fréchet derivative is the generalization of derivatives of differentiable function between finite dimensional vector spaces. However due to the abstract nature of the underlying spaces, we shall not expect to write down explicit formulae for Fréchet derivatives in general, but we can take advantage of their analytical properties.

Hamilton’s principle

**Axiom** (Hamilton’s principle). A conservative mechanical system, as it evolves during the time from $t_1$ to $t_2$, will choose a trajectory $q(t)$ in $C^1[t_1,t_2]$ so that

$$D(J)_q = 0, \quad \text{where} \quad J(q) = \int_{t_1}^{t_2} L(q(t)) \, dt$$

(2.11)

and the Lagrangian $L(q(t))$ is the system’s kinetic energy less its potential energy at time $t$.

The value $J(q)$ is called the action of the trajectory $q$. Hamilton’s principle essentially says that a system in motion must take the trajectory with the least action. It turns out Hamilton’s principle is equivalent to Newton’s laws of motion, but we may find this axiom more efficient in formulating physics problems in functionals.

**Theorem 2.1** (Newton’s laws of motion). Hamilton’s principle holds if and only all of the following statements are true:

1. In an inertial reference frame, unless acted upon by a force, a body remains at rest or moving at a constant velocity.

2. In an inertial reference frame, the vector sum of all forces $F$ acting on a body is equal to the change in the body’s momentum in time; i.e.,

$$D\vec{F}_t = mD\vec{v}_t, \quad \text{for all} \ t$$
If a body \( i \) exerts a force \( \vec{F}_{ij} \) on a body \( j \), the body \( j \) simultaneously exerts a force \( \vec{F}_{ji} = -\vec{F}_{ij} \) equal in magnitude and opposite in sign on the first body.

Suppose \( D(J)_q = 0 \). Let \( f \) be any non-zero trajectory with \( f(t_1) = f(t_2) = 0 \). By adding \( \varepsilon \cdot f \) for real numbers \( \varepsilon \) with small absolute values, we can perturb \( J \) by small amounts. Define \( \ell(\varepsilon) := J(q + \varepsilon f) \). Note that since \( D(J)_q = 0 \), by (2.10)

\[
\lim_{\varepsilon \to 0} \frac{\|J(q + \varepsilon f) - J(q) - D(J)_q(\varepsilon f)\|}{\|\varepsilon f\|} = 0
\]

\[
\Rightarrow \ell'(0) = \lim_{\varepsilon \to 0} \frac{|\ell(\varepsilon) - \ell(0)|}{|\varepsilon|} \cdot \frac{1}{\|f\|} = 0
\]

(2.12)

For any fixed direction \( f \) of perturbation, (2.12) simplifies finding stationary trajectory in Hamilton’s principle into a single-variable calculus problem. However the dependence on the directions of perturbation will be treated carefully.

**Problem F** (Vibration of a spring-mass system). What is the trajectory of a mass sliding along a horizontal frictionless rail restrained by an ideal spring?

Suppose a mass \( m \) is free to slide along a horizontal frictionless rail restrained by an ideal spring of spring constant \( k \). Hooke’s law states that the restoring force \( F \) of the spring is proportional to the displacement \( q \) \((q(t)\)

![Figure 2.6: A spring-mass system](image)

is also the trajectory here) from its natural length:

\[ F = -kq \]

Hence if the mass is displaced \( \tilde{q} \) from rest, the work done must be

\[ V = -\int_0^{\tilde{q}} F dq = \int_0^{\tilde{q}} kq dq = \frac{k\tilde{q}^2}{2} \]

Note that the kinetic energy at time \( t \) is given by \( T = \frac{m(q'(t))^2}{2} \). Therefore the spring-mass system must move by Hamilton’s principle in the stationary trajectory \( q \) of the functional

\[ J(q) = \int_{t_1}^{t_2} L dt = \int_{t_1}^{t_2} \left[ \frac{m(q'(t))^2}{2} - \frac{kq(t)^2}{2} \right] dt \]

I.e., \( D(J)_q = 0 \).
Problem G (The planar pendulum). What is the trajectory of a bob connected by a massless rigid rod to a point of pivot?

Suppose a bob of mass $m$ is connected by a massless rigid rod of length $a$ to a point of pivot which constrains the resulting pendulum’s motion to one plane. The trajectory of the system can be described by the angle $\theta(t)$ from the vertical direction. The kinetic energy of the bob at time $t$ is given by

$$T = \frac{mv^2}{2} = \frac{a^2 m \cdot \theta'(t)^2}{2}$$

whereas the potential energy at time $t$ is given by

$$V = \text{weight} \times \text{height} = mga(1 - \cos(\theta(t)))$$

Therefore the planar pendulum must move by Hamilton’s principle in the stationary trajectory $\theta$ of the functional

$$J(\theta) = \int_{t_1}^{t_2} L \, dt = \int_{t_1}^{t_2} \left[ \frac{a^2 m \cdot \theta'(t)^2}{2} - mga(1 - \cos(\theta(t))) \right] \, dt$$

2.2 Inverse function theorem ($\mathbb{R}^n$ version)

The inverse function theorem and its consequences are ones of the most important results in mathematical analysis. They laid some analytical foundations for differential geometry and optimization theory. Also, inverse function theorem and implicit function theorem generalize nicely to complete normed linear spaces, which we will cover later in the program. (The proofs are almost the same.) In this meeting, we will focus on the $\mathbb{R}^n$ versions of the theorems. We will frequently make use of them in solving variational problems.

Exercise 2.4. Let $f$ be a real-valued continuously differentiable function on $\mathbb{R}$ and $a \in \mathbb{R}$. Suppose $f(a) \neq 0$.

1. Prove that there exists some $\delta > 0$ so that the restriction $f|_{(a-\delta, a+\delta)}$ is injective. [Hint: By the continuity of $f'$ at 0, there exists some $\delta > 0$ so that $f'(x) \neq 0$ for all $x \in (a-\delta, a+\delta)$ and then use the mean value theorem to prove $f$ is injective on $(a-\delta, a+\delta)$.]
2. For an injective function, we can define its inverse from its range to its domain. Prove that the inverse function of \( f\bigr|_{(a-\delta,a+\delta)} \) is differentiable with derivative \( \frac{1}{f'(b)} \). [Hint: You will need to show the inverse function of \( f\bigr|_{(a-\delta,a+\delta)} \) is continuous first; it will be used to prove the differentiability.]

The core of the inverse function theorem is to extract local information about the function from its derivative. The derivative can be considered as the local linearization of the function. Essentially the theorem says that if the derivative is invertible, then the function is locally invertible (there exists an open neighborhood of the point where the function is invertible) and its local inverse preserves its smoothness.

**Theorem 2.2** (Inverse function theorem for \( \mathbb{R}^n \)). Let \( U \) be an open set in \( \mathbb{R}^n \), \( x_0 \in U \), and \( F : U \to \mathbb{R}^n \) be a continuous differentiable (\( C^1 \)) map. If \( DF_{x_0} : \mathbb{R}^n \to \mathbb{R}^n \) is invertible, then there exist open sets \( V \) and \( W \) containing \( x_0 \) and \( F(x_0) \) respectively such that the restriction \( F|_V : V \to W \) is a bijection and \( F|_V^{-1} : W \to V \) is also \( C^1 \). Moreover, if \( F \) is \( k \)-time continuously differentiable (\( C^k \)) at \( x_0 \), then its local inverse is also \( C^k \).

**Remark.** Under the assumptions of the theorem, by the chain rule,

\[
D(F|_V^{-1} \circ F) = I \quad \Rightarrow \quad D(F|_V^{-1})F(x_0)D(F)_{x_0} = I \quad \Rightarrow \quad D(F|_V^{-1})F(x_0) = (D(F)_{x_0})^{-1} \tag{2.13}
\]

**Proof.** Without the loss of generality, with shifting by constants we may assume \( x_0 = F(x_0) = 0 \), for otherwise we could prove for \( G(x) := F(x + x_0) - F(x_0) \) instead. To prove the local invertibility of \( F \) at 0, we need to show that for \( y \) sufficiently close to 0, the equation \( F(x) = y \) has a unique solution on a neighborhood of 0. Unfortunately unlike in \( \mathbb{R} \), we don’t have the **mean value theorem**, so we take another approach. Recall the proof of the contraction mapping theorem from Math 165a, and we will use some contractive iteration to generate a unique solution. By continuity of the first derivative, there exists some neighborhood \( U \) of 0 so that \( DF_x \) is invertible for all \( x \in U \). Now fix \( y \), and define \( T : U \to \mathbb{R}^n \) by

\[
T(x) := DF_0^{-1}(DF_0(x) - F(x) + y) \tag{2.14}
\]

If \( \bar{x} \) is a fixed point of \( T \), then \( F(\bar{x}) = y \). Now we need to show that for \( y \) sufficiently close to 0, the corresponding map \( T \) is a contraction that can be applied iteratively on an open neighborhood of 0; i.e.,

- There exists some open set \( V \subset U \) so that \( T(\overline{V}) \subset \overline{V} \), where \( \overline{V} \) is the closure of \( V \).
- There is some \( 0 \leq p < 1 \) so that for all \( x_1, x_2 \in \overline{V} \), \( \|T(x_1) - T(x_2)\| \leq p\|x_1 - x_2\| \)

Note that for \( x \in U \),

\[
\|Tx\| = \left\| DF_0^{-1}(DF_0(x) - F(x) + y) \right\|
\leq \left\| DF_0^{-1} \right\|_{\text{op}} \left\| DF_0(x) - F(x) + y \right\|
\leq \left\| DF_0^{-1} \right\|_{\text{op}} \left( \left\| DF_0(x) - F(x) \right\| + \|y\| \right)
= \left\| DF_0^{-1} \right\|_{\text{op}} \left( \left\| \int_0^1 \frac{d}{dt} F(tx) \, dt - DF_0(x) \right\| + \|y\| \right)
= \left\| DF_0^{-1} \right\|_{\text{op}} \left( \left\| \int_0^1 (DF_{tx} - DF_0) \, dt \right\| + \|y\| \right)
\]
Since $DF$ is continuous at 0, there exists some $r_1 > 0$ so that

$$\|DF_x - DF_0\|_B \leq \frac{1}{2}\|DF_0^{-1}\|_B, \quad \text{for all } \|x\| \leq r_1 \tag{2.15}$$

Take $r_2 = \frac{r_1}{2\|DF_0^{-1}\|_B}$. Then for all $y \in B(0, r_2) := \{z \in \mathbb{R}^n : \|z - 0\| < r_2\}$, and for all $x \in B(0, r_1) := \{z \in \mathbb{R}^n : \|z - 0\| \leq r_1\}$,

$$\|T(x)\| \leq \|DF_0^{-1}\|_B \left( \left( \int_0^1 \|DF_{tx} - DF_0\|_B \, dt \right) \cdot \|x\| + \|y\| \right) \leq \frac{1}{2}\|DF_0^{-1}\|_B \left( \frac{1}{2}\|DF_0^{-1}\|_B \cdot \|x\| + \|y\| \right) < \frac{r_1}{2} + \frac{r_1}{2} = r_1$$

In conclusion, this paragraph showed that for all $y \in B(0, r_2)$, the corresponding $T$ maps $\overline{B}(0, r_1)$ into itself, i.e., $T(\overline{B}(0, r_1)) \subset \overline{B}(0, r_1)$.

To show $T$ is a contraction on $\overline{B}(0, r_1)$, let $x_1, x_2 \in \overline{B}(0, r_1)$.

$$\|T(x_1) - T(x_2)\| = \left\| DF_0^{-1}(DF_0(x_1) - F(x_1) + y) - DF_0^{-1}(DF_0(x_2) - F(x_2) + y) \right\| \leq \|DF_0^{-1}\|_B \|F(x_2) - F(x_1) + DF_0(x_1 - x_2)\| \leq \|DF_0^{-1}\|_B \left( \int_0^1 \frac{d}{dt}F(x_1 + t(x_2 - x_1)) \, dt + DF_0(x_1 - x_2) \right) \leq \|DF_0^{-1}\|_B \left( \int_0^1 DF_{x_1 + t(x_2 - x_1)} - DF_0 \, dt \right) \|x_2 - x_1\| \leq \frac{1}{2}\|DF_0^{-1}\|_B \left( \frac{1}{2}\|DF_0^{-1}\|_B \cdot \|x_2 - x_1\| \right) = \frac{1}{4}\|x_1 - x_2\| \tag{2.16}$$

So far we have shown that for all $y \in B(0, r_2)$, $T : \overline{B}(0, r_1) \rightarrow \overline{B}(0, r_1)$ is a contraction. Then by the contraction mapping theorem, there exists a unique fixed point of $T$ in $\overline{B}(0, r_1)$. Hence for all $y \in B(0, r_2)$, there is exactly one point $x \in \overline{B}(0, r_1)$ solving the equation $F(x) = y$; this allows us to define a map $G : B(0, r_2) \rightarrow \overline{B}(0, r_1)$ by mapping $y = F(x)$ back to $x$.

$$F|_{F^{-1}B(0, r_2) \cap \overline{B}(0, r_1)} : F^{-1}B(0, r_2) \cap \overline{B}(0, r_1) \rightarrow B(0, r_2) \tag{2.18}$$

is a continuous bijection and its inverse function is $G$.

To get a bijective mapping from an open neighborhood of 0, we need to exclude the points on the boundary $\partial\overline{B}(0, r_1) = \{z \in \mathbb{R}^n : \|z - 0\| = r_1\}$. Note that $\partial\overline{B}(0, r_1)$ is compact, so the set $\|F(\partial\overline{B}(0, r_1))\|$ is also compact, and hence $\|F(\partial\overline{B}(0, r_1))\|$ has a minimum, say $r_3$. Note that we have proved that there exists a unique point $x_0$ in $\overline{B}(0, r_1)$ so that $F(x_0) = 0$; that unique point $x_0$ is 0. Meanwhile note that 0 $\not\in \partial\overline{B}(0, r_1)$, hence $F(x) \neq 0$ for all $x \in \partial\overline{B}(0, r_1)$ and therefore $r_3 > 0$. Then for all $x \in \partial\overline{B}(0, r_1)$, $\|F(x)\| \geq r_3 > 0$; i.e., for a $y \in B(0, r_2)$, if
\[ \|y\| < r_3 \] then the corresponding unique point \( x \in \overline{B}(0, r_1) \) with \( F(x) = y \) cannot be on the boundary \( \partial \overline{B}(0, r_1) \). Take \( r_4 = \min\{r_2, r_3\} \). Then \( F^{-1}B(0, r_4) \cap \overline{B}(0, r_1) \) \( \subset \) \( B(0, r_1) \), so \( F^{-1}B(0, r_4) \cap \overline{B}(0, r_1) = F^{-1}B(0, r_4) \cap B(0, r_1) \) is open in \( \mathbb{R}^n \). In conclusion,

\[ F|_{F^{-1}B(0, r_4) \cap \overline{B}(0, r_1)} : F^{-1}B(0, r_4) \cap \overline{B}(0, r_1) \rightarrow B(0, r_4) \quad (2.19) \]

is a continuous bijection between open sets in \( \mathbb{R}^n \).

We have proved the local invertibility. It remains to show that the local inverse still satisfies the \( C^1 \)-smoothness. Like the proof in \( \mathbb{R} \) (Exercise 2.4), this will likely use the continuity of \( G \), so we show it first. Let \( y_1, y_2 \in B(0, r_2) \). Then there exist unique \( x_1, x_2 \in \overline{B}(0, r_1) \) so that \( F(x_1) = y_1, F(x_2) = y_2 \). Then

\[
\| G(x_1) - G(x_2) \| = \| x_1 - x_2 \| \\
= \| D\mathcal{F}_F^{-1} \| \| (DF_0(x_1) - F(x_1) + y_1) - (DF_0(x_2) - F(x_2) + y_2) \| \\
\leq \| D\mathcal{F}_F^{-1} \|_B \| (DF_0(x_1 - x_2) - (F(x_1) - F(x_2)) + (y_1 - y_2)) \| \\
\leq \frac{1}{2} \| G(y_1) - G(y_2) \| + \| D\mathcal{F}_F^{-1} \|_B \| y_1 - y_2 \|
\]

so \( G \) is continuous on \( B(0, r_2) \).

Finally we prove the differentiability of \( G \) on \( B(0, r_2) \), and by (2.13) in Remark the higher-order \( C^k \)-smoothness follow directly from the matrix inversion formula. Take any \( y \in B(0, r_2) \) and take \( h \in \mathbb{R}^n \) so that \( y + h \in B(0, r_2) \). By (2.13) we know if \( G \) is differentiable at \( y \), then \( DG_y = (DF_{G(y)})^{-1} \), so we want to prove that

\[
\| G(y + h) - G(y) - (DF_{G(y)})^{-1}h \| \rightarrow 0 \quad (h \rightarrow 0)
\]

\[
\begin{align*}
h &= F(G(y + h)) - F(G(y)) \\
&= \left( \int_0^1 D\mathcal{F}_{G(y) + t(G(y + h) - G(y))} \, dt \right) (G(y + h) - G(y))
\end{align*}
\]

so

\[
G(y + h) - G(y) - (DF_{G(y)})^{-1}h = (DF_{G(y)})^{-1} \int_0^1 \left( D\mathcal{F}_{G(y) + t(G(y + h) - G(y))} \right) dt(G(y + h) - G(y))
\]

(2.23)

Since \( G \) and \( DF \) are continuous, for all \( \varepsilon > 0 \), there exists \( \delta > 0 \) so that

\[
\| D\mathcal{F}_{G(y_1)} - D\mathcal{F}_{G(y_2)} \|_B \leq \varepsilon \quad \text{whenever} \; \|y_1 - y_2\| < \delta
\]

(2.24)
and consequently for \( \|h\| < \delta \),

\[
(DF_{G(y)})^{-1} \int_{0}^{1} \left( DF_{G(y)}(y+h)-DF_{G(y)}(y+h)-G(y) \right) dt
\]

\[
\leq \|DF_{G(y)}(y+h)-DF_{G(y)}(y)\|_{B} dt
\]

\[
\leq \varepsilon \cdot \|DF_{G(y)}(y+h)-DF_{G(y)}(y)\|_{B}
\]

Hence for any arbitrarily small \( \varepsilon > 0 \), there is some \( \delta > 0 \) so that for all \( \|h\| < \delta \),

\[
\frac{\|G(y+h)-G(y)-(DF_{G(y)})^{-1}h\|}{\|h\|} \leq \frac{\varepsilon \cdot \|DF_{G(y)}(y+h)-DF_{G(y)}(y)\|_{B}}{\|h\|}
\]

\[
\leq \varepsilon \cdot \|DF_{G(y)}(y+h)-DF_{G(y)}(y)\|_{B} \left( 2 \|DF_{0}^{-1}\|_{B} \right)
\]

i.e., \( \lim_{\|h\| \to 0} \frac{\|G(y+h)-G(y)-(DF_{G(y)})^{-1}h\|}{\|h\|} = 0 \), so \( G \) is differentiable at \( y \) with \( DG_{y} = (DF_{G(y)})^{-1} \). For smoothness, note that if \( F \) is \( C^{k} \) on \( F^{-1}B(0, r_{2}) \cap \overline{B}(0, r_{1}) \) then the entries of \( DF_{G(y)} \) are \( C^{k} \) in the variable \( y \), and because the entries of inverse matrix \( A \) can be written as rational functions in entries of \( A \) by the matrix inversion formula, the entries of \( DG_{y} = (DF_{G(y)})^{-1} \) are rational combinations of the entries of \( DF_{G(y)} \), so the \( C^{k} \)-smoothness is indeed preserved.

To conclude, given that \( F \) is \( k \)-time continuously differentiable at \( 0 \) with \( F'(0) \) invertible, we have proved that there are \( r_{1}, r_{2} > 0 \) so that

\[
F|_{F^{-1}B(0, r_{4}) \cap \overline{B}(0, r_{1})} : F^{-1}B(0, r_{4}) \cap \overline{B}(0, r_{1}) \to B(0, r_{4})
\]

is invertible and its inverse function is also \( C^{k} \). \( \square \)

**Exercise 2.5.** Show that there exists an open neighborhood \( U \) of \( (0, 0) \) so that for all \( (y_{1}, y_{2}) \in U \) the system of equations

\[
\begin{cases}
    x_{1} + x_{1}^{3} - x_{2}^{2} = y_{1} \\
    x_{1}^{2} - x_{2} + x_{2}^{2} = y_{2}
\end{cases}
\]

has a unique solution \( (x_{1}, x_{2}) \). Challenge: Find such an open set \( U \).

### 2.3 Implicit function theorem (\( \mathbb{R}^{n} \) version)

In calculus, we have seen the technique of implicit differentiation. Consider the following solution that uses the implicit differentiation.

**Problem 2.1.** Find the tangent line to the curve \( C = \{ (x, y) : x^{3} \sin y + y = 4x + 3 \} \) at the point \( (-\frac{3}{4}, 0) \).

