Chapter 3

The Contraction Mapping Principle

The notion of a complete space is introduced in Section 1. The importance of complete metric spaces partly lies on the Contraction Mapping Principle, which is proved in Section 2. Two major applications of the Contraction Mapping Principle are subsequently given, first a proof of the Inverse Function Theorem in Section 3 and, second, a proof of the fundamental existence and uniqueness theorem for the initial value problem of differential equations in Section 4.

3.1 Complete Metric Space

In $\mathbb{R}^n$ a basic property is that every Cauchy sequence converges. This property is called the completeness of the Euclidean space. The notion of a Cauchy sequence is well-defined in a metric space. Indeed, a sequence $\{x_n\}$ in (X, d) is a **Cauchy sequence** if for every $\varepsilon > 0$, there exists some n_0 such that $d(x_n, x_m) < \varepsilon$, for all $n, m \geq n_0$. A metric space (X, d) is **complete** if every Cauchy sequence in it converges. A subset E is **complete** if $(E, d\big|_{E\times E})$ is complete, or, equivalently, every Cauchy sequence in E converges with limit in E.

Proposition 3.1. *Let (X, d) be a metric space.*

(a) Every complete set in X is closed.

(b) Every closed set in a complete metric space is complete.

In particular, this proposition shows that every subset in a complete metric space is complete if and only if it is closed.

Proof. (a) Let $E \subset X$ be complete and $\{x_n\}$ a sequence converging to some x in X. Since every convergent sequence is a Cauchy sequence, $\{x_n\}$ must converge to some z in E. By the uniqueness of limit, we must have $x = z \in E$, so E is closed.

(b) Let (X, d) be complete and E a closed subset of X. Every Cauchy sequence $\{x_n\}$ in E is also a Cauchy sequence in X. By the completeness of X, there is some x in X to which $\{x_n\}$ converges. However, as E is closed, x also belongs to E. So every Cauchy sequence in E has a limit in E.

□

Example 3.1. In 2050 it was shown that the space $\mathbb{R}$ is complete. Consequently, as the closed subsets in $\mathbb{R}$, the intervals $[a, b], (-\infty, b]$ and $[a, \infty)$ are all complete sets. In contrast, the set $[a, b)$, $b \in \mathbb{R}$, is not complete. For, simply observe that the sequence $\{b - 1/k\}, k \geq k_0$, for some large k_0, is a Cauchy sequence in $[a, b)$ and yet it does not have a limit in $[a, b)$ (the limit is b, which lies outside $[a, b)$). The set of all rational numbers, $\mathbb{Q}$, is also not complete. Every irrational number is the limit of some sequence in $\mathbb{Q}$, and these sequences are Cauchy sequences whose limits lie outside $\mathbb{Q}$.

Example 3.2. In 2060 we learned that every Cauchy sequence in $C[a, b]$ with respect to the sup-norm implies that it converges uniformly, so the limit is again continuous. Therefore, $C[a, b]$ is a complete space. The subset $E = \{f : f(x) \geq 0, \ \forall x\}$ is also complete. Indeed, let $\{f_n\}$ be a Cauchy sequence in E, it is also a Cauchy sequence in $C[a, b]$ and hence there exists some $f \in C[a, b]$ such that $\{f_n\}$ converges to f uniformly. As uniform convergence implies pointwise convergence, $f(x) = \lim_{n\to\infty} f_n(x) \geq 0$, so f belongs to E, and E is complete. Next, let $P[a, b]$ be the collection of all polynomials restricted to $[a, b]$. It is not complete. For, taking the sequence $h_n(x)$ given by

$$h_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!},$$

$\{h_n\}$ is a Cauchy sequence in $P[a, b]$ which converges to e^x. As e^x is not a polynomial, $P[a, b]$ is not complete.

To obtain a typical non-complete set, we consider the closed interval $[0, 1]$ in $\mathbb{R}$. Take away one point z from it to form $E = [a, b] \setminus \{z\}$. E is not complete, since every sequence in E converging to z is a Cauchy sequence which does not converge in E. In general, you may think of sets with "holes" being non-complete ones. Now, given a non-complete metric space, can we make it into a complete metric space by filling out all the holes? The answer turns out to affirmative. We can always enlarge a non-complete metric space into a complete one by putting in sufficiently many ideal points.

Theorem 3.2 (**Completion Theorem**). *Every metric space has a completion.*

This theorem will be further explained and proved in the appendix.

3.2 The Contraction Mapping Principle

Solving an equation $f(x) = 0$, where f is a function from $\mathbb{R}^n$ to itself frequently comes up in application. This problem can be turned into a problem for fixed points. Literally, a fixed point of a mapping is a point which becomes unchanged under this mapping. By introducing the function $g(x) = f(x) + x$, solving the equation $f(x) = 0$ is equivalent to finding a fixed point for g. This general observation underlines the importance of finding fixed points. In this section we prove the Contraction Mapping Principle, one of the oldest fixed point theorems and perhaps the most well-known one. As we will see, it has a wide range of applications.

A map $T : (X, d) \to (X, d)$ is called a **contraction** if there is a constant $\gamma \in (0, 1)$ such that $d(Tx, Ty) \leq \gamma d(x, y)$, $\forall x, y \in X$. A point x is called a **fixed point** of T if $Tx = x$. Usually we write Tx instead of $T(x)$.

Theorem 3.3 (**Contraction Mapping Principle**). *Every contraction in a complete metric space admits a unique fixed point.*

This theorem is also called **Banach's Fixed Point Theorem**.

Proof. Let T be a contraction in the complete metric space (X, d). Pick an arbitrary $x_0 \in X$ and define a sequence $\{x_n\}$ by setting $x_n = Tx_{n-1} = T^n x_0$, $\forall n \geq 1$. We claim that $\{x_n\}$ forms a Cauchy sequence in X. First of all, by iteration we have

$$\begin{aligned} d(T^{n+1}x_0, T^n x_0) &\leq \gamma d(T^n x_0, T^{n-1}x_0) \\ &\cdot \\ &\cdot \\ &\leq \gamma^n d(Tx_0, x_0). \end{aligned} \tag{3.1}$$

Next, for $n