**Solution.** We first solve \( \frac{dy}{dx} \) by implicit differentiation. Differentiate the curve equation with respect to the variable
$x$ on both sides to get

$$
\frac{d}{dx} \left( x^3 \sin y + y \right) = \frac{d}{dx} (4x + 3)
$$

$$
3x^2 \cdot \sin(y) + x^3 \cdot \left( \frac{dy}{dx} \cdot \cos(y) \right) + \frac{dy}{dx} = 4
$$

$$
\frac{dy}{dx} = \frac{4 - 3x^2 \cdot \sin(y)}{x^3 \cdot \cos(y) + 1}
$$

$$
\left. \frac{dy}{dx} \right|_{x=\frac{3}{4}, y=0} = \frac{4 - 3 \cdot \left( \frac{3}{4} \right)^2 \cdot \sin(0)}{\left( \frac{3}{4} \right)^3 \cdot \cos(0) + 1} = \frac{4}{\left( \frac{3}{4} \right)^3 + 1}
$$

Hence the tangent line to \( \{(x, y) : x^3 \sin y + y = 4x + 3\} \) at the point \( (-\frac{3}{4}, 0) \) is

$$
y = \frac{4}{\left( \frac{3}{4} \right)^3 + 1} \left( x + \frac{3}{4} \right).
$$

In this solution by “implicit differentiation”, what actually happened is that we assumed near the point \( (-\frac{3}{4}, 0) \) the set \( C \) can be locally written as \( y = f(x) \) for some differentiable function \( f \); precisely, there exist an open neighborhood \( U \) of \( (-\frac{3}{4}, 0) \) in \( \mathbb{R}^2 \), some \( \varepsilon > 0 \) and a differentiable function \( f : (\frac{3}{4} - \varepsilon, \frac{3}{4} + \varepsilon) \to \mathbb{R} \) so that

$$
U \cap C = \left\{ (x, f(x)) : x \in \left( \frac{3}{4} - \varepsilon, \frac{3}{4} + \varepsilon \right) \right\}
$$

When we write the computations in terms of the function \( f \),

![Figure 2.8: Near \((-\frac{3}{4}, 0)\), \( C \) can be expressed as \( y = f(x) \) for some smooth function \( f \)](image-url)
\[ x^3 \sin(f(x)) + f(x) = 4x + 3 \]
\[ \frac{d}{dx} [x^3 \sin(f(x)) + f(x)] = \frac{d}{dx} [4x + 3] \]

It becomes clear why our familiar differentiation rules apply, which teachers sometimes skip when teaching students in a first course in calculus. This is a quite strong assumption; for sets defined by certain equations, not only the sets are locally smooth curves, but also the curves can locally be written in such a nice form \((x, f(x))\) with \(f\) differentiable.

But what kind of equations can locally yield such nice curves? Consider a \(C^k\)-function
\[ F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m, \quad (x, y) \mapsto F(x, y), x \in \mathbb{R}^n, y \in \mathbb{R}^m \]
and \(F(x_0, y_0) = 0\) for some \(x_0 \in \mathbb{R}^n, y_0 \in \mathbb{R}^m\). In Problem 2.1, \(F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}, F(x, y) = x^3 \sin y + y - (4x + 3)\) and \(F \left(-\frac{3}{4}, 0\right) = 0\). The implicit function theorem gives a sufficient condition, depending on \(DF(x_0, y_0)\), of the existence of a \(C^k\)-function \(f : U \subset \mathbb{R}^n \to \mathbb{R}^m\) for some open neighborhood \(U\) of \(x_0\) so that
\[ F(x, f(x)) = 0, \quad \text{for all } x \in U \]
and \(f\) is unique up to restrictions of domain.

Before we state the implicit function theorem, let’s review the notation of block matrices. A block matrix is a matrix that is defined using smaller matrices (blocks). For example, we can write
\[
M = \begin{pmatrix}
0 & 1 & 2 & 3 \\
4 & 5 & 6 & 7 \\
8 & 9 & 10 & 11 \\
12 & 13 & 14 & 15 \\
16 & 17 & 18 & 19
\end{pmatrix} = \begin{pmatrix}
A & B \\
C & D
\end{pmatrix}
\]
where
\[
A = \begin{pmatrix}
0 & 1 & 2 \\
4 & 5 & 6 \\
8 & 9 & 10
\end{pmatrix}, \quad B = \begin{pmatrix}
3 \\
7 \\
11
\end{pmatrix}, \quad C = \begin{pmatrix}
12 & 13 & 14 \\
16 & 17 & 18
\end{pmatrix}, \quad D = \begin{pmatrix}
15 \\
19
\end{pmatrix}
\]
The addition and multiplication of block matrices can be carried out by treating the blocks as entries, but beware that unlike real or complex numbers, matrix multiplication is not commutative.

The construction of block matrices can be understood as the decomposition into the linear transformations from linear subspaces to quotient spaces. In particular, suppose \(V, U\) are finite dimensional vector spaces with bases \(\{v_1, v_2, \ldots, v_n\}\) and \(\{u_1, u_2, \ldots, u_m\}\) respectively. Then any linear transformation \(T : V \to U\) has a \(n \times m\)-matrix representation
\[
M = \begin{pmatrix}
Tv_1 & Tv_2 & \cdots & Tv_n
\end{pmatrix}
\]
We can decompose \(V, U\) into direct sum of their linear subspaces
\[
V = \text{Span}\{v_1, v_2, \ldots, v_k\} \oplus \text{Span}\{v_{k+1}, v_{k+2}, \ldots, v_n\} =: V_1 \oplus V_2 \subseteq V
\]
\[
U = \text{Span}\{u_1, u_2, \ldots, u_l\} \oplus \text{Span}\{u_{l+1}, u_{l+2}, \ldots, u_j\} \oplus \text{Span}\{u_{j+1}, u_{j+2}, \ldots, u_m\} =: U_1 \oplus U_2 \oplus U_3 \subseteq U
\]
Then the blocks of $M$ are exactly the linear transformations between subspaces

$$M = \begin{pmatrix} T/(U_2 \oplus U_3)|_{V_1} : V_1 \to U/(U_2 \oplus U_3) \simeq U_1 & T/(U_2 \oplus U_3)|_{V_2} : V_2 \to U/(U_2 \oplus U_3) \simeq U_1 \\ T/(U_1 \oplus U_3)|_{V_1} : V_1 \to U/(U_1 \oplus U_3) \simeq U_2 & T/(U_1 \oplus U_3)|_{V_2} : V_2 \to U/(U_1 \oplus U_3) \simeq U_2 \\ T/(U_1 \oplus U_2)|_{V_1} : V_1 \to U/(U_1 \oplus U_2) \simeq U_3 & T/(U_1 \oplus U_2)|_{V_2} : V_2 \to U/(U_1 \oplus U_2) \simeq U_3 \end{pmatrix}$$

Let $F(x, y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$ be a differentiable function. For the Jacobian matrix of $DF_z : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$ at $z \in \mathbb{R}^n \times \mathbb{R}^m$, we denote its blocks by

$$DF = \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \end{pmatrix}$$

Now we can state and prove the implicit function theorem. It is a fairly quick consequence of the inverse function theorem.

**Theorem 2.3 (Implicit function theorem for $\mathbb{R}^n$).** Let $U$ be an open set in $\mathbb{R}^n \times \mathbb{R}^m$, $(x_0, y_0) \in U$, and $F(x, y) : U \to \mathbb{R}^m$ be a continuously differentiable ($C^1$) map. If $\frac{\partial F}{\partial y}(x_0, y_0) : \mathbb{R}^m \to \mathbb{R}^m$ is invertible, then there exist an open set $V_1 \times V_2$ in $U$ containing $(x_0, y_0)$ and a unique $C^1$-function $\varphi : V_1 \to V_2$ such that $\varphi(x_0) = y_0$ and

$$F(x, \varphi(x)) = F(x_0, y_0), \quad \text{for all } x \in V_1$$

$\varphi$ is unique in the sense that the local level set of $F(x_0, y_0)$ is exactly the graph of $\varphi$; i.e.,

$$\{(x, y) : F(x, y) = F(x_0, y_0)\} \cap V_1 \times V_2 = \{(x, \varphi(y)) : x \in V_1\}$$

Moreover, if $F$ is $C^k$, then $\varphi$ is also $C^k$.

**Proof.** Without the loss of generality, if you try to prove the theorem in the future, you could assume $(x_0, y_0) = (0, 0)$ and $F(x_0, y_0) = 0$, for otherwise we could prove for $G(x) := F(x + x_0, y + y_0) - F(x_0, y_0)$ instead. For a clearer indication of coordinates, I will not do so. Define

$$L : U \to \mathbb{R}^n \times \mathbb{R}^m, \quad (x, y) \mapsto (x, F(x, y)) \quad (2.25)$$

Then

$$DL = \begin{pmatrix} I_{n \times n} & 0 \\ \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \end{pmatrix}$$

Note that because $\left.\frac{\partial F}{\partial y}\right|_{(x_0, y_0)}$ is invertible, $DL$ is invertible at $(x_0, y_0)$ with

$$\left(DL_{(x_0, y_0)}\right)^{-1} = \begin{pmatrix} I_{n \times n} & 0 \\ -\left.\frac{\partial F}{\partial x}\right|_{(x_0, y_0)} & \left.\frac{\partial F}{\partial y}\right|_{(x_0, y_0)}^{-1} \end{pmatrix}^{-1} \left(\left.\frac{\partial F}{\partial y}\right|_{(x_0, y_0)}^{-1}\right)$$

(2.26)

By the inverse function theorem, $L$ is locally bijective near $(x_0, y_0)$ with $C^1$-inverse; i.e., there exists a continuously differentiable local inverse $\Phi = (\Phi_1, \Phi_2)$ from some open set $W_1 \times W_2$ in $\mathbb{R}^n \times \mathbb{R}^m$ containing $(x_0, F(x_0, y_0))$ mapping onto $W_1 \times V_2$, for some open set $V_2$ containing $y_0$. Note that for all $(x, z) \in W_1 \times W_2$, we have

$$L(\Phi_1(x, z), \Phi_2(x, z)) = (x, z)$$

25
Note that $L$ is identity on the first $n$ coordinates, so $\Phi_1(x) = x$ for all $x \in W_1$; on the other hand, $L$ is $F$ on the last $m$ coordinates, so

$$F(x, \Phi_2(x, z)) = F(\Phi_1(x, z), \Phi_2(x, z)) = z, \quad \text{for all } x \in W_1, z \in W_2$$

In particular, we can take $z = F(x_0, y_0)$ and then

$$F(x, \Phi_2(x, F(x_0, y_0))) = F(x_0, y_0), \quad \text{for all } x \in W_1$$

(2.27)

Now define $\varphi(x) := \Phi_2(x, F(x_0, y_0))$ on from $W_1$ to $V_2$. Then $\varphi$ satisfies that $F(x, \varphi(x)) = F(x_0, y_0)$ for all $x \in W_1$ and $\varphi$ has the same $C^k$-smoothness as $\Phi_2$ and hence as $F$. Additionally

$$L(x_0, \varphi(x_0)) = L(\Phi_1(x_0, y_0), \Phi_2(x_0, y_0)) = (x_0, F(x_0, y_0)) = L(x_0, y_0)$$

Since $L$ is a bijective function from $W_1 \times V_2$ onto $W_1 \times W_2$ and $y_0 \in V_2$, $\varphi(x)$ is $y_0$ as desired.

It remains to prove the uniqueness of $\varphi$. Another way to state the uniqueness of $\varphi$ is that if $\psi$ is another $C^1$-function from some open set $V_1$ containing $x_0$ to $V_2$ satisfying $F(x, \psi(x)) = F(x_0, y_0)$ for all $x \in V_1$ and $\psi(x_0) = y_0$, then $\psi$ coincides with $\varphi$ in $V_1 \cap W_1$.

Since $\frac{\partial F}{\partial y}$ is continuous at $(x_0, y_0)$, we can find open sets $U_1, U_2$ in $\mathbb{R}^n, \mathbb{R}^m$ containing $x_0, y_0$ respectively so that $\frac{\partial F}{\partial y}(x, y)$ is invertible for all $x \in U_1, y \in U_2$. Then by the mean value theorem,

$$\int_0^1 \left. \frac{\partial F}{\partial y} \right|_{(x, y, t(y_2 - y_1))} dt$$

is invertible for all $x \in U_1$ and $y_1, y_2 \in U_2$. Suppose $\psi : U_1 \to U_2$ is another $C^1$-function satisfying $F(x, \psi(x)) = F(x_0, y_0)$ for all $x \in U_1$ and $\psi(x_0) = y_0$. Then fix $x \in U_1$,

$$\left. F(x, \varphi(x)) - F(x, \psi(x)) \right|_{= 0} = \left( \int_0^1 \left. \frac{\partial F}{\partial y} \right|_{x, \psi(x) + t(\varphi(x) - \psi(x))} dt \right) (\varphi(x) - \psi(x))$$

(2.28)

Note that $F(x, \varphi(x)) - F(x, \psi(x)) = 0$ and $\int_0^1 \left. \frac{\partial F}{\partial y} \right|_{x, \psi(x) + t(\varphi(x) - \psi(x))} dt$ is invertible, so $\varphi(x) - \psi(x) = 0$, proving the uniqueness of $\varphi$.

\begin{proof}

Remark. Neither the implicit function theorem nor its proof gives an immediate way of computing an explicit formula for the implicit function $\varphi$, but (2.26) directly implies that

$$D\varphi_{x_0} = - \left. \frac{\partial F}{\partial x} \right|_{(x_0, y_0)} \left( \left. \frac{\partial F}{\partial y} \right|_{(x_0, y_0)} \right)^{-1}$$

(2.29)

\end{proof}

**Exercise 2.6.** Let $C = \{(x, y) : x^3 \sin y + y = 4x + 3\}$. Prove that around each point $(x_0, y_0) \in C$, $C$ can be locally written as a curve in the form of either $y = f(x)$ or $x = f(y)$ for a $C^\infty$-function $f$.

**Exercise 2.7.** Let $C = \{(x, y, z) : x^3 + y^3 + z^3 + z = 0\}$. Prove that around each point $(x_0, y_0, z_0) \in C$, $C$ can be locally written as a curve in the form of $z = F(x, y)$ for a $C^\infty$-function $F$.

An important consequence of the implicit function theorem is the method of Lagrange multipliers.
Corollary 2.3.1 (Lagrange multipliers). For \( n \geq 2 \), let \( U \) be an open set in \( \mathbb{R}^n \) and \( f : U \to \mathbb{R} \) and \( g : U \to \mathbb{R} \) be \( C^1 \)-functions. Let \( x_0 \in U \) with \( Dg_{x_0} \neq 0 \). If \( f(x) \) has a local extremum at \( x_0 \) in \( U \) with respect to the constraint \( g(x) = g(x_0) \), then there is a \( \lambda \in \mathbb{R} \) so that

\[
Df_{x_0} = \lambda Dg_{x_0}
\]  

(2.30)

Proof. Since \( Dg_{x_0} \) is not zero, there exists some \( 1 \leq k \leq n \) so that \( \frac{\partial g}{\partial x_k} \bigg|_{x_0} \) is invertible; now switch the \( k \)th coordinate with the \( n \)th coordinate. By the implicit function theorem, there exists open neighborhoods \( V \) in \( \mathbb{R}^{n-1} \) and \( W \) in \( \mathbb{R}^n \) and a \( C^1 \)-function \( \varphi : \mathbb{R}^{n-1} \to \mathbb{R} \) so that

\[
g^{-1}(g(x_0)) \cap W = \{ (z, \varphi(z)) : z \in V \}, \quad (z_0, \varphi(z_0)) = x_0
\]  

(2.31)

where \( z_0 \) is the first \( (n-1) \) coordinates of \( x_0 \). (2.31) implies that optimizing \( f(x) \) with respect to the constraint \( g(x) = g(x_0) \) is equivalent to optimizing

\[
h(z) := f(z, \varphi(z))
\]  

(2.32)

Suppose \( f(x) \) has a local extremum at \( x_0 \) in \( U \) with respect to the constraint \( g(x) = g(x_0) \). Then \( Dh_{x_0} = 0 \).

For the rest of the proof, denote the first \( (n-1) \) coordinates by \( z \). Note that

\[
0 = Dh_{x_0} = \frac{df(z, \varphi(z))}{dz} \bigg|_{z=z_0} = Df_{x_0} \frac{d(z, \varphi(z))}{dz} \bigg|_{z=z_0} = Df_{x_0} \left( \frac{I_{(n-1) \times (n-1)}}{D\varphi_{z_0}} \right)
\]

(2.29)

\[
= Df_{x_0} \left( -\frac{\partial g}{\partial z} \bigg|_{x_0} \left( \frac{\partial g}{\partial x_n} \bigg|_{x_0} \right)^{-1} \right) = \frac{\partial f}{\partial z} \bigg|_{x_0} + \frac{\partial f}{\partial x_n} \bigg|_{x_0} \left( -\frac{\partial g}{\partial z} \bigg|_{x_0} \left( \frac{\partial g}{\partial x_n} \bigg|_{x_0} \right)^{-1} \right)
\]  

(2.33)

If \( \frac{\partial f}{\partial x_n} \bigg|_{x_0} = 0 \), then \( \frac{\partial f}{\partial z} \bigg|_{x_0} = 0 \) and the corollary is proved, so assume that \( \frac{\partial f}{\partial x_n} \bigg|_{x_0} \neq 0 \). Hence By (2.33),

\[
\frac{\partial g}{\partial z} \bigg|_{x_0} \left( \frac{\partial g}{\partial x_n} \bigg|_{x_0} \right)^{-1} = \frac{1}{\frac{\partial f}{\partial z} \bigg|_{x_0}} \frac{\partial f}{\partial x_n} \bigg|_{x_0} \Rightarrow \frac{\partial f}{\partial z} \bigg|_{x_0} = \frac{\partial f}{\partial x_n} \bigg|_{x_0} \frac{\partial g}{\partial z} \bigg|_{x_0}
\]

\[
Df_{x_0} = \left( \frac{\partial f}{\partial x_n} \bigg|_{x_0} \right) \left( \frac{\partial g}{\partial x_n} \bigg|_{x_0} \right) = \left( \frac{\partial f}{\partial x_n} \bigg|_{x_0} \frac{\partial g}{\partial x_n} \bigg|_{x_0} \right) = \frac{\partial f}{\partial x_n} \bigg|_{x_0} Dg_{x_0}
\]  

(2.34)

proving the corollary.

Remark. In (2.30), \( \lambda = \frac{\partial f}{\partial x_k} \bigg|_{x_0} \) for any non-vanishing \( \frac{\partial g}{\partial x_k} \bigg|_{x_0} \).
3

Variational Optimization

In Problems C (the catenary) and E (Dido’s problem), beside minimizing or maximizing the cost functionals $J$, there are also constraints prescribed by another functionals (string length and perimeter respectively). In Problems A, B, D, F and G, the boundary conditions are only given by the endpoints’ values. In this unit, we will solve variational problems that don’t have “sophisticated” boundary conditions like functional constraints. We will first introduce the famous Euler-Lagrange equation which resolves some classic variational problems with no functional constraints, and then develop Hamilton’s equation which predicts the trajectories of conservative mechanical systems.

3.1 The Euler-Lagrange equation

3.1.1 Formulation

Consider a $C^2$-function $L(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ and the functional $J$ on $C^2[a, b]$ defined by

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

We wish to locate the extreme values for $J$. Suppose at $f \in C^1[a, b]$, $J$ attains a local extreme value. Then its derivative at $f$ vanishes

$$D(J)_f = 0$$

Consider applying the perturbations $\varepsilon \eta$ with $\varepsilon \in \mathbb{R}$, for some $\eta \in C^1[a, b]$, $\eta(a) = \eta(b) = 0$. Then we can define a differentiable function $\ell : \mathbb{R} \to \mathbb{R}$ by

$$\ell(\varepsilon) := J(f + \varepsilon \eta) = \int_a^b L(t, f(t) + \varepsilon \eta(t), f'(t) + \varepsilon f''(t)) \, dt$$

By the chain rule,

$$\ell'(0) = DJ_f \left. \frac{d(f + \varepsilon \eta)}{d\varepsilon} \right|_{\varepsilon=0} = DJ_f(\eta) = 0$$

On the other hand,

$$\ell'(0) = \frac{d}{d\varepsilon} \int_a^b L(t, f(t) + \varepsilon \eta(t), f'(t) + \varepsilon f''(t)) \, dt \bigg|_{\varepsilon=0}$$

$^1$The Euler-Lagrange equation holds for assumptions weaker than $L$ being $C^2$, but for simplicity we assume $L$ is $C^2$. 
To push the differential operator \( \frac{d}{dx} \) into the integral, consider the following lemma:

**Lemma 3.1** (Leibniz integral rule). If \( g(\varepsilon, t): \mathbb{R} \times [a, b] \rightarrow \mathbb{R} \) is a \( C^1 \)-function, then

\[
\frac{d}{d\varepsilon} \int_a^b g(\varepsilon, t) \, dt \bigg|_{\varepsilon_0} = \int_a^b \frac{\partial g}{\partial \varepsilon} \bigg|_{(\varepsilon_0, t)} \, dt
\]  

(3.5)

**Proof.** For any \( h \in [-1, 1] \),

\[
\int_a^b [g(x_0 + h, t) - g(x_0, t)] \, dt = \int_a^b \int_{x_0}^{x_0 + h} \frac{\partial g}{\partial \varepsilon} \, d\varepsilon \, dt = \int_a^b \int_{x_0}^{x_0 + h} \frac{\partial g}{\partial \varepsilon} \, d\varepsilon \, dt
\]

Then

\[
\frac{d}{d\varepsilon} \int_a^b g(\varepsilon, t) \, dt \bigg|_{\varepsilon_0} = \lim_{h \to 0} \frac{\int_a^b g(x_0 + h, t) \, dt - \int_a^b g(x_0, t) \, dt}{h} = \lim_{h \to 0} \frac{1}{h} \int_{x_0}^{x_0 + h} \frac{\partial g}{\partial \varepsilon} \, dt \, d\varepsilon
\]

\[
= \frac{d}{d\varepsilon} \int_a^b \frac{\partial g}{\partial \varepsilon} \bigg|_{(\varepsilon_0, t)} \, dt
\]

Fundamental theorem of calculus

\[
\int_a^b \frac{\partial g}{\partial \varepsilon} \bigg|_{(\varepsilon_0, t)} \, dt
\]

\]

Continue on (3.4),

\[
\ell'(0) = \frac{d}{d\varepsilon} \int_a^b L(t, f(t) + \varepsilon \eta(t), f'(t) + \varepsilon f'(t)) \, dt \bigg|_{\varepsilon=0}
\]

\[
= \int_a^b \frac{\partial K(\varepsilon, t)}{\partial \varepsilon} \bigg|_{(0,t)} \, dt \quad \text{where } K(\varepsilon, t) = L(t, f(t) + \varepsilon \eta(t), f'(t) + \varepsilon f'(t))
\]

\[
= \int_a^b \left[ \frac{\partial L}{\partial x} \bigg|_{y=f(t)+\varepsilon \eta(t), z=f'(t)+\varepsilon \eta'(t)} \frac{\partial f(t) + \varepsilon \eta(t)}{\partial \varepsilon} \bigg|_{(0,t)} \right] \, dt
\]

Recall we denoted the coordinates by \( L(x, y, z) \)

\[
= \int_a^b \left[ \frac{\partial L}{\partial y} \bigg|_{y=f(t), z=f'(t)} \cdot \eta(t) + \frac{\partial L}{\partial z} \bigg|_{y=f(t), z=f'(t)} \cdot \eta'(t) \right] \, dt
\]

by parts

\[
= \int_a^b \left[ \frac{\partial L}{\partial y} \bigg|_{y=f(t), z=f'(t)} \cdot \eta(t) - \left( \frac{d}{dt} \frac{\partial L}{\partial z} \bigg|_{y=f(t), z=f'(t)} \right) \cdot \eta(t) \right] \, dt + \left[ \eta(x) \cdot \frac{\partial L}{\partial z} \bigg|_{y=f(t), z=f'(t)} \right]_a^b
\]

\[
= \eta(a) = \eta(b) \int_a^b \left[ \frac{\partial L}{\partial y} \bigg|_{y=f(t), z=f'(t)} - \left( \frac{d}{dt} \frac{\partial L}{\partial z} \bigg|_{y=f(t), z=f'(t)} \right) \right] \cdot \eta(t) \, dt
\]  

(3.6)

Through the work above, we are close to a complete derivation of the Euler-Lagrange equations, with the help of the next lemma.
Lemma 3.2 (The fundamental lemma). For all $g \in C[0,1]$ and $k \geq 1$, if

$$\int_a^b g(t) \eta(t) \, dt = 0, \quad \text{for all } \eta \in C^k[a,b] \text{ with } \eta(a) = \eta(b) = 0$$

then $g$ is identically 0 on $[a,b]$.

**Remark.** If we merely require $\eta$ to be continuous with no restriction on the values of $\eta(a)$ and $\eta(b)$, then simply take $\eta = g$ and use Exercise 1.5.

Suppose we now require $\eta$ to be continuous and $\eta(a) = \eta(b) = 0$. We prove by contradiction that $g$ is identically 0 on $[a,b]$. Assume not; $g(x_0) \neq 0$; by continuity of $g$, we can assume that $x_0$ is not the endpoint $a$ or $b$; without the loss of generality, assume $g(x_0) > 0$, for otherwise prove for $-g(x)$. By continuity of $g$, there exists some $\varepsilon > 0$ so that for all $x \in (x_0 - \varepsilon, x_0 + \varepsilon) \subset (a,b)$,

$$|g(x_0) - g(x)| < \frac{g(x_0)}{2} \quad \Rightarrow \quad g(x) > \frac{g(x_0)}{2}$$

Now take

$$\eta(x) = \begin{cases} \varepsilon - |x - x_0| & \text{for } x \in (x_0 - \varepsilon, x_0 + \varepsilon) \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

Figure 3.1: The graph for $\eta(x) = \begin{cases} \varepsilon - |x - x_0| & \text{for } x \in (x_0 - \varepsilon, x_0 + \varepsilon) \\ 0 & \text{otherwise} \end{cases}$

Then $g(t)\eta(t) \geq 0$ for all $t \in (x_0 - \varepsilon, x_0 + \varepsilon)$. Furthermore clearly

$$\eta(x) > \frac{\varepsilon}{2} \quad \text{for all } x \in \left(x_0 - \frac{\varepsilon}{2}, x_0 + \frac{\varepsilon}{2}\right)$$

However

$$0 = \int_a^b g(x) \eta(x) \, dx = \int_{(x_0 - \varepsilon, x_0 + \varepsilon)} g(x) \eta(x) \, dx \geq \int_{(x_0 - \frac{\varepsilon}{2}, x_0 + \frac{\varepsilon}{2})} g(x) \eta(x) \, dx \geq \varepsilon \cdot \frac{g(x_0)}{2} \cdot \frac{\varepsilon}{2} > 0$$

which is a contradiction; hence $g$ must be identically zero on $[a,b]$.

**Proof.** To prove the lemma, we can strengthen the results in the remark by restricting $\eta$ to $C^k[a,b]$ for any $k \geq 1$ and in addition to $\eta(a) = \eta(b) = 0$. To do so, simply replacing the $\eta$ in (3.7) by the $C^\infty$-smooth bump function

$$\eta(x) = \begin{cases} \varepsilon \exp \left(1 - (x - x_0)^2 - 1\right) & \text{for } x \in (x_0 - \varepsilon, x_0 + \varepsilon) \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

and the rest of the arguments is the same as above.

□
With (3.6) and Lemma 3.2, we have derived that

**Theorem 3.3** (Euler-Lagrange equation). Suppose $L(x,y,z) : \mathbb{R}^3 \to \mathbb{R}$ is $C^2$ and $\hat{f} \in C^2[a,b]$. Let

$$J : C^2[a,b] \to \mathbb{R}, \quad f \mapsto \int_a^b L(t, f(t), f'(t)) \, dt$$

If $J$ is differentiable at $\hat{f}$ and $DJ_{\hat{f}} = 0$, then

$$\frac{\partial L(x,y,z)}{\partial y} \bigg|_{x=\hat{f}(t), y=\hat{f}(t), z=\hat{f}'(t)} - \left( \frac{d}{dt} \frac{\partial L(x,y,z)}{\partial z} \bigg|_{x=\hat{f}(t), y=\hat{f}(t), z=\hat{f}'(t)} \right) = 0 \quad \text{for all } t \in [a,b] \quad (\text{E-L})$$

![Figure 3.2: The graph for $\eta(x) = \begin{cases} \varepsilon \exp \left( 1 - (\varepsilon^{-2}(x-x_0)^2 - 1)^{-2} \right) & \text{for } x \in (x_0-\varepsilon, x_0+\varepsilon) \\ 0 & \text{otherwise} \end{cases}$]

**Remark.** (E-L) only makes sense for $C^2$ functions $\hat{f}$ but $J$ can be defined on $C^1[a,b]$. In fact there are functionals whose optimizer are not $C^2$, but most variational problems that arise in real-world applications have sufficiently regular (at least $C^2$) solutions.

### 3.1.2 Solution to the planar geodesics (Problem A)

**Problem A** (Shortest distance between two points). *What is the shortest distance between two points $p_0, p_1$ on the plane $\mathbb{R}^2$?*

**Solution.** Recall that in §2.1.1, we set up the problem as minimizing the cost functional

$$J(f) = \int_0^a \sqrt{1 + (f'(t))^2} \, dt$$

Suppose the minimizer $\hat{f}$ is $C^2$. To apply Euler-Lagrange equation, set

$$J(f) = \int_0^a L(t, f(t), f'(t)) \, dt, \quad \text{where } L(x,y,z) = \sqrt{1 + z^2}$$

Then by (E-L),

$$\frac{d}{dt} \left[ \frac{2z}{\sqrt{1 + z^2}} \right]_{z=\hat{f}'(t)} = 0$$

So $\frac{\hat{f}'(t)}{\sqrt{1 + (\hat{f}'(t))^2}}$ is a constant, say it equals $c$. Then

$$\hat{f}'(t) = \frac{c^2}{1 - c^2}$$

is a constant, so $\hat{f}(t)$ is a straight line.
Remark. Note that satisfying the Euler-Lagrange equation is merely a necessary condition for a $C^2$ optimizer. In fact, (E-L) is insufficient even in some practical problems; like in calculus, having zero derivatives does not guarantee a minimizer or a maximizer.

3.1.3 Solution to the least surface of revolution (Problem B)

**Problem B** (Graph with least surface of revolution). Revolve the curve $y = f(x)$ between $(a, c)$ and $(b, d)$ about the $y$-axis. Which function $f$ yields the least surface area?

**Solution.** Recall that in §2.1.2, we set up the problem as minimizing the cost functional

$$J(f) = \int_a^b t \sqrt{1 + f'(t)^2} \, dt$$

Suppose the minimizer $\hat{f}$ is $C^2$. Set

$$L(x, y, z) = x \sqrt{1 + z^2}$$

Then by (E-L),

$$\frac{d}{dt} \left[ \frac{xz}{\sqrt{1 + z^2}} \right]_{x=\hat{f}(t)} = 0$$

So $\frac{x \hat{f}'(t)}{\sqrt{1 + \hat{f}'(t)^2}}$ is constant, say it equals $k$. Then

$$\hat{f}'(t) = \frac{k}{\sqrt{t^2 - k^2}}$$

Looking up this anti-derivative to get

$$\hat{f}(t) = \cosh^{-1} \left( \frac{x}{k} \right) + C$$

(3.9)

I will leave you to find the constants $k, C$ in (3.9) according to $\hat{f}(a) = c, \hat{f}(b) = d$. ◀

3.1.4 Solution to the brachistochrone (Problem D)

**Problem D** (The brachistochrone). What is the shape of a frictionless slide from one place to a lower one that yields the fastest transit time?

Recall that in §2.1.4, we set up the problem as minimizing the cost functional

$$J(f) = \int_0^b \frac{\sqrt{1 + f'(t)^2}}{\sqrt{f(t)}} \, dt$$

Suppose the minimizer $\hat{f}$ is $C^2$. Set

$$L(x, y, z) = \frac{\sqrt{1 + z^2}}{\sqrt{y}}$$

33
Then by (E-L),
\[
\left[-\frac{\sqrt{1+z^2}}{2y^{3/2}}\right]_{y=f(t), z=f'(t)} - \frac{d}{dt} \left[\frac{z}{\sqrt{y(1+z^2)}}\right]_{y=f(t), z=f'(t)} = 0
\]
(3.10)

Simplify (3.10) to get the differential equation
\[
2\hat{f}(t)\hat{f}''(t) + \hat{f}'(t)^2 + 1 = 0
\]
(3.11)

(3.11) is not easy to solve for an explicit formula, so we consider the alternative approach as follows.

**Corollary 3.3.1.** Suppose \(L(y, z) : \mathbb{R}^2 \to \mathbb{R}\) is \(C^2\) and \(\hat{f} \in C^2[a, b]\). Let
\[
J : C^2[a, b] \to \mathbb{R}, \quad f \mapsto \int_a^b L(f(t), f'(t)) \, dt
\]
If \(J\) is differentiable at \(\hat{f}\) and \(DJ_{\hat{f}} = 0\), then there exists a constant \(c\) such that
\[
L(\hat{f}(t), \hat{f}'(t)) - \hat{f}'(t) \cdot \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} = c, \quad \text{for all } t \in (a, b)
\]
(E-L, no \(x\))

**Proof.** We show that \(L(\hat{f}(t), \hat{f}'(t)) - \hat{f}'(t) \cdot \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)}\) has zero derivative, so it is constant.

\[
\begin{align*}
\frac{d}{dt} \left(L(\hat{f}(t), \hat{f}'(t)) - \hat{f}'(t) \cdot \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)}\right) &= \frac{\partial L}{\partial y} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} \cdot \hat{f}'(t) + \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} \cdot \hat{f}''(t) - \hat{f}'(t) \cdot \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} - \frac{d}{dt} \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} \\
&= \hat{f}'(t) \cdot \left(\frac{\partial L}{\partial y} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)} - \frac{d}{dt} \frac{\partial L}{\partial z} \bigg|_{y=\hat{f}(t), z=\hat{f}'(t)}\right) \\
&= 0
\end{align*}
\]

\(\square\)

**Solution.** [Problem D] Let \(L(x, y, z) = \frac{\sqrt{1+z^2}}{\sqrt{y}}\) and suppose the minimizer \(f\) of \(J\) is \(C^2\). By (E-L, no \(x\)), there exists some constant \(c\) so that
\[
c = \frac{\sqrt{1+f'(t)^2}}{\sqrt{f(t)}} - f'(t) \cdot \frac{\partial}{\partial z} \frac{\sqrt{1+z^2}}{\sqrt{y}} \bigg|_{y=f(t), z=f'(t)} = \frac{\sqrt{1+f'(t)^2}}{\sqrt{f(t)}} - \frac{f'(t)^2}{\sqrt{f(t)(1+f'(t)^2)}}
\]
\[
f(t)(1+f'(t)^2) = k \quad \text{here } k = \frac{1}{c^2}
\]

Since the object is sliding down, we have \(f'(t) \geq 0\), so
\[
f'(t) = \sqrt{\frac{k-f(t)}{f(t)}}
\]
which is a separable ordinary differential equation.

\[
\int \sqrt[\frac{f(t)}{k - f(t)}} df(t) = t + C
\]

We use trigonometric substitution to find the integral on the left-hand side: substitute \( f(t) = \frac{\theta}{2} (1 - \cos \theta) \) to get

\[
\int \sqrt[\frac{f(t)}{k - f(t)}} df(t) = \frac{k}{2} \int \frac{1 - \cos \theta}{1 + \cos \theta} \sin \theta d\theta = \frac{k}{2} \int (1 - \cos \theta) d\theta = \frac{k}{2} (\theta - \sin \theta)
\]

where \( \theta = \arccos \left( 1 - \frac{2f(t)}{k} \right) \). Then

\[
t + C = \frac{k}{2} (\theta - \sin \theta) = \frac{k}{2} \left( \arccos \left( 1 - \frac{2f(t)}{k} \right) - \sin \arccos \left( 1 - \frac{2f(t)}{k} \right) \right) 
\]

(3.12)

By \( f(0) = 0 \), we know \( C = 0 \). Rather than inverting the function in (3.12), we have a much nicer parametrization of the curve by \( \theta \) as

\[
\{(r(\theta - \sin \theta), r(1 - \cos \theta)) : 0 \leq \theta \leq \theta_0\}
\]

(3.13)

where \( \theta_0 \) is the unique solution in \([0, 2\pi]\) to

\[
\frac{1 - \cos \theta_0}{\theta_0 - \sin \theta_0} = \frac{d}{b}
\]

and

\[
r = \frac{d}{1 - \cos \theta_0}
\]

The parametrized curve in (3.13) is called a \textit{cycloid}, which can be obtained by tracing a point on the edge of a rolling wheel of radius \( r \).
3.2 Discontinuities in $||\cdot||_\infty$

Consider the $\mathbb{R}$-vector space $C^2[a, b]$ of all the continuously differentiable functions over the interval $[a, b]$. There are various norms we can equip $C^2[a, b]$ with:

$$
||f||_\infty := \sup_{x \in [a, b]} |f(x)| \quad \text{(The } L^\infty/C^0\text{-norm)}
$$

$$
||f||_{C^1, \infty} := \sup_{x \in [a, b]} |f(x)| + \sup_{x \in [a, b]} |f'(x)| \quad \text{(The } C^{1, \infty}\text{-norm)}
$$

$$
||f||_{L^p} := \left( \int_a^b |f(x)|^p \, dx \right)^{1/p} \quad \text{(The } L^p\text{-norm)}
$$

None of these norms are equivalent$^2$. In this section, we will be mainly examining why the $C^0$-norm is often a bad choice for studying stationary trajectories of some reasonably constructed functionals.

Consider the problem of minimizing the functional on $C^2[0, 1]$

$$
J(f) = \int_0^1 \left[ f(t)^2 + \left( f'(t)^2 - 1 \right)^2 \right] \, dt
$$

with $f(0) = f(1) = 0$. Note that $f(t) = 0$ is a solution to the Euler-Lagrange equation

$$
4 \left( 1 - 3f'(t)^2 \right) f''(t) + 2f(t) = 0
$$

However, we can show that $J$ is not even continuous at $f(t) = 0$ with respect to $||\cdot||_\infty$.

**Exercise 3.1.** Let $(V, ||\cdot||_V)$ and $(W, ||\cdot||_W)$ be normed linear spaces. Prove that if $J : V \to W$ is Fréchet differentiable at $q$ in $V$, then $J$ is continuous at $q$.

**Exercise 3.2.** Prove that $(C[0, 1], ||\cdot||_\infty)$ is a complete metric space.

**Exercise 3.3.** Prove that $(C^1[0, 1], ||\cdot||_{C^1})$ is a complete metric space.

**Exercise 3.4.** Show that $(C[0, 1], ||\cdot||_{L^p})$ is not a complete metric space for all $p$.

---

$^2$Two norms $||\cdot||_1$ and $||\cdot||_2$ on a vector space $X$ are called equivalent if there exist $c, d > 0$ so that $c||x||_1 \leq ||x||_2 \leq d||x||_1$ for all $x \in X$. In fact, two norms are equivalent if and only if they generate the same topology.
In this section, we will formulate a differential equation which the optimizers under the constraint must satisfy, and many important optimization problems have constraints, like Problems C (the catenary) and E (Dido’s problem).

3.3 Integral Constraints

Note that \( J \geq 0 \), but there is no continuously differentiable function \( f \) so that \( J(f) = 0 \) because it would simultaneously require \( f'(t) = 1 \) or \(-1\) and \( f(t) = 0 \). On the other hand, we can find a Cauchy sequence of functions \( \{f_n\}_n \) (Cauchy in the \( \|\cdot\|_\infty \) sense) so that \( \{J(f_n)\}_n \) converges to 0. Consider

\[
    f_n(t) = \begin{cases} 
    t - \frac{k}{n}, & \text{when } \frac{k}{n} \leq t \leq \frac{2k+1}{2n} \\
    \frac{k+1}{n} - t, & \text{when } \frac{2k+1}{2n} \leq t \leq \frac{k+1}{n} 
    \end{cases} \quad k = 0, 1, \ldots, n - 1
\]

(3.14)

You may argue that the \( f_n \)'s are not differentiable, namely the corners at the points \( t_j = \frac{j}{2n} \). But it is easy to smooth out the corners, replacing each \( f_n \) by a \( C^2 \)-approximation \( \tilde{f}_n \). For example, for each \( \varepsilon > 0 \), we can smooth out each corner at \( t_j = \frac{j}{2n} \) by replacing the function values over the neighborhood \( \left(\frac{j}{2n} - \frac{\varepsilon}{2n}, \frac{j}{2n} + \frac{\varepsilon}{2n}\right) \) by the cubic spline interpolations as follows

\[
    \tilde{f}_n(t) = \begin{cases} 
    f_n(t), & \text{if } t \not\in \bigcup_{j=0}^{n-1} \left(\frac{j}{2n} - \frac{\varepsilon}{2n}, \frac{j}{2n} + \frac{\varepsilon}{2n}\right) \\
    f_n(a_j)(2t - (3a_j - b_j))(t - b_j)^2 + f_n(b_j)(2t - (3b_j - a_j))(t - a_j)^2 \\
    \quad \quad \quad + \frac{(t - a_j)(t - b_j)^2 + (t - b_j)(t - a_j)^2}{(\varepsilon/n)^2} & \text{if } t \in (a_j, b_j), a_j = \frac{j}{2n} - \frac{\varepsilon}{2n}, b_j = \frac{j}{2n} + \frac{\varepsilon}{2n} \\
    \end{cases} 
\]

\( j = 0, 1, \ldots, n - 1 \)

Then \( \|f_n - \tilde{f}_n\|_\infty < \varepsilon \) and \( |J(f_n) - J(\tilde{f}_n)| < \varepsilon \). So we will complete the arguments “pretending” the \( f_n \)'s to be \( C^1 \).

Note that \( \|f_n\|_\infty \rightarrow 0 \), so

\[
    f_n \xrightarrow{n \to \infty} 0 \quad \text{in } \|\cdot\|_\infty
\]

(3.15)

On the other hand,

\[
    J(f_n) = \int_0^1 f_n(x)^2 \, dx = \sum_{j=0}^{2n-1} \int_{j/(2n)}^{(j+1)/(2n)} f_n(x)^2 \, dx = \frac{2n}{4n^2} = \frac{1}{n} \xrightarrow{n \to \infty} 0 \neq J(0)
\]

(3.16)

\( f_n \rightarrow 0 \) in \( \|\cdot\|_\infty \) but \( J(f_n) \not\rightarrow J(0) \), so \( J \) is not continuous at \( f(t) = 0 \), and hence not Fréchet differentiable, with respect to the \( C^0 \)-norm.

Remark. \( \{f_n\}_n \) does not converge to \( f(t) = 0 \) in the \( C^1 \)-norm; indeed, the sequence is not even Cauchy in \( \|\cdot\|_{C^1} \) because \( \|f_m - f_n\|_{C^1} = 2 \) for any \( m \neq n \).

3.3 Integral Constraints

Many important optimization problems have constraints, like Problems C (the catenary) and E (Dido’s problem). In this section, we will formulate a differential equation which the optimizers under the constraint must satisfy, and
3.3.1 Formulation

Consider $C^2$-functions $L(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ and $M(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$. We are set to locate the extreme values for $J(f) := \int_a^b L(t, f(t), f'(t)) \, dt$ with respect to the constraint $C(f) = c_0$, where

$$C(f) := \int_a^b M(t, f(t), f'(t)) \, dt$$

I.e., we want to find $f \in C^2[a, b]$ so that $J(f)$ is a local minimum/maximum on the level set $C^{-1}(\{c_0\})$.

Suppose at $f \in C^2[a, b] \cap C^{-1}(\{c_0\})$, $J$ attains a local extreme value. Let $\eta, \xi \in C^2[a, b]$ be linearly independent with $\eta(a) = \eta(b) = \xi(a) = \xi(b) = 0$. We perturb $f$ in the span of $\eta$ and $\xi$. Let $\varepsilon, \delta \in \mathbb{R}$. Define

$$\ell(\varepsilon, \delta) := J(f + \varepsilon \eta + \delta \xi) = \int_a^b L(t, f(t) + \varepsilon \eta(t) + \delta \xi(t), f'(t) + \varepsilon \eta'(t) + \delta \xi'(t)) \, dt \quad (3.17)$$

$$m(\varepsilon, \delta) := C(f + \varepsilon \eta + \delta \xi) = \int_a^b M(t, f(t) + \varepsilon \eta(t) + \delta \xi(t), f'(t) + \varepsilon \eta'(t) + \delta \xi'(t)) \, dt \quad (3.18)$$

$$F(\varepsilon, \delta) := (\ell(\varepsilon, \delta), m(\varepsilon, \delta)) \quad (3.19)$$

Suppose $DF(0, 0)$ is invertible. Then by the inverse function theorem, $F$ is a local $C^2$-bijection from $(0, 0)$ to $(J(f), c_0)$; i.e., there exists open sets $U, V$ in $\mathbb{R}^2$ containing $(0, 0), (J(f), c_0)$ respectively so that $F|_U : U \to V$ is a $C^2$-bijection. However, consider the curve $\{m = c_0\} \cap V$. Since $V$ is open, there exists some $\tau > 0$ so that $(J(f) - \tau, J(f) + \tau) \times \{c_0\} \subset \{m = c_0\} \cap V$. $F^{-1}((J(f) - \tau, J(f) + \tau) \times \{c_0\})$ is a $C^2$-curve passing through 0, 0, so $J$ cannot achieve a local extreme value at $f$. Therefore,

$$DF(0, 0) = \begin{bmatrix} \frac{\partial \ell}{\partial \varepsilon} \bigg|_{(0, 0)} & \frac{\partial \ell}{\partial \delta} \bigg|_{(0, 0)} \\ \frac{\partial m}{\partial \varepsilon} \bigg|_{(0, 0)} & \frac{\partial m}{\partial \delta} \bigg|_{(0, 0)} \end{bmatrix} \quad (3.20)$$
cannot be invertible. Equivalently, the rows of $DF_{(0,0)}$ are linearly dependent; i.e., there exists some $\lambda \in \mathbb{R}$ so that

$$
\left( \frac{\partial \ell}{\partial x} \bigg|_{(0,0)}, \frac{\partial \ell}{\partial y} \bigg|_{(0,0)} \right) = \lambda \left( \frac{\partial m}{\partial x} \bigg|_{(0,0)}, \frac{\partial m}{\partial y} \bigg|_{(0,0)} \right)
$$

(3.21)

Therefore by the derivation of the Euler-Lagrange equation,

$$
\int_a^b \left[ \frac{\partial L(x, y, z)}{\partial y} \bigg|_{y=g(t), z=f(t)} - \left( \frac{d}{dt} \frac{\partial L(x, y, z)}{\partial z} \bigg|_{y=g(t), z=f(t)} \right) \right] dt = 0
$$

(\text{E-L, one integral constraint})

Then by Lemma 3.2,

$$
\frac{\partial L(x, y, z)}{\partial y} \bigg|_{y=g(t), z=f(t)} - \left( \frac{d}{dt} \frac{\partial L(x, y, z)}{\partial z} \bigg|_{y=g(t), z=f(t)} \right) - \lambda \left( \frac{\partial M(x, y, z)}{\partial y} \bigg|_{y=g(t), z=f(t)} \right) + \lambda \left( \frac{d}{dt} \frac{\partial M(x, y, z)}{\partial z} \bigg|_{y=g(t), z=f(t)} \right) = 0
$$

(3.24)

for all $t \in (a, b)$. Now we conclude that

**Theorem 3.4** (Euler-Lagrange equation with one integral constraint). Suppose $L(x, y, z), M(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ are $C^2$ and $f \in C^2[a, b]$. Let

$J : C^2[a, b] \to \mathbb{R}, \quad f \mapsto \int_a^b L(t, f(t), f'(t)) dt$

$C : C^2[a, b] \to \mathbb{R}, \quad f \mapsto \int_a^b M(t, f(t), f'(t)) dt$

If $J$ attains extreme values at $\hat{f}$ over $C(f) = c_0$ for some $c_0 \in \mathbb{R}$, then there exists some $\lambda \in \mathbb{R}$ so that

$$
\left. \frac{\partial N_\lambda(x, y, z)}{\partial y} \right|_{y=\hat{f}(t), z=f'(t)} - \left( \frac{d}{dt} \left. \frac{\partial N_\lambda(x, y, z)}{\partial z} \right|_{y=\hat{f}(t), z=f'(t)} \right) = 0 \quad \text{for all } t \in [a, b] \quad (\text{E-L, one integral constraint})
$$

where $N_\lambda = L + \lambda M$.

**3.3.2 Solution to Dido’s problem (Problem E)**

**Problem E** (Dido’s problem). Granted a portion of a coastline of Africa as border, what is the largest country that can be enclosed by a given remaining perimeter?
Solution. We are supposed to maximize the functional

\[ J(f) = \int_a^b f(t) \, dt \]  \hspace{1cm} (3.25)

with respect to the constraint

\[ \int_a^b \sqrt{1 + f'(x)^2} \, dx = \gamma \]  \hspace{1cm} (3.26)

for some \( \gamma > 0 \). Applying (E-L) to

\[ N_\lambda(t, f(t), f'(t)) = f(t) + \lambda \sqrt{1 + f'(t)^2} \]  \hspace{1cm} (3.27)

yields that

\[
1 - \lambda \frac{d}{dt} \left[ (1 + f'(t)^2)^{-1/2} f'(t) \right] = 0
\]

\[
\frac{d}{dt} \left[ (1 + f'(t)^2)^{-1/2} f'(t) \right] = \frac{1}{\lambda}
\]

so there exists some \( k \in \mathbb{R} \) so that

\[
\frac{f'(t)}{\sqrt{1 + f'(t)^2}} = \frac{1}{\lambda} t + k \]  \hspace{1cm} (3.28)

Squaring both sides of (3.28) gives

\[
\frac{f'(t)^2}{1 + f'(t)^2} = \frac{1}{\lambda^2} (t + \lambda k)^2 
\]

\[
f'(t) = \frac{(t + \lambda k)}{\sqrt{\lambda^2 - (t + \lambda k)^2}} 
\]

\[
f(t) = \int \frac{(t + \lambda k)}{\sqrt{\lambda^2 - (t + \lambda k)^2}} \, dt = \sqrt{\lambda^2 - (t + \lambda k)^2} + C \quad \text{for some } C \in \mathbb{R}
\]

Hence \( f \) has graph of an arc of a circle of radius \( \lambda \). Note that the arclength \( \gamma \) over the chord of length \( b - a \) determines the radius \( \lambda \) uniquely by solving

\[
b - a = 2 \cdot \lambda \cdot \sin \frac{\lambda}{2 \gamma} \]  \hspace{1cm} (3.29)

I don’t think the solution of \( \lambda \) in terms of \( \gamma \) in (3.29) has explicit expression by elementary functions. \( \blacktriangleright \)

3.3.3 Solution to the catenary (Problem C)

Problem C (The catenary). Suspend a string of length \( \gamma \) between two points. What shape will the string take?

Solution. We are to minimize the functional

\[ J(f) = \int_{-a}^a f(t) \sqrt{1 + f'(t)^2} \, dt \]  \hspace{1cm} (3.30)

with respect to the constraint

\[ \int_{-a}^a \sqrt{1 + f'(t)^2} \, dt = \gamma \]  \hspace{1cm} (3.31)
Applying (E-L, no x) from Corollary 3.3.1 to
\[ N_\lambda(t, f(t), f'(t)) = f(t)\sqrt{1 + f'(x)^2} + \lambda\sqrt{1 + f'(x)^2} = (f(t) + \lambda)\sqrt{1 + f'(t)^2} \] (3.32)
yields that
\[ (f(t) + \lambda)\sqrt{1 + f'(t)^2} - f'(t) \cdot (f(t) + \lambda)(1 + f'(t))^{-1/2} f'(t) = C_1 \]
\[ f(t) + \lambda \frac{1}{\sqrt{1 + f'(t)^2}} (1 + f'(t)^2 - f'(t)^2) = C_1 \]
\[ \frac{f(t) + \lambda}{\sqrt{1 + f'(t)^2}} = C_1 \]
for some constant \( C_1 \). Rearranging,
\[ f'(t) = \pm \sqrt{\left(\frac{f(t) + \lambda}{C_1} - 1\right)} \] (3.33)
We solve the positive part and the negative part can be argued similarly. Note that (3.33) is separable with
\[ t = \int \frac{C_1 \, du}{\sqrt{u^2 - C_1^2}} \text{ here } u = f'(t) + \lambda \]
Note that \( \int \frac{1}{\sqrt{x^2 - 1}} \, dx = \cosh^{-1} +\text{constants} \), so
\[ t = C_1 \cosh^{-1} \left( \frac{u}{C_1} \right) + C_2 \text{ for some constant } C_2 \]
\[ f(t) = \lambda + C_1 \cosh \left( \frac{t - C_2}{C_1} \right) \] (3.34)
By the symmetry of catenary, \( f \) is an even function so \( C_2 = 0 \). Moreover,
\[ \gamma = \int_{-a}^{a} \sqrt{1 + f'(t)^2} \, dt = C_1 \sinh(a/C_1) - C_1 \sinh(-a/C_1) \] (3.35)
which uniquely determines \( C_1 \). By the boundary condition \( f(a) = b \), \( \lambda \) can also be determined.

3.4 The Second Variation

The solutions to variational problems given by solving the Euler-Lagrange equations are the critical points (stationary trajectories) for the variational principle, meaning that the functional derivatives at them vanish. We know that for finite dimensional differentiable optimization problems, having zero derivatives (as known as gradients/Jacobian) is merely a necessary condition for extremity. In particular, if the principle is twice differentiable, we can impose additional conditions, based on the second derivative of the cost/objective at the critical point, in order to determine that it is a local minimum, a local maximum, or a saddle point. Similarly, the solutions to the Euler-Lagrange equation may also include local minima, local maxima, and non-extremal stationary trajectories. To distinguish between the possibilities, we want to formulate a second derivative test for the objective functional. This process is known as the second variation. In this unit, we will construct and analyze it in its simplest manifestation.
3.4.1 Review of finite dimensional second derivative test

Recall that for a twice continuously differentiable objective function $F: \mathbb{R}^n \rightarrow \mathbb{R}$, its second derivative $H\bar{a} = D(DF_{\bar{a}}) = \left[ \frac{\partial^2 F}{\partial x_i \partial x_j} \right]_{i,j=1}^n$ (as known as the Hessian matrix, which is symmetric by Fubini’s theorem) is the unique linear transformation that satisfies the second Taylor expansion formula

$$F(\bar{x}) = F(\bar{a}) + D F_{\bar{a}} (\bar{x} - \bar{a}) + \frac{1}{2} \langle HF_{\bar{a}} (\bar{x} - \bar{a}), (\bar{x} - \bar{a}) \rangle + O(|\bar{x} - \bar{a}|^3)$$ \hspace{1cm} (3.36)

The second derivative test is based on the positive definiteness of its Hessian matrix. Recall that an $n \times n$ symmetric matrix $A$ is called positive definite if $\langle A\bar{v}, \bar{v} \rangle > 0$ for all non-zero $\bar{v} \in \mathbb{R}^n$. You can prove that a symmetric matrix is positive definite if and only if all of its eigenvalues are positive.

**Theorem 3.5** (Finite dimensional second derivative test). Let $F: \mathbb{R}^n \rightarrow \mathbb{R}$ be a continuously twice differentiable function and $\bar{a} \in \mathbb{R}^n$. Suppose $DF_{\bar{a}} = 0$ and $HF_{\bar{a}}$ is invertible. Then

1. If $HF_{\bar{a}}$ is positive definite, then $F$ attains a local minimum at $\bar{a}$.
2. If $HF_{\bar{a}}$ is negative definite, then $F$ attains a local maximum at $\bar{a}$.
3. $\bar{a}$ is a saddle point otherwise.

Remark. Item 3 of the Theorem 3.5 holds more generally. If $DF_{\bar{a}} = 0$ and $HF_{\bar{a}}$ has both positive and negative eigenvalues, then $\bar{a}$ is a saddle point. This is true even if $HF_{\bar{a}}$ is not invertible.

3.4.2 Formulation

Let $J: V \rightarrow \mathbb{R}$ be a twice Fréchet continuously differentiable\(^3\) functional on a normed linear space $V$ and let $u \in V$. In an analogous fashion, we wish to expand $J(u)$ near a critical point $u$. However, we will only expand the perturbation $J(u + \varepsilon \eta)$ for some $\eta \in V$ in the real variable $\varepsilon$ near $0$. Let $u, \eta \in V$ and $\varepsilon \in \mathbb{R}$. Consider the real valued function

$$\ell(\varepsilon) = J(u + \varepsilon \eta)$$ \hspace{1cm} (3.38)

Then $\ell$ has Taylor expansion near $0$

$$\ell(\varepsilon) = J(u + \varepsilon \eta) = J(u) + \varepsilon K_1(u, \eta) + \varepsilon^2 K_2(u, \eta) + O(\varepsilon^3)$$ \hspace{1cm} (3.39)

We have found that the first order term $K_1(u, \eta)$ is linear in $\eta$ as

$$K_1(u, \eta) = \ell'(0) = DJ_u(\eta)$$ \hspace{1cm} (3.40)

In particular, when $u$ is a critical point, $K_1(u, \eta) = DJ_u(\eta) = 0$. Furthermore, note that if $u$ is a local minimizer, then $K_2(u, \eta) \geq 0$ for all non-zero $\eta \in V$. Conversely, if $K_2(u, \eta) > 0$ for all non-zero $\eta \in V$, then $u$ is a local minimizer. We call $K_2(u, \eta)$ the second variation.

---

\(^3\) $J: V \rightarrow W$ is called twice Fréchet continuously differentiable at $u$ if there exists a bounded bi-linear form $D^2 F_u : V \times V \rightarrow W$ such that for all $v \in V$

$$\lim_{h \rightarrow 0} \frac{\|DF_{u+h}(v) - DF_u(v) - D^2 F_u(v, h)\|_W}{\|h\|_V} = 0$$ \hspace{1cm} (3.37)

For our sake of deriving a second derivative test, we shall omit the formality of this definition here and simply regard that the perturbation $F(u + \varepsilon \eta)$ is twice continuously differentiable in $\varepsilon$ for all $\eta \in V$. 

42
Let us explicitly evaluate the second variation for the functional

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

(3.41)

with \( L : \mathbb{R}^3 \to \mathbb{R} \) being twice continuously differentiable. Then

\[ \ell(\varepsilon) = \int_a^b L(t, u(t) + \varepsilon \eta(t), u'(t) + \varepsilon \eta'(t)) \, dt \]  

(3.42)

Suppose \( \eta(a) = \eta(b) = 0 \). By Lemma 3.5,

\[ K_2(u, \eta) = \ell''(0) = \frac{d^2}{d\varepsilon^2} \int_a^b L(t, u(t) + \varepsilon \eta(t), u'(t) + \varepsilon \eta'(t)) \, dt \]

\[ = \int_a^b \left. \frac{\partial^2}{\partial \varepsilon^2} \right|_{\varepsilon = 0} L(t, u(t) + \varepsilon \eta(t), u'(t) + \varepsilon \eta'(t)) \, dt \]

\[ = \int_a^b \left[ \frac{\partial^2 L}{\partial y \partial z} \right|_{x=t, u'(x) = u(u'(x)), \, z'=u'(x)} \eta(t)^2 + 2 \frac{\partial^2 L}{\partial y^2} \right|_{x=t, u(u'(x)), \, z'=u'(x)} \eta(t) \eta'(t) + \frac{\partial^2 L}{\partial z^2} \right|_{x=t, u(u'(x)), \, z'=u'(x)} \eta'(t)^2 \right] \, dt \]  

(3.43)

**Example 3.1.** In Problem A, for the arc length minimization functional

\[ J(f) = \int_0^a \sqrt{1 + f'(t)^2} \, dt \]

where \( L(x, y, z) = \sqrt{1 + z^2} \). To apply the second variation test, we compute the second variation

\[ K_2(u, \eta) = \int_0^a \eta'(t)^2 \, dt \]

So the second variation \( K_2(u, \eta) \) vanishes if and only if \( \eta(t) \) is a constant function. However, we set the perturbation \( \eta(t) \) to satisfy the homogeneous boundary condition that \( \eta(0) = \eta(a) = 0 \), and \( \eta \) is a non-zero \( C^1 \) function, so we must have \( K_2(u, \eta) > 0 \). Therefore we conclude that the solution to the Euler-Lagrange equation (straight line) is a local minimizer of the arc length functional \( J \).
4

Higher Dimensional Variational Problems

Throughout the unit, “.” will exclusively mean scalar multiplication and dot products will be expressed in the notation “⟨,⟩”. I want to remark that throughout the notes, I use the symbols “d,D” precisely for total derivatives and “∂” for partial derivatives. Please review Example 1.1 for total derivatives for finite dimensional differentiable functions. When a differentiable function has domain of more than one dimensions, its total derivative is a linear transformation with matrix representation of more than one columns, and a proper subcollection of these columns is called a partial derivative in this context. In particular, for finite dimensional normed linear spaces $U,V$ and $F(\vec{u},\vec{v}) : U \times V \to \mathbb{R}^k$ differentiable with $\vec{u} = (u_1, u_2, \ldots, u_n), \vec{v} = (v_1, v_2, \ldots, v_m)$, $DF(\vec{u},\vec{v}) = \left[ \frac{\partial F}{\partial u_1} \frac{\partial F}{\partial u_2} \ldots \frac{\partial F}{\partial u_n} \right] \left[ \frac{\partial F}{\partial v_1} \frac{\partial F}{\partial v_2} \ldots \frac{\partial F}{\partial v_m} \right]$.

4.1 The Euler-Lagrange equation for trajectories in higher dimensional linear spaces

4.1.1 Formulation

Suppose we have a time-dependent conservative mechanical system whose configuration is determined by $n$ independent $C^1$-coordinates $\vec{q}(t) : [t_1, t_2] \to \mathbb{R}^n$. For a trajectory $\vec{q}$, suppose at time $t$ it has kinetic energy $T( t, \vec{q}(t), \frac{d}{dt} \vec{q}(t))$ and potential energy $V( t, \vec{q}(t), \frac{d}{dt} \vec{q}(t))$, where $T(x, \vec{y}, \vec{z}), V(x, \vec{y}, \vec{z}) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Recall from §2.1.6 that the Lagrangian is $L = T - V$ and by Hamilton’s principle, all the trajectories the system actually takes must be critical points of the functional

$$J(\vec{q}) = \int_{t_1}^{t_2} L \left( t, \vec{q}(t), \frac{d}{dt} \vec{q}(t) \right) \, dt \quad (4.1)$$

A direct computation similar to the proof of the Euler-Lagrange equation (Theorem 3.3) shall prove the following theorem.
Theorem 4.1 (Euler-Lagrange equation for higher dimensional trajectories/Lagrange equations). Suppose \( L(x, \vec{y}, \vec{z}) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) is \( C^2 \) and \( \vec{q} : [t_1, t_2] \to \mathbb{R}^n \) has all coordinates \( q_1, q_2, \ldots, q_n \in C^2[t_1, t_2] \). Let
\[
J(\vec{q}) = \int_{t_1}^{t_2} L \left( t, \vec{q}(t), \frac{d}{dt} \vec{q}(t) \right) dt
\]

If \( J \) is differentiable at \( \vec{q} \) and \( DJ_\vec{q} = 0 \), then
\[
\frac{\partial L}{\partial \vec{y}} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} - \frac{d}{dt} \left( \frac{\partial L}{\partial \vec{z}} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \right) = 0 \quad \text{for all } t \in (t_1, t_2)
\]
(4.2)

Proof. We adapt the proof of the Euler-Lagrange equation (Theorem 3.3). Let \( \vec{\eta} : [t_1, t_2] \to \mathbb{R}^n \) have all of its coordinates in \( C^1[t_1, t_2] \), with \( \vec{\eta}(t_1) = \vec{\eta}(t_2) = \vec{0} \). Consider applying the perturbation in the direction \( \vec{\eta} \) by a scale of \( \varepsilon \in \mathbb{R} \). Then we can define a differentiable function
\[
\ell(\varepsilon) := J(\vec{q} + \varepsilon \vec{\eta}) = \int_{t_1}^{t_2} L \left( t, \vec{q}(t) + \varepsilon \vec{\eta}(t), \frac{d}{dt} \vec{q}(t) + \varepsilon \frac{d}{dt} \vec{\eta}(t) \right) dt
\]
(4.3)

By the chain rule,
\[
\ell'(0) = DJ_\vec{q} \frac{d(\vec{q} + \varepsilon \vec{\eta})}{d\varepsilon} = D\ell_\vec{q} \frac{d\vec{\eta}}{d\varepsilon} = 0
\]
(4.4)

On the other hand,
\[
\ell'(0) = \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} \int_{t_1}^{t_2} L \left( t, \vec{q}(t) + \varepsilon \vec{\eta}(t), \frac{d}{dt} \vec{q}(t) + \varepsilon \frac{d}{dt} \vec{\eta}(t) \right) dt
\]

(3.5)

\[
= \int_{t_1}^{t_2} \frac{\partial}{\partial \varepsilon} \bigg|_{\varepsilon=0} \left( \frac{\partial L}{\partial \vec{y}} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \cdot \vec{\eta}(t) + \frac{\partial L}{\partial \vec{z}} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \cdot \frac{d}{dt} \vec{\eta}(t) \right) dt
\]

(4.5)

\[
= \sum_{k=1}^{n} \int_{t_1}^{t_2} \left( \frac{\partial L}{\partial \eta_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \cdot \eta_k(t) + \frac{\partial L}{\partial \eta_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \cdot \frac{d}{dt} \eta_k(t) \right) dt
\]

(4.6)

For each \( k = 1, 2, \ldots, n \), take the perturbation direction \( \vec{\eta} \) to be vanishing except the \( k \)th coordinate; i.e., \( \eta_i = 0 \) for all \( i \neq k \), and then (4.6) becomes
\[
\int_{t_1}^{t_2} \eta_k(t) \left( \frac{\partial L}{\partial \eta_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} - \frac{d}{dt} \frac{\partial L}{\partial \eta_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \frac{d}{dt} \vec{q}(t)} \right) dt
\]
(4.7)
by the fundamental lemma (Lemma 3.2), (4.7) implies that
\[
\frac{\partial L}{\partial y_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \vec{q}(t)} = \frac{d}{dt} \frac{\partial L}{\partial \vec{x}_k} \bigg|_{\vec{y} = \vec{q}(t), \vec{z} = \vec{q}(t)}
\]
proving the theorem.

\[\square\]

4.1.2 Planar pendulum

4.1.3 The Hamiltonian

4.2 The Euler-Lagrange equation for higher dimensional domains

On a time-dependent curve, every state (position on the curve) can be uniquely identified by the single time variable. Rather than searching for curves that minimize cost functionals, in this section we turn our attention to the “trajectories” (manifolds) whose states (positions on the manifolds) are determined by several variables. For example, you can think about this as searching for optimal surfaces, whose states are determined by two variables. This leads naturally to minimal surfaces, Plateau’s problem, stable flows, Schrödinger’s equation and etc.

4.2.1 Formulation

Let \( U \) be an open and bounded (equivalently, \( U \) has compact closure) in \( \mathbb{R}^n \). Consider a functional \( J \) defined on \( C^1(U) = \{ u : U \to \mathbb{R} : u \text{ is continuously differentiable on } U, \partial U \text{ and continuous on } \bar{U} \} \) of the form
\[
J(u) = \int_U L(\vec{y}, u(\vec{t}), Du) d\vec{t}
\]
where \( L(\vec{x}, y, \vec{z}) : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \). Then a direct computation gives the following theorem.

**Theorem 4.2** (Euler-Lagrange equation for higher dimensional domains). *Let \( U \) be an open and bounded subset of \( \mathbb{R}^n \). Suppose that \( u \in C^2(U) \) is minimal in the sense that
\[
J(u) \leq J(v) \text{ for all } v \in C^2(U) \text{ satisfying } v = u \text{ on } \partial U
\]
Then
\[
\frac{\partial L}{\partial \vec{y}} \bigg|_{\vec{x} = \vec{y}(t), \vec{z} = Du_t} - \text{trace} \left( D_t \left[ \frac{\partial L}{\partial \vec{z}} \bigg|_{\vec{x} = \vec{y}(t), \vec{z} = Du_t} \right] \right) = 0 \text{ for all } \vec{t} \in U \quad \text{(E-L for higher dimensional domains)}
\]

**Proof.** We adapt the proof of the Euler-Lagrange equation (Theorem 3.3) with a few additional technicalities. Let \( \eta \) be compactly supported\(^1\), \( C^2 \) on \( U \) and continuous on \( \bar{U} \). Consider perturbing in the direction \( \eta \) by a (real) scale of \( \varepsilon \), namely \( u + \varepsilon \eta \). \( \eta = 0 \) on \( \partial U \), \( u = u + \varepsilon \eta \) on \( \partial U \), so by assumption
\[
J(u) \leq J(u + \varepsilon \eta) \quad \text{for all } t \in \mathbb{R}
\]

\(^1\)For a function \( f : X \to \mathbb{R} \), where \( X \) is a topological space, \( f \) is said to be *compactly supported* if there exists a compact set \( K \subset X \) so that \( f(x) = 0 \) for all \( x \in X \setminus K \). With the \( U \) defined previously as an open set in \( \mathbb{R}^n \) with compact closure, suppose \( \eta \) is supported on a compact set \( K \subset U \). Then you can prove that \( \partial U \cap K = \emptyset \), so \( \eta = 0 \) on \( \partial U \) by continuity.
Define \( \ell(\varepsilon) := J(u + \varepsilon \eta) \). Then \( \ell(\varepsilon) \) attains a (global) minimum at \( \varepsilon = 0 \), so if \( \ell \) is differentiable at 0, then \( \ell'(0) = 0 \).

On the other hand, we now try to express \( \ell'(0) \) in terms of \( L, u, \eta \).

\[
\ell(\varepsilon) = J(u + \varepsilon \eta) = \int_U L(\tilde{x}, u(\tilde{x}) + \varepsilon \eta(\tilde{x}) + Du(\tilde{x}) + \varepsilon D\eta) \, d\tilde{x}
\]

(4.12)

Note that the Leibniz integral rule (Lemma 3.1) will still hold if \( g(\varepsilon, \tilde{x}) \) is from \( \mathbb{R} \times U \) where \( U \) is an open set with compact closure; the proof is left as an easy exercise. Then we have

\[
\int_U L(\tilde{x}, u(\tilde{x}) + \varepsilon \eta(\tilde{x}) + Du(\tilde{x}) + \varepsilon D\eta) \, d\tilde{x} = \int_U \left[ \frac{\partial L}{\partial y} \bigg|_{\tilde{y} = u(\tilde{x}) + \varepsilon \eta(\tilde{x}) + Du(\tilde{x}) + \varepsilon D\eta} \right] \cdot \eta(\tilde{x}) d\tilde{x}
\]

(4.13)

In place of the by-part method, we use the divergence theorem for functions.

**Theorem 4.3** (Divergence theorem/integration by parts). Let \( U \) be an open and bounded subset of \( \mathbb{R}^n \) with \( C^1 \) boundary \( \partial U \). If \( \omega \in C^1(\overline{U}) \) and each coordinate of \( \vec{F} \) is \( C^1(\overline{U}) \), then

\[
\int_U \omega(\tilde{x}) \cdot \text{trace}(D\vec{F}) \, d\tilde{x} = \int_{\partial U} \left( \omega(\tilde{x}) \cdot \vec{F}(\tilde{x}), \vec{v}(\tilde{x}) \right) \, dS(\tilde{x}) - \int_U \left( D\omega, \vec{F}(\tilde{x}) \right) \, d\tilde{x}
\]

(Divergence theorem)

where \( dS \) is the \((n-1)\)-dimensional form, and for each \( \tilde{x} \in \partial U \), \( \vec{v}(\tilde{x}) \) is the vector perpendicular to \( \tilde{x} \) with length 1.

The proof of the divergence theorem can be found in most vector calculus textbooks, so we omit it here. By the divergence theorem (by setting \( \vec{F}(\tilde{x}) = \frac{\partial L}{\partial \vec{z}} \bigg|_{\tilde{z} = \tilde{x}} \) and \( \omega = \eta \)), we rewrite the second term of (4.13) to be

\[
\int_U \left( \frac{\partial L}{\partial \vec{z}} \bigg|_{\tilde{z} = \tilde{x}} \cdot \eta(\tilde{x}) \right) \, d\tilde{x} = \int_{\partial U} \left( \eta(\tilde{x}) \cdot \frac{\partial L}{\partial \vec{z}} \bigg|_{\tilde{z} = \tilde{x}} \right) \, dS(\tilde{x}) - \int_U \text{trace} \left( D\vec{v} \bigg|_{\tilde{z} = \tilde{x}} \right) \, d\tilde{x}
\]

(4.14)

Substituting (4.14) back to (4.13), we get

\[
\ell'(0) = \int_U \left[ \frac{\partial L}{\partial y} \bigg|_{\tilde{y} = u(\tilde{x}) + \varepsilon \eta(\tilde{x}) + Du(\tilde{x}) + \varepsilon D\eta} + \text{trace} \left( D\vec{v} \bigg|_{\tilde{z} = \tilde{x}} \right) \right] \cdot \eta(\tilde{x}) \, d\tilde{x}
\]

(4.15)

Now it suffices to prove a higher dimensional generalization of the fundamental lemma (Lemma 3.2).

**Lemma 4.4** (The fundamental lemma for higher dimensions). Let \( U \) be an open and bounded subset of \( \mathbb{R}^n \). For all \( g \in C(U) \) and \( k \geq 1 \), if

\[
\int_U g(\tilde{x}) \eta(\tilde{x}) \, d\tilde{x} = 0, \quad \text{for all compactly supported } \eta \in C^k(U)
\]

(4.16)

then \( g \) is identically 0 on \( U \).
Proof. Assume to the contrary that \( g \) satisfies (4.16) but \( g(t_0) \neq 0 \) at some \( t_0 \in U \). Without the loss of generality, assume \( g(t_0) > 0 \). By the continuity of \( g \), there exists \( \delta > 0 \) such that \( \overline{B}(t_0, \delta) \subset U \) and

\[
g(t) \geq \frac{g(t_0)}{2} \quad \text{for all } t \in B(t_0, \delta)
\]

Take \( \eta \) to be a smooth bump function at \( t_0 \), supported on \( B(t_0, \delta) \); specifically,

\[
\eta(t) = \begin{cases} 
\exp \left( -\frac{1}{\delta^2 - |t-t_0|^2} \right) & \text{if } t \in B(t_0, \delta) \\
0 & \text{otherwise}
\end{cases}
\]

(4.17)

Then

- \( \eta \) is supported on the compact set \( \overline{B}(t_0, \delta) \),
- \( \eta \) is infinitely differentiable, and
- \( \eta(t) > 0 \) for all \( t \in B(t_0, \delta) \) and \( \eta \) is zero on \( U \setminus B(t_0, \delta) \).

Therefore

\[
0 = \int_U g(t)\eta(t) \, dt = \int_{B(t_0, \delta)} g(t)\eta(t) \, dt \geq \frac{g(t_0)}{2} \int_{B(t_0, \delta)} \eta(t) \, dt > 0
\]

which is a contradiction. Thus \( g \) must be identically 0 on \( U \) if \( g \) satisfies (4.16).

Together with Lemma 4.4, (4.15) implies (E-L for higher dimensional domains), completing the proof.

4.2.2 Minimal surface

Like finding curves between two points with minimal arc length, called geodesics, another naïve question revealing some underlying geometry of Euclidean spaces can be asked.

**Problem H.** What surface \( \{(x, y, z) : z = f(x, y)\} \) with planar boundary curve \( \{(x, y, z) : x^2 + y^2 = 1, z = 0\} \) has the least surface area?

**Solution.** For a surface \( z = f(x, y) \), the surface area is the functional

\[
A(f) = \int_{B(0,1) \cap \{(t_1, t_2) : t_1^2 + t_2^2 \leq 1\}} \sqrt{1 + \left( \frac{\partial f}{\partial t_1} \right)^2 + \left( \frac{\partial f}{\partial t_2} \right)^2} \, dt_1 \, dt_2
\]

(4.19)

Hence

\[
L(x_1, x_2, y, z_1, z_2) = \sqrt{1 + z_1^2 + z_2^2}
\]

(4.20)

By (E-L for higher dimensional domains), for all \((t_1, t_2) \in B(0,1) \),

\[
\frac{\partial}{\partial y} \sqrt{1 + z_1^2 + z_2^2} \bigg|_{(x_1, x_2) = (t_1, t_2) \atop (z_1, z_2) = \left( \frac{\partial f}{\partial t_1}, \frac{\partial f}{\partial t_2} \right)} - \text{trace}
\]

\[
\begin{bmatrix}
\frac{\partial}{\partial t_1} \frac{\partial}{\partial z_1} + \frac{\partial}{\partial t_2} \frac{\partial}{\partial z_2} & \frac{\partial}{\partial t_1} \frac{\partial}{\partial z_1} + \frac{\partial}{\partial t_2} \frac{\partial}{\partial z_2} \\
\frac{\partial}{\partial t_1} \frac{\partial}{\partial z_1} + \frac{\partial}{\partial t_2} \frac{\partial}{\partial z_2} & \frac{\partial}{\partial t_1} \frac{\partial}{\partial z_1} + \frac{\partial}{\partial t_2} \frac{\partial}{\partial z_2}
\end{bmatrix}
\]

(4.21)
\[ \frac{\partial}{\partial y} \left( x_1 x_2 \right) - \frac{\partial^2 g}{\partial \zeta_1^2} - \frac{\partial^2 g}{\partial \zeta_2^2} = 0 \]

\[ \begin{align*}
\frac{\partial}{\partial t_1} \left( \frac{\partial f}{\partial t_1} \right) + \frac{\partial}{\partial t_2} \left( \frac{\partial f}{\partial t_2} \right) &= 0 \\
\frac{\partial^2 f}{\partial t_1 \partial t_2} &= 0
\end{align*} \]

(4.22)

\[ \frac{\partial^2 f}{\partial t_1 \partial t_2} = 0 \]

(4.23)

Let us employ an ingenious trick from the proof of the maximal principle for the heat equation to prove that \( f = 0 \) on \( \partial B(0,1) \). Let \( g(t_1, t_2) = f(t_1, t_2) + \varepsilon t_1^2 \) for some \( \varepsilon > 0 \). Then (4.23) becomes

\[ \left( \frac{\partial^2 g}{\partial t_1 \partial t_2} - 2\varepsilon \right) \left( 1 + \left( \frac{\partial g}{\partial t_2} \right)^2 \right) - 2 \left( \frac{\partial g}{\partial t_1} - 2\varepsilon t_1 \right) \frac{\partial^2 g}{\partial t_1 \partial t_2} + \frac{\partial^2 g}{\partial t_2^2} \left( 1 + \left( \frac{\partial g}{\partial t_1} \right)^2 \right) = 0 \]

(4.24)

By the extreme value theorem, since \( \partial B(0,1) \) is compact and \( g \) is continuous, \( g \) has both global maximum and global minimum over \( \partial B(0,1) \). Suppose the global maximum is attained at an interior point \( (t_1, t_2) \in B(0,1) \), we must have \( \frac{\partial g}{\partial t_1} \bigg|_{(t_1, t_2)} = \frac{\partial g}{\partial t_2} \bigg|_{(t_1, t_2)} = 0 \) and \( \frac{\partial^2 g}{\partial t_1^2} \bigg|_{(t_1, t_2)} \leq 0 \). Then (4.24) becomes

\[ \left( \frac{\partial^2 g}{\partial t_1 \partial t_2} - 2\varepsilon \right) + \frac{\partial^2 g}{\partial t_2^2} \left( 1 + 4\varepsilon^2 t_1^2 \right) = 0 \]

(4.25)

which is a contradiction because the expression on the left-hand side of (4.25) must be negative. Thus the maximizer of \( g \) must lie on the boundary \( S^1 = \{ (t_1, t_2) : t_1^2 + t_2^2 = 1 \} \). However \( f = 0 \) on the boundary \( S^1 \), so \( g \leq f + \varepsilon = \varepsilon \) on \( S^1 \), so

\[ \max_{(t_1, t_2) \in \partial B(0,1)} g(t_1, t_2) \leq \varepsilon \]

\[ f(t_1, t_2) \leq g(t_1, t_2) \leq \varepsilon \text{ for all } (t_1, t_2) \in \partial B(0,1) \]

(4.26)

Since (4.26) holds for all \( \varepsilon > 0 \), \( f \leq 0 \) on \( \partial B(0,1) \). On the other hand, the exactly same arguments hold for \( -f \), so \( f = 0 \) on \( \partial B(0,1) \).

\[ \text{\textit{Remark.}} \quad \text{You may observe from the solution that the shape of the boundary curve is irrelevant in this problem.} \]

4.2.3 A soap film

4.2.4 Stable flows

4.2.5 Schrödinger’s equation (unfinished)

The following is in the spirit of Schrödinger’s first derivation. Suppose a single electron of mass \( m \) with total energy \( E \) is orbiting its nucleus situated at the origin of the \( \mathbb{R}^3 \) space. At each position \( (t_1, t_2, t_3) \) in its orbit, the electron has an \( 1/r \) electrostatic potential energy \( V(t_1, t_2, t_3) \) and kinetic energy \( T(t_1, t_2, t_3) = \frac{p(t_1, t_2, t_3)^2}{2m} = E - V(t_1, t_2, t_3) \), where \( E \) is the constant total energy of the particle. \( U \subset \mathbb{R}^3 \) is open and \( \psi : U \to \mathbb{R} \).

\[ J(\psi) := \int_U \frac{\hbar^2}{2m} \left[ \left( \frac{\partial \psi}{\partial t_1} \right)^2 + \left( \frac{\partial \psi}{\partial t_2} \right)^2 + \left( \frac{\partial \psi}{\partial t_3} \right)^2 \right] + (V(t_1, t_2, t_3) - E)\psi(t_1, t_2, t_3)^2 d(t_1, t_2, t_3) \]

(4.27)
Here \( L(\vec{x}, y, \vec{z}) = \frac{\hbar^2}{2m}(z_1^2 + z_2^2 + z_3^2) + [V(\vec{x}) - E] \cdot y^2 \). Applying (E-L for higher dimensional domains) to get

\[
\left[ 2y \cdot (V(\vec{x}) - E) - \text{trace} \frac{d}{d(t_1, t_2, t_3)} \frac{\hbar^2}{2m} (2z_1, 2z_2, 2z_3) \right] \left[ \vec{x} = (t_1, t_2, t_3) \right] \left[ y = \psi(t_1, t_2, t_3) \right] = 0
\]

\[
2\psi(t_1, t_2, t_3) \cdot (V(t_1, t_2, t_3) - E) - \frac{2\hbar^2}{2m} \left( \frac{\partial^2 \psi}{\partial t_1 \partial t_1} + \frac{\partial^2 \psi}{\partial t_2 \partial t_2} + \frac{\partial^2 \psi}{\partial t_3 \partial t_3} \right) = 0 \tag{4.28}
\]

(4.28) is the equation for the stationary states of the wave function \( \psi \) of the hydrogen atom.
5

Analysis on Metric Spaces

5.1 Metric spaces

5.1.1 Basic definitions

A metric space is a pair \((X, d)\), where \(X\) is a set and \(d\) is a metric on \(X\), that is, a function defined \(X \times X\) such that for all \(x, y, z \in X\) we have:

(i) \(d\) is real-valued, finite and nonnegative; \(0 \leq d(x, y) < \infty\) for all \(x, y \in X\).

(ii) \(d(x, y) = 0\) if and only if \(x = y\).

(iii) (symmetry) \(d(x, y) = d(y, x)\).

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

Example 5.1 (Discrete metric spaces). Take any set \(X\), and for \(x, y \in X\) with \(x \neq y\), define \(d(x, x) = 0\) and \(d(x, y) = 1\).

Example 5.2 (\(\mathbb{R}^n\)). Let \(x, y \in \mathbb{R}^n\), where \(x = (x_1, x_2, \ldots, x_n)\) and \(y = (y_1, y_2, \ldots, y_n)\).

\[
d_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}
\]

\[
d_\infty(x, y) = \max\{ |x_k - y_k| : k = 1, 2, \ldots, n \}
\]

\[
d_1(x, y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|
\]

\[
d_p(x, y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \cdots + |x_n - y_n|^p} \quad \text{for } p \geq 1
\]

are all metrics on \(\mathbb{R}^n\). We show that \(d_p\) satisfies the triangle inequality.

1. To begin with, we first prove the Young’s inequality.

Lemma 5.1 (Young’s inequality). For real numbers \(p, q > 1\) satisfying \(\frac{1}{p} + \frac{1}{q} = 1\), and \(a, b > 0\),

\[
ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad (5.1)
\]
Proof. Let $\frac{1}{p} + \frac{1}{q} = 1$. Consider $\ln(ab) = \ln(a) + \ln(b) = \ln(a^{p/q}) + \ln(b^{q/p}) = \frac{\ln(a^{p})}{p} + \frac{\ln(b^{q})}{q}$. Let $s(x)$ be the line going through $(a^{p}, \ln(a^{p}))$ and $(b^{q}, \ln(b^{q}))$. Then

$$\frac{\ln(a^{p}) - \ln(b^{q})}{a^{p} - b^{q}}(x - a^{p}) = s(x) - \ln(a^{p})$$

$$s(x) = \frac{\ln(a^{p}) - \ln(b^{q})}{a^{p} - b^{q}}(x - a^{p}) + \ln(a^{p})$$

Since $\frac{1}{p} < 1$, $\frac{1}{p}a^{p} + \frac{1}{q}b^{q} = a^{p} + (b^{q} - a^{p})\frac{1}{q} \in [a^{p}, b^{q}]$. By the convexity (concave up) of ln-function, $s(\frac{1}{p}a^{p} + \frac{1}{q}b^{q}) \leq \ln(\frac{1}{p}a^{p} + \frac{1}{q}b^{q})$, so

$$\ln(\frac{1}{p}a^{p} + \frac{1}{q}b^{q}) \geq \frac{\ln(a^{p}) - \ln(b^{q})}{a^{p} - b^{q}}(\frac{1}{p}a^{p} + \frac{1}{q}b^{q}) + \ln(a^{p})$$

$$= \frac{\ln(a^{p}) - \ln(b^{q})}{a^{p} - b^{q}} \frac{1}{q}(b^{q} - a^{p}) + \ln(a^{p})$$

$$= \frac{1}{q}(-\ln(b^{q}) - \ln(a^{p})) + \ln(a^{p})$$

$$= \frac{1}{p} \ln(a^{p}) + \frac{1}{q} \ln(b^{q})$$

Then $e^{\ln(\frac{1}{p}a^{p} + \frac{1}{q}b^{q})} \geq e^{\frac{1}{p} \ln(a^{p}) + \frac{1}{q} \ln(b^{q})} = e^{\ln(ab)}, \frac{1}{p}a^{p} + \frac{1}{q}b^{q} \geq ab. \square$

2. Then we prove the Hőder’s Inequality.

**Lemma 5.2** (Hőder’s inequality). For real numbers $p, q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, and $x_{i}, y_{i} \in \mathbb{R}$ for $i = 1, 2, \ldots, n$,

$$\sum_{i=1}^{n} |x_{i}y_{i}| \leq \left( \sum_{i=1}^{n} |x_{i}|^{p} \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} |y_{i}|^{q} \right)^{\frac{1}{q}} \tag{5.2}$$

**Proof.** Let $\frac{1}{p} + \frac{1}{q} = 1$ and $x, y \in \mathbb{R}^{n}$. Assume $\sum_{i=1}^{n} |x_{i}|^{p} = \sum_{i=1}^{n} |y_{i}|^{q} = 1$. Then $(\sum_{i=1}^{n} |x_{i}|^{p})^{\frac{1}{p}}(\sum_{i=1}^{n} |y_{i}|^{q})^{\frac{1}{q}} = 1$ and $\sum_{i=1}^{n} |x_{i}|^{p} = 1, \sum_{i=1}^{n} |y_{i}|^{q} = 1$. Also by Young’s Inequality, $\sum_{i=1}^{n} |x_{i}y_{i}| \leq \sum_{i=1}^{n} \frac{1}{p} |x_{i}|^{p} + \sum_{i=1}^{n} \frac{1}{q} |y_{i}|^{q} = \frac{1}{p} \sum_{i=1}^{n} |x_{i}|^{p} + \frac{1}{q} \sum_{i=1}^{n} |y_{i}|^{q} = \frac{1}{p} + \frac{1}{q} = 1 \leq (\sum_{i=1}^{n} |x_{i}|^{p})^{\frac{1}{p}}(\sum_{i=1}^{n} |y_{i}|^{q})^{\frac{1}{q}}$.

Now let $x, y \in \mathbb{R}^{n}$ be two arbitrary non-zero vectors. Then by the results obtained from the last paragraph,

$$\sum_{i=1}^{n} \left| \frac{x_{i}}{\sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}}} \right| \leq \left( \sum_{i=1}^{n} \left| \frac{x_{i}}{\sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}}} \right|^{p} \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} \left| \frac{y_{i}}{\sqrt[\frac{1}{q}]{\sum_{i=1}^{n} |y_{i}|^{q}}} \right|^{q} \right)^{\frac{1}{q}}$$

$$\sum_{i=1}^{n} \sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}} \sqrt[\frac{1}{q}]{\sum_{i=1}^{n} |y_{i}|^{q}} |x_{i}y_{i}| \leq \left( \sum_{i=1}^{n} \left| \frac{x_{i}}{\sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}}} \right|^{p} \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} \left| \frac{y_{i}}{\sqrt[\frac{1}{q}]{\sum_{i=1}^{n} |y_{i}|^{q}}} \right|^{q} \right)^{\frac{1}{q}}$$

$$\sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}} \sqrt[\frac{1}{q}]{\sum_{i=1}^{n} |y_{i}|^{q}} (\sum_{i=1}^{n} |x_{i}y_{i}|) \leq \left( \sum_{i=1}^{n} \left| \frac{x_{i}}{\sqrt[\frac{1}{p}]{\sum_{i=1}^{n} |x_{i}|^{p}}} \right|^{p} \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} \left| \frac{y_{i}}{\sqrt[\frac{1}{q}]{\sum_{i=1}^{n} |y_{i}|^{q}}} \right|^{q} \right)^{\frac{1}{q}}$$

$$\sum_{i=1}^{n} |x_{i}y_{i}| \leq \left( \sum_{i=1}^{n} |x_{i}|^{p} \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} |y_{i}|^{q} \right)^{\frac{1}{q}}.$$

Assume either $x$ or $y$ is a zero vector. Without the loss of generality, let $x = 0$. Then $\sum_{i=1}^{n} |x_{i}y_{i}| = 0 \leq (\sum_{i=1}^{n} |x_{i}|^{p})^{\frac{1}{p}} (\sum_{i=1}^{n} |y_{i}|^{q})^{\frac{1}{q}}$. \square
3. Last we prove the Minkowski’s inequality.

**Lemma 5.3** (Minkowski’s inequality). For $p \geq 1$ and $x_i, y_i \in \mathbb{R}$,

$$
\left( \sum_{i=1}^{n} |x_i + y_i|^p \right)^{\frac{1}{p}} \leq \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} + \left( \sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}}
$$  \hspace{1cm} (5.3)

**Proof.** Let $p > 1$. Then

$$
\sum_{i=1}^{n} |x_i + y_i|^p = \sum_{i=1}^{n} |x_i + y_i||x_i + y_i|^{p-1}
$$

$$
\leq \sum_{i=1}^{n} (|x_i||x_i + y_i|^{p-1} + |y_i||x_i + y_i|^{p-1})
$$

$$
= \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1}
$$

$$
\leq \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} |x_i + y_i|^{(p-1)(\frac{1}{p})^{1-\frac{1}{p}}} \right)^{1-\frac{1}{p}}
$$

$$
+ \left( \sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} |x_i + y_i|^{(p-1)(\frac{1}{p})^{1-\frac{1}{p}}} \right)^{1-\frac{1}{p}}
$$

$$
= \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \left( \sum_{i=1}^{n} |x_i + y_i|^{(p-1)(\frac{1}{p})^{1-\frac{1}{p}}} \right)^{1-\frac{1}{p}}
$$

$$
\leq \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} + \left( \sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}}
$$

\[\square\]

**Exercise 5.1.** Check that on $\mathbb{R}^n$, $d_\infty$ satisfies the definition of metrics.

**Example 5.3** ($\ell^\infty$). Consider the space of all bounded real sequences $\ell^\infty$. For $(x_n), (y_n) \in \ell^\infty$, define

$$
d_\infty((x_n), (y_n)) := \sup_n |x_n - y_n|
$$  \hspace{1cm} (5.4)

$d_\infty$ is a metric on $\ell^\infty$.

**Exercise 5.2.** Check that on $\ell^\infty$, $d_\infty$ satisfies the definition of metrics.

**Example 5.4** ($\ell^p$). Let $p \geq 1$ be a fixed real number. Define $\ell^p$ to be the space of all real sequences $(x_n)$ such that $\sum_{n=1}^{\infty} |x_n|^p < \infty$;

$$
\ell^p = \left\{ (x_n) : x_n \in \mathbb{R}, \sum_{n=1}^{\infty} |x_n|^p < \infty \right\}
$$  \hspace{1cm} (5.5)

Define the metric $d_p$ on $\ell^p$ by

$$
d_p((x_n), (y_n)) = \left( \sum_{n=1}^{\infty} |x_n - y_n|^p \right)^{\frac{1}{p}}
$$  \hspace{1cm} (5.6)

**Exercise 5.3.** Check that on $\ell^p$, $d_p$ satisfies the definition of metrics. [Hint: To prove that $d_p$ satisfies the triangle inequality, follow the steps for that in $\mathbb{R}^n$.]
Example 5.5 ($C[0, 1]$). Consider the space of all real-valued and continuous functions on $[0, 1]$, denoted by $C[0, 1]$. For $f, g \in C[0, 1]$, define

$$d_\infty(f, g) = \max_{x \in [0, 1]} |f(x) - g(x)|$$

(5.7)

$$d_1(f, g) = \int_0^1 |f - g|$$

(5.8)

$$d_2(f, g) = \sqrt{\int_0^1 (f - g)^2}$$

(5.9)

$$d_p(f, g) = \int_0^1 |f - g|^p$$

(5.10)

d_\infty, d_p are metrics on $C[0, 1]$.

We check that on $C[0, 1]$, $d_p$ defined above satisfies the triangle inequality.

1. Hölder’s Inequality: $\int_0^1 |f(x)g(x)| \, dx \leq (\int_0^1 |f(x)|^p \, dx)^{\frac{1}{p}} (\int_0^1 |g(x)|^q \, dx)^{\frac{1}{q}}$ for $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. Let $f, g$ be continuous functions on $[0, 1]$. Let $p > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then by Young’s Inequality,

$$\frac{|f(x)|}{(\int_0^1 |f(x)|^p \, dx)^{\frac{1}{p}}} \cdot \frac{|g(x)|}{(\int_0^1 |g(x)|^q \, dx)^{\frac{1}{q}}} \leq \frac{|f(x)|^p}{p \int_0^1 |f(x)|^p \, dx} + \frac{|g(x)|^q}{q \int_0^1 |g(x)|^q \, dx}$$

$$= \int_0^1 \left( \frac{|f(x)|^p}{p \int_0^1 |f(x)|^p \, dx} + \frac{|g(x)|^q}{q \int_0^1 |g(x)|^q \, dx} \right) \, dx$$

$$= \frac{1}{p} \int_0^1 |f(x)|^p \, dx + \frac{1}{q} \int_0^1 |g(x)|^q \, dx$$

$$= \frac{1}{p} + \frac{1}{q} = 1.$$

$$\frac{\int_0^1 |f(x)g(x)| \, dx}{(\int_0^1 |f(x)|^p \, dx)^{\frac{1}{p}}(\int_0^1 |g(x)|^q \, dx)^{\frac{1}{q}}} \leq 1$$

$$\int_0^1 |f(x)g(x)| \, dx \leq (\int_0^1 |f(x)|^p \, dx)^{\frac{1}{p}} (\int_0^1 |g(x)|^q \, dx)^{\frac{1}{q}}.$$

2. Minkowski’s Inequality: $(\int_0^1 |f(x) + g(x)|^p \, dx)^{1/p} \leq (\int_0^1 |f(x)|^p \, dx)^{\frac{1}{p}} + (\int_0^1 |g(x)|^p \, dx)^{\frac{1}{p}}$.
Proof. Let \( f, g \) be defined above. Let \( p > 1 \).
\[
\int_0^1 |f(x) + g(x)|^p \, dx = \int_0^1 |f(x) + g(x)| |f(x) + g(x)|^{p-1} \, dx \\
\leq \int_0^1 (|f(x)| + |g(x)|) |f(x) + g(x)|^{p-1} \, dx \\
= \int_0^1 |f(x)||f(x) + g(x)|^{p-1} \, dx + \int_0^1 |g(x)||f(x) + g(x)|^{p-1} \, dx \\
\leq \left( \int_0^1 |f(x)| \, dx \right)^{\frac{1}{p}} \left( \int_0^1 |f(x) + g(x)|^{(p-1)\left(1 - \frac{1}{p}\right)} \, dx \right)^{1 - \frac{1}{p}} \\
+ \left( \int_0^1 |g(x)| \, dx \right)^{\frac{1}{p}} \left( \int_0^1 |f(x) + g(x)|^{(p-1)\left(1 - \frac{1}{p}\right)} \, dx \right)^{1 - \frac{1}{p}} \\
= \left( \int_0^1 |f(x)| \, dx \right)^{\frac{1}{p}} + \left( \int_0^1 |g(x)| \, dx \right)^{\frac{1}{p}}
\]
\( \left( \int_0^1 |f(x) + g(x)|^p \, dx \right)^{1/p} \leq \left( \int_0^1 |f(x)| \, dx \right)^{\frac{1}{p}} + \left( \int_0^1 |g(x)| \, dx \right)^{\frac{1}{p}} \)
\( \square \)

**Exercise 5.4.** Check that on \( C[0,1], d_\infty, d_p \) satisfy the definition of metrics.

### 5.1.2 Topologies of metric spaces

For a metric space \((X,d)\), the collection of all \(r\)-balls \(B(x,r)\), for \(x \in X\) and \(r > 0\), is a basis for a topology on \(X\), called the metric topology induced by the metric \(d\). A subset \(U\) of a metric space \((X,d)\) is open if and only if it contains a ball about each of its points.

For a function between two metric spaces, we can define continuity locally: Let \((X,d_X)\) and \((Y,d_Y)\) be metric spaces. A function \(f : X \to Y\) is said to be **continuous at a point** \(x \in X\) if for every \(\varepsilon > 0\), there is a \(\delta > 0\) such that whenever \(d_X(x,y) < \delta\), \(d_Y(f(x), f(y)) < \varepsilon\).

A sequence \((x_n)\) in a metric space \((X,d)\) is said to converge if there exist an \(x \in X\) such that \(\lim_{n \to \infty} d(x_n, x) = 0\). \(x\) is called the **limit** of \((x_n)\) and we write \(\lim_{n \to \infty} (x_n) = x\) or \((x_n) \to x\).

**Exercise 5.5.** Show that in a metric space \((X,d)\), a convergent sequence is bounded.

**Exercise 5.6.** Show that in a metric space \((X,d)\), if \((x_n) \to x\) and \((y_n) \to y\), then \(d(x_n, y_n) \to d(x,y)\).

Note that in the space \(C[0,1]\), convergence with respect to \(d_\infty\) is uniform convergence. Recall that a sequence of function \((f_n)\) on a metric space \((X,d)\) is said to converge uniformly to the limit \(f\) if for all \(\varepsilon > 0\), there exists \(N \in \mathbb{N}\) such that for all \(x \in X\) and for all \(n > N\), \(d(f(x), f_n(x)) < \varepsilon\). We shall distinguish uniform convergence from pointwise convergence, which is defined as: a sequence of function \((f_n)\) on a metric space \((X,d)\) is said to converge pointwise to the limit \(f\) if for all \(\varepsilon > 0\) and for all \(x \in X\), there exists \(N \in \mathbb{N}\) such that for all \(n > N\), \(d(f(x), f_n(x)) < \varepsilon\). Note that if a sequence in \(C[0,1]\) converges uniformly, then it converges uniformly to a continuous function; however, if a sequence in \(C[0,1]\) converges pointwise, then it may converge pointwise to a discontinuous function.
Exercise 5.7. Give an example of a sequence of functions in $C[0,1]$ that converges pointwise to a discontinuous function.

**Theorem 5.4.** Let $(X,d)$ be a metric space. Then the following are equivalent:

(i) $X$ is compact; i.e., every open cover of $X$ has a finite subcover.

(ii) $X$ is limit point compact; i.e., every infinite subset of $X$ has a limit point.

(iii) $X$ is sequentially compact; i.e., every sequence in $X$ has a convergent subsequence.

A sequence $(x_n)$ in a metric space $(X,d)$ is said to be *Cauchy* if for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for all $m, n > N$, $d(x_m, x_n) < \varepsilon$. The metric space $(X,d)$ is said to be *complete* if every Cauchy sequence in $X$ converges to an element in $X$.

Exercise 5.8. Show that every convergent sequence in a metric space is Cauchy.

Exercise 5.9. Show that $\mathbb{R}$ is complete.

**Example 5.6.** $\mathbb{R}^n$ is complete. To show this, consider any Cauchy sequence $(x_k)$ in $\mathbb{R}^n$ and denote $x_k = (x_k(1), x_k(2), \ldots, x_k(n))$. Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that for all $m, r > N$,

$$d(x_m, x_r) = \sqrt{\sum_{j=1}^{n} (x_m(j) - x_r(j))^2} < \varepsilon \quad (5.11)$$

This shows that for each fixed $j$, the sequence $(x_1(j), x_2(j), \ldots)$ is a Cauchy sequence in $\mathbb{R}$; by completeness of $\mathbb{R}$, the sequence $(x_1(j), x_2(j), \ldots)$ is convergent, say its limit is $x(j)$. Define $x = (x(1), x(2), \ldots, x(n))$. Then by letting $r \to \infty$ in (5.11), for all $m > N$,

$$d(x_m, x) \leq \varepsilon$$

This shows that $(x_k) \to x$, proving that $\mathbb{R}^n$ is complete.

**Example 5.7.** The space $\ell^\infty$ is complete.

**Example 5.8.** The space $\ell^p$ for $p \geq 1$ is complete.

Exercise 5.10. Prove that the space $C[0,1]$ with the metric $d_\infty$ is complete.

Exercise 5.11. Prove that the space $C^1[0,1]$ with the metric $d_{1,\infty}(f,g) = \sup_{x \in [0,1]} |f(x) - g(x)| + \sup_{x \in [0,1]} |f'(x) - g'(x)|$ is complete.

**Example 5.9.** The space $C[0,1]$ with the metric $d_p$ for $p \geq 1$ is not complete. Consider the sequence of functions $(f_n) \subseteq C([0,1])$ defined by

$$f_1(x) := 0$$

if $n \geq 2$

$$f_n(x) := \begin{cases} 0 & x \in [0, 1 - \frac{1}{n}) \\ n(x - \frac{1}{2}) + \frac{1}{2} & x \in [\frac{1}{2} - \frac{1}{n}, \frac{1}{2}] \\ 1 & x \in [\frac{1}{2}, 1] \end{cases}$$

$(f_n)$ is Cauchy, but not convergent in $C[0,1]$. 

58
For any metric space \((X,d)\), a map \(T : X \to X\) is called a **contraction mapping** on \(X\) if there exists \(q \in [0,1)\) such that \(d(T(x), T(y)) \leq qd(x, y)\) for all \(x, y \in X\).

**Theorem 5.5** (Contraction mapping theorem). For a non-empty complete metric space \((X,d)\), if \(T : X \to X\) is a contraction mapping on \(X\), then \(T\) has a unique fixed point \(x\) in \(X\); i.e., \(T(x) = x\).

**Proof.** Let \(q \in [0,1)\) such that \(d(T(x), T(y)) \leq qd(x, y)\) for all \(x, y \in X\). Take \(x_0 \in X\) and define a sequence \((x_n)\) inductively by \(x_{n+1} = T(x_n - 1)\). Then for all \(n \in \mathbb{N}\),

\[
d(x_{n+1}, x_n) \leq q^n d(x_1, x_0).
\] (5.12)

To see \((x_n)\) is a Cauchy sequence, let \(\epsilon > 0\) and take \(N \in \mathbb{N}\) such that \(q^N < \frac{\epsilon (1 - q)}{d(x_1, x_0)}\); (5.13)

if \(m > n > N\), then

\[
d(x_m, x_n) \leq \sum_{k=n}^{m-1} d(x_{k+1}, x_k)
\leq \sum_{k=n}^{m-1} q^k d(x_1, x_0)
=q^n d(x_1, x_0) \sum_{k=0}^{m-n-1} q^k
\leq q^n d(x_1, x_0) \sum_{k=0}^{\infty} q^k
=q^n d(x_1, x_0) \frac{1}{1 - q} < \epsilon.
\]

By the completeness of \(X\), \((x_n)\) has a limit in \(X\), say \(x\). Then \(x\) is a fixed point of \(T\), since by continuity of \(T\), \(x = \lim_{n \to \infty} x_n = \lim_{n \to \infty} T(x_{n-1}) = T(\lim_{n \to \infty} x_{n-1}) = T(x)\). Say \(T\) has another fixed point \(y\), then \(d(x, y) = d(T(x), T(y)) \leq qd(x, y)\) forces that \(d(x, y) = 0\) since \(q < 1\).

**5.1.3 Completion of metric spaces**

A lot of results in real analysis only hold for complete metric spaces. When we have an incomplete metric space, we want to know how we can find a complete metric space and “embed” the incomplete one into the complete space. This process is the **completion of metric spaces**.

A mapping \(T : X \to Y\) between metric spaces \((X,d_X)\) and \((Y,d_Y)\) is called an **isometry** if \(T\) preserves distances; i.e., for all \(x, y \in X\), \(d_Y(Tx, Ty) = d_X(x, y)\). The space \(X\) is said to be **isometric** with the space \(Y\) if there exists a bijective isometry of \(X\) onto \(Y\).

**Exercise 5.12.** Show that every isometry is injective.

We now state and prove the theorem that every metric space can be “completed”.

59
Theorem 5.6 (Completion of metric spaces). For a metric space \((X,d_X)\) there exists a complete metric space \((Y,d_Y)\) which has a subspace \(Z\) that is isometric with \(X\) and is dense in \(Y\). The space \(Y\) is unique up to isometries; that is, if \((W,d_W)\) is any complete metric space having a dense subspace \(Z'\) isometric with \(X\), then \(W\) and \(Y\) are isometric.

Proof. 1. We construct \((Y,d_Y)\) first. Let \((x_n)\) and \((y_n)\) be Cauchy sequences in \(X\). Define that \((x_n)\) is said to be equivalent to \((y_n)\), written as \((x_n) \sim (y_n)\), if \(\lim_{n \to \infty} d(x_n, y_n) = 0\). Let \(Y\) be the set of all equivalence classes of Cauchy sequences in \(X\). For \([[(x_n)]], [[y_n]] \in Y\), define

\[
d_Y([[x_n]], [[y_n]]) = \lim_{n \to \infty} d_X(x_n, y_n)
\]

This limit always exists since

\[
d(x_n, y_n) \leq d(x_n, x_m) + d(x_m, y_m) + d(y_m, y_n)
\]

\[
d(x_n, y_n) - d(x_m, y_m) \leq d(x_n, x_m) + d(y_m, y_n)
\]

(5.14)

Since \((x_n)\) and \((y_n)\) are Cauchy, (5.14) shows that \((d(x_n, y_n))_n\) is a Cauchy sequence of real numbers, so it is convergent; i.e., the limit \(d_Y([[x_n]], [[y_n]]) = \lim_{n \to \infty} d_X(x_n, y_n)\) always exists. [We have used the completeness of \(\mathbb{R}\) here; this is one of the reason why the construction of \(\mathbb{R}\) should be done before the completion of metric spaces.]

We also need to show that \(d_Y([[x_n]], [[y_n]])\) is independent of the choices of representatives of the equivalence classes. Let \((x_n) \sim (x'_n)\) and \((y_n) \sim (y'_n)\). Then

\[
|d(x_n, y_n) - d(x'_n, y'_n)| \leq d(x_n, x'_n) + d(y_n, y'_n) \to 0
\]

(5.15)

Exercise 5.13. Show that \(d_Y\) satisfies the definition of metrics.

2. Then we construct the isometry from \(X\) to a dense subspace \(Z\) of \(Y\). The isometry \(T : X \to Y\) is given by \(T : x \mapsto (x, x, x, \ldots)\), mapping \(x\) to the equivalent class that contains the constant sequence of \(x\). \(T\) is clearly an isometry and any isometry is injective (previous exercise). Hence \(X\) and \(T(X)\) are isometric.

Now we show that \(T(X)\) is dense in \(Y\). Let \(\varepsilon > 0\). Take any \([[x_n]] \in Y\). Then there exists \(N \in \mathbb{N}\) such that for all \(n > N\), \(d(x_n, x_N) < \frac{\varepsilon}{2}\). Then \(d_Y([[x_n]], [T(x_N)]) = \lim_{n \to \infty} d_X(x_n, x_N) \leq \frac{\varepsilon}{2} < \varepsilon\), so \(T(X)\) is dense in \(Y\).

3. We need to show that \(Y\) is complete. Let \([[x_n]]_k\) be a Cauchy sequence in \(Y\). Since \(T(X)\) is dense in \(Y\), for each \([[x_n]]_k\), there exists \(z_k \in X\) such that \(d_Y([[x_n]]_k, Tz_k) < \frac{1}{k}\). Then

\[
d_Y(Tz_k, Tz_\ell) \leq d_Y(Tz_k, [[x_n]]_k) + d_Y([[x_n]]_k, [[x_n]]_\ell) + d_Y([[x_n]]_\ell, Tz_\ell) < \frac{1}{k} + d_Y([[x_n]]_k, [[x_n]]_\ell) + \frac{1}{\ell}
\]

Therefore \((Tz_k)\) is Cauchy. Consider the equivalent class that \((z_k)\) belongs to, \([[z_k]]\). We show that \([[x_n]]_k \to [[z_k]]\).

\[
d_Y([[x_n]]_k, [[z_k]]) \leq d_Y([[x_n]]_k, Tz_k) + d_Y(Tz_k, [[z_k]]) < \frac{1}{k} + d_Y(Tz_\ell, [[z_k]]) = \frac{1}{k} + \lim_{\ell \to \infty} d_X(z_\ell, z_\ell)
\]
Take the limit as $k \to \infty$, we get
\[
\lim_{k \to \infty} d_Y([(x_n)]_k, [(z_k)]) \leq \lim_{k \to \infty} \frac{1}{k} + \lim_{k,\ell \to \infty} d_X(z_k, z_\ell) = 0
\]
showing the convergence of $([(x_n)]_k)$.

4. Now we show that $Y$ is unique up to isometries. Let $(W, d_W)$ be another complete metric space with a dense subspace $Z'$ that is isometric with $X$. Since $Z'$ is dense in $W$, for any $x, y \in W$, there exist sequences $(x_n), (y_n)$ in $Z'$ such that $(x_n) \to x$ and $(y_n) \to y$. Hence
\[
|d_W(x, y) - d_W(x_n, y_n)| \leq d_W(x, x_n) + d_W(y, y_n)
d_W(x, y) = \lim_{n \to \infty} d_W(x_n, y_n)
\]
Since $Z'$ is isometric with the dense subspace $Z$ in $Y$, the distances on $Y$ and $W$ must be the same; i.e., $Y$ and $W$ are isometric.

The space of equivalence classes of Cauchy sequences in $X$, with the metric defined by $d_Y([(x_n)], [(y_n)]) = \lim_{n \to \infty} d_X(x_n, y_n)$, is the completion of $X$. □

Note that elements in the completion $(\overline{X}, \overline{d})$ of a metric space $(X, d)$ are equivalence classes of elements of $X$. In particular, the completion of $C[0, 1]$ with respect to the metric $d_p$ is denoted by $L^p$. Elements of $L^p$ are not necessarily “continuous functions”; in the strictest sense, they are not even functions, as two elements in $L^p$ may be equal even if their values on certain points vary. For example, $f(x) = \chi_{\{1/2\}}$ and $g(x) = 0$ are in the same equivalence class because
\[
d_p(f, g) = \left( \int_0^1 |f - g|^p \right)^{1/p} = \left( \int_0^1 |\chi_{\{1/2\}}|^p \right)^{1/p} = 0
\]
However, $f(1/2) = 1$ but $g(1/2) = 0$. Hence you should keep in mind that elements in $L^p$ do not have well defined pointwise values over $[0, 1]$. We may observe that they behave more like distributions with density functions (all absolutely continuous distributions have density functions and the converse also holds) in probability theory, as you wouldn’t distinguish two distributions by the pointwise values of their density functions over a countable set. You can think about elements of $L^p$ as distributions and their representatives as density functions. Indeed, the $L^p$ spaces are closely related to probability theory/measure theory, which we will not detail here.

5.2 Normed linear spaces

Throughout this chapter, $F$ stands for $\mathbb{C}$ or $\mathbb{R}$.

5.2.1 Vector spaces

A vector space over a field $F$ is a nonempty set $X$ together with two algebraic operations, vector addition and multiplication of vectors by scalars (elements of $F$).

Example 5.10. $\mathbb{R}^n$, $\mathbb{C}^n$, $C[0,1]$, $\ell^\infty$ and $\ell^p$ are vector spaces. To see the vector addition in $\ell^2$, use Minkowski inequality.
A **subspace** of a vector space $X$ is a nonempty subset $Y$ of $X$ such that $Y$ is itself a vector space, with the two algebraic operations induced from $X$. A special subspace of $X$ is the **improper subspace** $Y = X$. Every other subspace of $X$ is called **proper**.

A **linear combination** of vectors $x_1, x_2, \ldots, x_m$ of a vector space $X$ is an expression of **finite** sum

$$\alpha_1x_1 + \alpha_2x_2 + \cdots + \alpha_mx_m$$

where the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_m$ are elements of $\mathbb{F}$.

For any nonempty subset $M \subset X$, the set of all linear combinations of vectors of $M$ is called the **span** of $M$, written $\text{Span}M$.

Linear independence and dependence of a **finite** set $M$ of vectors $x_1, x_2, \ldots, x_r$ in a vector space $X$ are defined by means of the equation

$$\alpha_1x_1 + \alpha_2x_2 + \cdots + \alpha_rx_r = 0$$

(5.16)

where $\alpha_1, \ldots, \alpha_r$ are scalars. Clearly equation (5.16) holds for $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. If this is the only solution to (5.16), the set $M$ is said to be **linearly independent**; otherwise $M$ is said to be **linearly dependent**.

An arbitrary subset $N$ of $X$ is said to be **linearly independent** if every nonempty finite subset of $M$ is linearly independent. If $N$ is linearly dependent, then at least one vector of $N$ can be written as a linear combination of the others.

A vector space $X$ is said to be **finite dimensional** if there is a positive integer $n$ such that $X$ contains a linearly independent set of $n$ vectors whereas any set of $n + 1$ or more vectors of $X$ is linearly dependent. $n$ is called the **dimension** of $X$, written $n = \dim X$. By definition, $X = \{0\}$ is finite dimensional and $\dim X = 0$. If $X$ is not finite dimensional, it is said to be **infinite dimensional**.

**Example 5.11.** $\mathbb{R}^n$, $\mathbb{C}^n$ are finite dimensional and $C[0, 1]$, $\ell^\infty$ and $\ell^p$ are infinite dimensional.

If $\dim X = n$, a linearly independent set of cardinality $n$ of vectors of $X$ is called a **basis** for $X$. If $\{e_1, e_2, \ldots, e_n\}$ is a basis for $X$, every $x \in X$ has a unique representation as a linear combination of the basis vectors $x = \alpha_1e_1 + \alpha_2e_2 + \cdots + \alpha_ne_n$.

If $X$ is any vector space, not necessarily finite dimensional, and $B$ is a linearly independent subset of $X$ which spans $X$, then $B$ is called a **Hamel basis** for $X$. Hence if $B$ is a Hamel basis for $X$, then every nonzero $x \in X$ has a unique representation as a linear combination of **finitely many** elements of $B$ with nonzero scalars as coefficients. It is clear that every finite dimensional vector space has a Hamel basis. For infinite dimensional vector spaces, an existence proof will be given by the use of Zorn’s lemma.

**Theorem 5.7.** Every vector space $X \neq \{0\}$ has a Hamel basis.

**Proof.** Let $\mathcal{M}$ be the set of all linearly independent subsets of $X$. Since $X \neq \{0\}$, it has an element $x \neq 0$ and $\{x\} \in \mathcal{M}$, so $\mathcal{M} \neq \emptyset$. Set inclusion defines a partial order on $\mathcal{M}$. Every chain $C \subset \mathcal{M}$ has an upper bound, namely the union of all subsets of $X$ which are elements of $C$. By Zorn’s lemma, $\mathcal{M}$ has a maximal element of $B$. We now show that $B$ is a Hamel basis for $X$. Let $Y = \text{Span}B$. Then $Y$ is a subspace of $X$, and $Y = X$ since otherwise for
$z \in X \setminus Y$, $B \cup \{z\}$ would be a linearly independent set containing $B$ as a proper subset, contrary to the maximality of $B$.

All Hamel bases for a given vector space have the same cardinality. This number is called the dimension of $X$.

**Exercise 5.14.** Let $X$ be an $n$-dimensional vector space. Show that any proper subspace $Y$ of $X$ has dimension less than $n$.

### 5.2.2 Normed linear spaces

Spaces like $\mathbb{F}^n$, $\ell^p$, $\ell^\infty$, $C[0,1]$ are both vector spaces and metric spaces. To combine algebraic and metric concepts of those spaces, we introduce an auxiliary concept, the norm.

A *normed space* $X$ is a vector space, with a real-valued function on $X$ whose value at an $x \in X$ is denoted by $\|x\|$ with the following properties

(i) $\|x\| \geq 0$ for all $x \in X$;

(ii) $\|x\| = 0 \iff x = 0$;

(iii) for all $\alpha \in \mathbb{F}$ and $x \in X$, $\|\alpha x\| = |\alpha|\|x\|$;

(iv) for all $x, y \in X$, $\|x + y\| \leq \|x\| + \|y\|$.

A norm on $X$ defines a metric $d$ on $X$ which is given by

$$d(x, y) = \|x - y\|, \quad x, y \in X$$

and is called the *metric induced by the norm*. A *Banach space* is a complete normed space with respect to the induced metric.

**Exercise 5.15.** Show that the norm is continuous; that is, $x \mapsto \|x\|$ is a continuous function from $(X, \|\cdot\|)$ into $\mathbb{R}$.

**Example 5.12.** The $p$-norm on $\mathbb{F}^n$ is defined as

$$\|x\|_p = \left( \sum_{k=1}^n |x_k|^p \right)^{\frac{1}{p}}$$

$(\mathbb{F}^n, \|\cdot\|_p)$ is a Banach space.

**Example 5.13.** The $\infty$-norm on $\mathbb{F}^n$ is defined as

$$\|x\|_\infty = \max_k |x_k|$$

$(\mathbb{F}^n, \|\cdot\|_\infty)$ is a Banach space.

**Example 5.14.** The space $\ell^p$ has a norm

$$\|x\|_p = \left( \sum_{k=1}^\infty |x_k|^p \right)^{\frac{1}{p}}$$

$(\ell^p, \|\cdot\|_p)$ is a Banach space.
**Example 5.15.** The space $\ell^\infty$ has a norm
\begin{equation}
\|x\|_\infty = \sup_n |x_n|
\end{equation}
$(\ell^\infty,\|\cdot\|_\infty)$ is a Banach space.

**Example 5.16.** The space $C[0,1]$ has a norm
\begin{equation}
\|f\|_p = \left(\int_0^1 |f(x)|^p \, dx\right)^{1/p}
\end{equation}
$(C[0,1],\|\cdot\|_p)$ is not a Banach space.

**Example 5.17.** The space $C[0,1]$ has a norm
\begin{equation}
\|f\|_\infty = \max_{x \in [0,1]} |f(x)|
\end{equation}
$(C[0,1],\|\cdot\|_\infty)$ is a Banach space.

**Exercise 5.16.**
1. Let $X$ be a normed space with induced metric $d$. Show that for all $x, y, z \in X$, $d(x+a, y+a) = d(x, y)$.

2. Let $S$ be the set of all $F$-valued sequences. Define
\begin{equation}
d(x, y) = \sum_{k=1}^\infty \frac{1}{2^k} \frac{|x_k - y_k|}{1 + |x_k - y_k|}
\end{equation}
Show that $(S, d)$ is a metric space and a vector space.

3. Show that no norm can induce the metric $d$ defined in part 2.

The convergence of sequences in normed spaces follow readily from the corresponding definitions for metric spaces. With the addition in vector spaces, we can define infinite series. A series $\sum_{k=1}^\infty x_k$ is said to converge absolutely if the series $\sum_{k=1}^\infty \|x_k\|$ is convergent. The concept of convergence of a series can be used to define a “basis” as follows. If a normed space $X$ contains a sequence $(e_n)$ with the property that for every $x \in X$ there is a unique sequence of scalars $(\alpha_n)$ such that
\begin{equation}
\lim_{n \to \infty} \left\| x - \sum_{k=1}^n \alpha_k e_k \right\| = 0;
\end{equation}
equivalently,
\begin{equation}
\sum_{k=1}^n \alpha_k e_k \xrightarrow{n \to \infty} x
\end{equation}
then $(e_n)$ is called a Schauder basis for $X$.

Like metric spaces, we can “complete” an incomplete normed space. Although the completion process is essentially the same as for metric spaces, we will need to check the “completed” space has algebraic operations for vector spaces and these agree with the space we started with.

**Theorem 5.8** (Completion of normed spaces). For a normed space $(X,\|\cdot\|_X)$ there exists a complete normed space $(Y,\|\cdot\|_Y)$ which has a subspace $Z$ that is isometric with $X$ and is dense in $Y$. The space $Y$ is unique up to isometries.
Proof. Exercise.

Example 5.18 (Lebesgue spaces). Let \(1 \leq p < \infty\). We can define the \(L^p\)-norm on the space \(C[0,1]\) by

\[
\|f\|_{L^p} = \left(\int_U |f(x)|^p \, dx\right)^{1/p}
\]  

(5.26)

The completion of \(C[0,1]\) with respect to the \(L^p\)-norm is called the \(p\)-Lebesgue space, denoted by \(L^p[0,1]\).

Exercise 5.17 (Important exercise). Prove that if \(0 < p < q\), then \(L^p[0,1] \supset L^q[0,1]\). [Hint: Use Hölder’s inequality on \(k_{f}k_{L^p} = \int f^p \, dx \leq \|f\|_{L^p}^p = \|f\|_{L^q}^p\); I will leave the details to you.]

5.2.3 Linear operators

For vector spaces \(X, Y\) with the same scalar field \(K\), linear operator \(T\) is an function from \(X\) to \(Y\) such that for all \(x, y \in X\) and \(\alpha \in K\),

\[
T(x + y) = Tx + Ty
\]

and

\[
T(\alpha x) = \alpha Tx.
\]

We define the null space (or kernel) of \(T\) to be the set \(\{x \in X : Tx = 0\}\), denoted by \(N(T)\).

Theorem 5.9. Let \(T : X \to Y\) be a linear operator. Then

1. The range \(\{Tx : x \in X\}\) is a vector space; if \(\dim X = n < \infty\), then the dimension of the range is less or equal to \(n\).

2. The null space \(N(T)\) is a vector space.

Proof. Exercise.

Example 5.19. On \(C[0,1]\), we can define the integration operator

\[
T[f(t)] = \int_0^t f(x) \, dx
\]  

(5.27)

for every \(f \in X\). \(T\) is a linear operator.

Now let’s consider two normed spaces \(X, Y\) and \(T : X \to Y\) a linear operator. The operator \(T\) is said to be bounded if there exists a real number \(c\) such that for all \(x \in X\),

\[
\|Tx\| \leq c\|x\|
\]  

(5.28)

The space of all bounded linear operators from \(X\) to \(Y\) forms a vector space (Exercise), denoted by \(\mathcal{B}(X,Y)\). We can define the operator norm on \(\mathcal{B}(X,Y)\) by

\[
\|T\|_B = \sup_{x \in X, \|x\| = 1} \|Tx\| = \sup_{x \in X, x \neq 0} \frac{\|Tx\|}{\|x\|}
\]  

(5.29)

Exercise 5.18. 1. Prove the second equality in (5.29).

2. Show that \((\mathcal{B}(X,Y),\|\cdot\|_B)\) is a normed space.

Exercise 5.19. Let \(X, Y, Z\) be normed spaces and \(T_1 \in \mathcal{B}(X,Y)\) and \(T_2 \in \mathcal{B}(Y,Z)\). Show that \(\|T_1 \circ T_2\| \leq \|T_1\| \|T_2\|\).
Exercise 5.20. Let $X$ be the vector space of all polynomials on $[0, 1]$. We can define the differentiation operator

$$T(f(t)) = f'(t)$$  \hspace{1cm} (5.30)

for every $f \in X$. Show that $T$ defined above is a linear operator but it is not a bounded linear operator.

Exercise 5.21. For a given continuous function $\varphi$ in $[0, 1] \times [0, 1]$, we can define an integral operator $T : C[0, 1] \to C[0, 1]$ by

$$T_\varphi[f(t)] := \int_0^1 \varphi(t, x)f(x) \, dx$$  \hspace{1cm} (5.31)

We show that $T_\varphi$ is bounded with respect to the $\infty$-norm on $C[0, 1]$. Since $[0, 1] \times [0, 1]$ is compact, $\varphi([0, 1] \times [0, 1])$ is compact. By the Heine-Borel Theorem, $\varphi$ is bounded; say $\varphi(x, t) \leq k$ for all $x, t \in [0, 1]$. Then

$$\|T_\varphi(f)\| = \max_{t \in [0, 1]} \left| \int_0^1 \varphi(t, x)f(x) \, dx \right| \leq \max_{t \in [0, 1]} \int_0^1 |\varphi(t, x)||f(x)| \, dx \hspace{1cm} (5.32)$$

$$\leq k \|f\|$$  \hspace{1cm} (5.33)

Theorem 5.10 (Continuity and boundedness). Let $T : X \to Y$ be a linear operator. Then $T$ is continuous if and only if $T$ is bounded.

Proof. Exercise.

Corollary 5.10.1. Let $T : X \to Y$ be a linear operator. If $T$ is continuous at a single point, it is continuous.

Proof. Continuity of $T$ at a point implies boundedness of $T$; indeed, assume $T$ is continuous at an arbitrary $x_0 \in X$. Then given any $\varepsilon > 0$, there is a $\delta > 0$ such that $\|Tx - Tx_0\| \leq \varepsilon$ whenever $\|x - x_0\| \leq \delta$. Take any nonzero $y \in X$ and set $x = x_0 + \frac{\delta}{\|y\|}y$; then $\|x - x_0\| = \left\| \frac{\delta}{\|y\|}y \right\| = \delta$. Then $\|Tx - Tx_0\| = \|T(x - x_0)\| = \frac{\delta \|Ty\|}{\|y\|};$ equivalently $\|Ty\| \leq \frac{\delta}{\delta} \|y\|$.

Corollary 5.10.2. The null space of any bounded linear operator $T$ is closed.

Proof. Let $x \in \overline{\mathcal{N}(T)}$. Then there exists a sequence $(x_n) \subset \mathcal{N}(T)$ that converges to $x$. Note that

$$\|Tx_n - Tx\| = \|T(x_n - x)\| \leq \|T\| \|x_n - x\| \xrightarrow{n \to \infty} 0$$  \hspace{1cm} (5.35)

Hence $(Tx_n)_n \to Tx$, so $Tx = 0$ and $x \in \mathcal{N}(T)$.

5.3 Sobolev spaces

5.3.1 Weak derivatives

Let $f \in C^1[a, b]$ and $\varphi \in C^\infty[a, b]$ with $\varphi(a) = \varphi(b) = 0$. Recall the product rule that

$$D(f \cdot \varphi) = (Df) \cdot \varphi + f \cdot D\varphi$$  \hspace{1cm} (5.36)
or equivalently
\[ \int_a^b \frac{df}{dx} \cdot \varphi(x) \, dx = - \int_0^1 f(x) \cdot \frac{d\varphi}{dx} \, dx \]  \tag{5.37}

Conversely, fix \( f \in C^1[a, b] \); if \( g \in C[a, b] \) so that
\[ \int_a^b g(x) \cdot \varphi(x) \, dx = - \int_a^b f(x) \frac{d\varphi}{dx} \, dx \quad \text{for all } \varphi \in C^\infty[a, b] \text{ with } \varphi(a) = \varphi(b) = 0 \]  \tag{5.38}

Then by subtracting (5.37) from (5.38), we have
\[ \int_a^b \left( g(x) - \frac{df}{dx} \right) \cdot \varphi(x) \, dx = 0 \quad \text{for all } \varphi \in C^\infty[a, b] \text{ with } \varphi(a) = \varphi(b) = 0 \]  \tag{5.39}

Then by the Fundamental Lemma (Lemma 3.2), \( g(x) = \frac{df}{dx} \), due to the continuity of both of the functions. However, note that (5.38) makes sense as long as \( f \) and \( g \) are integrable. Loosing the continuity of \( f, g \) to mere integrability, we can generalize the idea of “derivatives” to integrable functions.

**Definition 1** (Weak derivatives). Let \( f, g \in L^1[a, b] \). We say that \( g \) is the weak derivative of \( f \) if
\[ \int_a^b g(x) \cdot \varphi(x) \, dx = - \int_a^b f(x) \frac{d\varphi}{dx} \, dx \quad \text{for all } \varphi \in C^\infty[a, b] \text{ with } \varphi(a) = \varphi(b) = 0 \]  \tag{6.38}

The weak derivative, when it exists, is unique in the \( L^1 \)-sense. Also, it coincides with the (regular) variational derivative for differentiable functions (in the \( L^1 \)-sense of course). For a weakly differentiable function \( f \), we denote the weak derivative of \( f \) by \( Wf \).

**Exercise 5.22.** Check that the space of weakly differentiable \( L^p \)-functions with \( L^p \)-weak derivatives is a vector space.

On the space of weakly differentiable \( L^p \)-functions with \( L^p \)-weak derivatives, we can equip it the following norm
\[ \|f\|_{W^{1,p}} = \left[ \int_a^b \left( |f|_p + |Wf|_p \right) \right]^{1/p} \]  \tag{5.40}

I will leave you to check that \( \|\cdot\|_{W^{1,p}} \) is indeed a norm. The space of weakly differentiable \( L^p \)-functions with \( L^p \)-weak derivatives together with the norm \( \|\cdot\|_{W^{1,p}} \) is called a Sobolev space, denoted by \( W^{1,p}[a, b] \). In fact, Sobolev spaces are Banach spaces; we will not prove the fact here. Hence an equivalent definition of Sobolev spaces is that \( W^{1,p}[a, b] \) is the completion of \( C^1[a, b] \) with respect to the \( W^{1,p} \)-norm.

**Example 5.20.** \( f(x) = |x| \) defined \([-1, 1]\) is weakly differentiable with weak derivative \( Wf(x) = \begin{cases} -1 & \text{if } x \in [-1, 0) \\ 1 & \text{if } x \in (0, 1] \end{cases} \).

To see this, take any \( \varphi \in C^\infty[-1, 1] \) with \( \varphi(-1) = \varphi(1) = 0 \),
\[ - \int_{-1}^1 |x| \frac{d\varphi}{dx} \, dx = \int_{-1}^0 x \cdot \frac{d\varphi}{dx} \, dx - \int_0^1 x \cdot \frac{d\varphi}{dx} \, dx \\
= - \int_{-1}^0 \varphi(x) \, dx + \int_0^1 \varphi(x) \, dx = \int_{-1}^1 \begin{cases} -1 & \text{if } x \in [-1, 0) \\ 1 & \text{if } x \in (0, 1] \end{cases} \cdot \varphi(x) \, dx \]
Example 5.21. \( f(x) = \begin{cases} 
-1 & \text{if } x \in [-1,0) \\
1 & \text{if } x \in (0,1] 
\end{cases} \) defined \([-1,1]\) is not weakly differentiable. To see this, suppose to the contrary \( f \) were weakly differentiable. Then for all compactly supported \( \varphi \in C^\infty[-1,1] \),

\[
- \int_{-1}^{1} Wf \cdot \varphi = \int_{-1}^{1} f \cdot D\varphi = [\varphi(1) - \varphi(0)] - [\varphi(0) - \varphi(-1)] = 2\varphi(0) \tag{5.41}
\]

However, this is impossible because we can find a sequence of functions \((\varphi_n)_n \subset C^\infty[-1,1]\) so that \( \varphi_n(0) = 1 \) but \( \varphi_n \rightharpoonup 0 \), and then \(- \int_{-1}^{1} Wf \cdot \varphi_n \to 0 \) but \( 2\varphi_n(0) = 2 \).

5.3.2 Properties of weak derivatives and Sobolev spaces

Note that for “functions” in \( L^p[a,b] \) or \( W^{1,p}[a,b] \), pointwise values are not determined. However, it turns out that the values of an element in \( W^{1,p}[a,b] \) can be well defined in the following sense.

Theorem 5.11 (Fundamental theorem of calculus). Every element in \( W^{1,p}[a,b] \) has a unique continuous representative on \([a,b]\). I.e., for any \( f \in W^{1,p}[a,b] \) with \( 1 \leq p \leq \infty \), there exists a function \( \tilde{f} \in C[a,b] \) such that \( f \sim_{L^1} \tilde{f} \). Furthermore,

\[
\tilde{f}(x) - \tilde{f}(y) = \int_y^x Wf(t) \, dt \tag{5.42}
\]

Hence in the rest of the notes, unless otherwise advised we will use the unique continuous representatives to represent elements in \( W^{1,p}[a,b] \).

You can think of Theorem 5.11 as a generalization to the part 2 of the fundamental theorem of calculus in single-variable calculus. We usually use the following two ingredients to prove the part 2 of the fundamental theorem of calculus:

- If \( \frac{df}{dx} = 0 \) over \((a,b)\), then \( f \) is a constant function over \((a,b)\).
- The part 1 of the fundamental theorem of calculus: if \( f \) is continuous over \([a,b]\), then for all \( c \in [a,b) \), \( \frac{df}{dx} \int_c^x f(t) \, dt = f(x) \).

It turns out that an analogous process works just as well for weak derivatives and \( L^1 \)-functions. To prove Theorem 5.11, we will need the following lemma about \( L^1 \)-functions.

Lemma 5.12. Suppose \( f \in L^1[a,b] \) has the property that

\[
\int_a^b [f(x)\varphi'(x)] \, dx = 0 \quad \text{for all compactly supported } C^1 \text{-function } \varphi \text{ on } (a,b) \tag{5.43}
\]

Then there exists a constant \( C \) such that \( f \sim_{L^1} C \) on \([a,b]\).

Remark. Lemma 5.12 is a generalization of the fact that a differentiable function with zero derivative is constant. We shall notice the difference between Lemma 5.12 and the fundamental lemma (Lemma 3.2); a compactly supported function \( \varphi \) has compactly supported antiderivative if and only if \( \int_a^b \varphi = 0 \), but the bump functions we used in the proof of Lemma 3.2 do not satisfy this property.
Proof. Fix a compactly supported continuous function $\psi$ on $(a, b)$ with $\int_a^b \psi = 1$. Let $w$ be a compactly supported continuous function on $(a, b)$. Consider the function

$$h(x) := w(x) - \psi(x) \cdot \int_a^b w$$

(5.44)

Note that (i) $h$ is continuous, (ii) $h$ is compactly supported in $(a, b)$, (iii) $\int_a^b h = \int_a^b w(x) - \int_a^b \psi \cdot \int_a^b w = 0$. Hence $h$ has a unique antiderivative that is compactly supported in $(a, b)$. Then by the assumption (5.43),

$$\int_a^b f(x) \cdot \left[w(x) - \psi(x) \cdot \int_a^b w\right] \, dx = 0 \quad \text{for all compactly supported continuous function } w \text{ on } (a, b)$$

(5.45)

Rewrite (5.45) into for all compact supported continuous function $w$ on $(a, b)$,

$$\int_a^b \left[f(x) \cdot w(x) - f(x) \cdot \psi(x) \cdot \left(\int_a^b w\right)\right] \, dx = 0$$

$$\int_a^b \left[f(x) \cdot w(x) - \left[\int_a^b (f \cdot \psi)\right] \cdot w(x)\right] \, dx = 0$$

$$\int_a^b \left[f(x) - \int_a^b (f \cdot \psi)\right] \cdot w(x) \, dx = 0$$

(5.46)

Thus by the fundamental lemma (Lemma 3.2) for all $f \in C[a, b], f - \int_a^b (f \cdot \psi) = 0$ on $[a, b]$.

Since any $L^1$-function can be approximated by continuous functions in $L^1$, we can pass the result from $C[a, b]$ to $L^1[a, b]$ by dominated convergence theorem from measure theory. We omit the details here.

\[ \square \]

**Corollary 5.12.1.** For all weakly differentiable $L^1$-functions $f$, if $Wf = 0$, then $f$ is $L^1$-equivalent to a constant function.

**Lemma 5.13** (Fundamental theorem of calculus, first part). Let $g \in L^1[a, b]$ and let $c \in (a, b)$. Define

$$G(x) := \int_c^x g(t) \, dt$$

(5.47)

Then $G$ is uniformly continuous and weakly differentiable with weak derivative $g$.

**Proof.** To prove continuity of $G$, let $\varepsilon > 0$. Then by the density of $C[a, b]$ in $L^1[a, b]$, there exists some $\varphi \in C[a, b]$ so that

$$\|g - \varphi\|_{L^1} < \frac{\varepsilon}{2}$$

(5.48)

Let $M = \max_{x \in [a, b]} |\varphi|$. Take $\delta = \min \left\{ \frac{\varepsilon}{2M}, b - x \right\}$. Then for all $x \in [a, b)$, and then

$$|G(x) - G(x + \delta)| = \left| \int_x^{x+\delta} g(x) \, dx \right|$$

$$\leq \int_x^{x+\delta} \varphi(x) \, dx + \frac{\varepsilon}{2}$$

$$\leq \delta \cdot M + \frac{\varepsilon}{2} < \varepsilon$$

A similar argument (left to you as an exercise) proves that there exists some $\delta > 0$ so that for all $x \in (a, b)$,

$$|G(x) - G(x - \delta)| < \varepsilon.$$ 

Thus $G$ is uniformly continuous.
To prove that $g$ is the weak derivative of $G$, we need to show that for all compactly supported $C^\infty$-function $\varphi$ on $(a,b)$,
\[
\int_a^b g(x) \cdot \varphi(x) \, dx = \int_a^b G(x) \cdot \varphi'(x) \, dx
\]  
(5.49)

To see this,
\[
\int_a^b G(x) \cdot \varphi'(x) \, dx = \int_a^b \left[ \int_c^x g(t) \, dt \right] \cdot \varphi'(x) \, dx
\]
\[
= - \int_a^c \int_x^c g(t) \cdot \varphi'(x) \, dt \, dx + \int_c^b \int_x^c g(t) \cdot \varphi'(x) \, dt \, dx
\]
\[
= - \int_a^c \int_c^t g(t) \cdot \left[ \int_t^x \varphi'(x) \, dx \right] \, dt + \int_c^b \int_c^t g(t) \cdot \left[ \int_t^b \varphi'(x) \, dx \right] \, dt
\]
\[
= - \int_a^b g(t) \cdot \varphi \, dt
\]

Remark. Note that Lemma 5.13 tells us that any $L^1$-function $f$ can be the weak derivative of some continuous function, namely $F(x) = \int_a^x f(t) \, dt$, so weak derivatives may not have continuous or even pointwise defined representatives.

The remaining of the proof of Theorem 5.11 is essentially the same as the version we learned in calculus.

**Proof of Theorem 5.11.** Fix $c \in (a,b)$ and let $\widetilde{f}(x) = \int_c^x Wf(t) \, dt$. By Lemma 5.13,
\[
\int_a^b \widetilde{f}(x) \cdot \varphi'(x) \, dx = - \int_a^b Wf(x) \cdot \varphi(x) \, dx = \int_a^b f(x) \cdot \varphi'(x) \, dx \quad \text{for all compactly supported } C^1 \text{-function } \varphi \text{ on } (a,b)
\]  
(5.50)

so $\int_a^b (f(x) - \widetilde{f}(x)) \cdot \varphi'(x) \, dx = 0$. By Lemma 5.12, $f - \widetilde{f}$ is $L^1$-equivalent to a constant function; say the constant is $C$. Then $\widetilde{f} + C$ is the desired function. 

5.3.3 Weak compactness in Sobolev spaces

$L^p$-duality

Recall Hölder’s inequality that
\[
\|f \cdot g\|_{L^1} \leq \|f\|_{L^p} \cdot \|g\|_{L^q} \quad \text{for all } \frac{1}{p} + \frac{1}{q} = 1, f \in L^p[a,b], g \in L^q[a,b]
\]  
(5.51)

Thus we can define a function $\Phi : L^p \times L^q \to \mathbb{R}$ by $\Phi(f,g) = \int_a^b f \cdot g$. When $p = q = 2$, $\Phi$ defines the inner product on $L^2$.

**Weak topology of $L^p$ / weak convergence of $L^p$-functions**

In general Banach spaces, under the topology generated by the norm, compact sets are hard to come by, so we need to find some weaker topology so that convergence happens more often.
**Definition 2** (Weak convergence in $L^p$). Let $1 \leq p < \infty$ and $(f_n)_n \subset L^p[a,b], f \in L^p[a,b]$. Let $\frac{1}{p} + \frac{1}{q} = 1$. We say that $(f_n)_n$ converges weakly to $f$ in $L^p[a,b]$, denoted by $(f_n)_n \rightharpoonup f$, if

$$\lim_{n \to \infty} \Phi(f_n, g) = \Phi(f, g) \quad \text{for all } g \in L^q[a,b] \quad (5.52)$$

For each positive pair $\frac{1}{p} + \frac{1}{q} = 1$, for all $g \in L^q[a,b]$, $\Phi(\cdot, g)$ is a linear functional on $L^p[a,b]$. Furthermore, by Hölder’s inequality, $\Phi(\cdot, g)$ is bounded with operator norm smaller or equal to $\|g\|_{L^q}$.

**Lemma 5.14.** Let $1 \leq p < \infty$. Let $(f_n)_n \subset L^p[a,b], f \in L^p[a,b], (g_n)_n \subset L^q[a,b], g \in L^q[a,b]$. If $(f_n)_n$ converges weakly to $f$ in $L^p[a,b]$ and $(g_n)_n$ converges strongly to $g$ in $L^q[a,b]$, then

$$\lim_{n \to \infty} \Phi(f_n, g_n) = \Phi(f, g) \quad (5.53)$$

** Remark.** If both $(f_n)_n, (g_n)_n$ converge weakly to $f, g$ respectively, we don’t necessarily have $\lim_{n \to \infty} \Phi(f_n, g_n) = \Phi(f, g)$. For example, let $p = q = 2$ and $f_n = g_n = \sqrt{2} \sin(\pi nx)$. Then $(f_n)_n$ converges weakly to 0 but $\Phi(f_n, f_n) = \int_0^1 (1 - 2 \cos(2\pi nx)) \, dx \xrightarrow{n \to \infty} 1$.

**Proof.** Note that

$$|\Phi(f_n, g_n) - \Phi(f, g)| \leq |\Phi(f_n, g) - \Phi(f, g)| + |\Phi(f_n, g_n) - \Phi(f_n, g)|$$

$$= |\Phi(f_n, g) - \Phi(f, g)| + |\Phi(f_n, g_n - g)|$$

$$\leq |\Phi(f_n, g) - \Phi(f, g)| + \|f_n\|_{L^p} \|g_n - g\|_{L^q} \quad (5.54)$$

where in (5.54), $|\Phi(f_n, g) - \Phi(f, g)| \xrightarrow{n \to \infty} 0$ by weak convergence of $(f_n)_n$ to $f$ and $\|g_n - g\|_{L^q} \xrightarrow{n \to \infty} 0$ by strong convergence of $(g_n)_n$ to $g$. It is a fact that every weakly convergent sequence is bounded in the norm, but proving it will require some extra work, so we omit the proof here.

**Weak compactness in $L^p$**

**Definition 3** (Weak compactness in $L^p$). Let $1 \leq p < \infty$. We say a subset $K$ of $L^p[a,b]$ is weakly compact if every bounded sequence in $K$ converges weakly in $K$.

**Theorem 5.15** (Banach-Alaoglu Theorem). For $1 < p < \infty$, every closed ball $B = \{ f \in L^p[a,b] : \|f\|_{L^p} \leq r \}$ is weakly compact.

**Remark.** Unfortunately the Banach-Alaoglu Theorem does not hold for $p = 1$.

**Exercise 5.23.** Let $f_n(x) = \begin{cases} \frac{m}{2} & \text{if } |x| \leq \frac{1}{m} \\ 0 & \text{otherwise} \end{cases}$ with domain $[-1, 1]$. Show that

1. $\|f_n\|_{L^1} = 1$, so $(f_n)_n \subset B(0,1)$.
2. For any function $\varphi \in C[-1,1], \Phi(f_n, \varphi) = \varphi(0)$.
Weak convergence in $W^{1,p}$ and compactness of $W^{1,p}$

**Definition 4 (Weak convergence in $W^{1,p}$).** Let $1 \leq p < \infty$ and $(f_n)_n \subset W^{1,p}[a,b], f \in W^{1,p}[a,b]$. We say that $(f_n)_n$ converges weakly to $f$ in $W^{1,p}[a,b]$, denoted by $(f_n)_n \rightharpoonup W f$, if

$$f_n \rightharpoonup f \quad \text{and} \quad W f_n \rightharpoonup W f \quad \text{in} \quad L^p$$

(5.55)

**Theorem 5.16 (Compactness of Sobolev spaces).** Let $1 < p < \infty$. For any bounded sequence $(f_n)_n \subset W^{1,p}[a,b],\ (i)\ \text{There exists a subsequence} \ (f_{n_k})_k \ \text{of} \ (f_n)_n \ \text{so that} \ (f_{n_k})_k \ \text{converges strong to a function} \ f \in L^p[a,b] \ \text{in} \ L^p.

(ii) \ \text{There exists a subsequence} \ (f_{m_k})_k \ \text{of} \ (f_n)_n \ \text{so that} \ (f_{m_k})_k \ \text{converges strong to a function} \ g \in W^{1,p}[a,b] \ \text{in} \ W^{1,p}.$$

An immediate and important consequence of the compactness theorem (Theorem 5.16) is the Poincaré’s inequality.

**Theorem 5.17 (Poincaré’s inequality).** Let $1 < p < \infty$. There exists a constant $C > 0$ so that

$$\|f\|_{L^p} \leq C \|Wf\|_{L^p} \quad \text{for all} \quad f \in W^{1,p}[a,b] \ \text{with} \ f(a) = f(b) = 0$$

(5.56)

**Theorem 5.18.** Let $J(f) := \int_a^b L(t, f(t), f'(t)) \, dt$ with $L$ bounded below and $y_a, y_b \in \mathbb{R}$. Define

$$\mathcal{A} = \{ f \in W^{1,p}[a,b] : f(a) = y_a, f(b) = y_b \}$$

Suppose that there exist $\alpha > 0, \beta \geq 0$ such that

$$L(x, y, z) \geq \alpha |z|^p - \beta \quad \text{for all} \quad x \in [a,b], y, z \in \mathbb{R}$$

(5.57)

and $L(x, y, z)$ is convex (concave up) in $z$. Then there exists $\hat{f} \in \mathcal{A}$ so that

$$J(\hat{f}) = \min_{f \in \mathcal{A}} J(f)$$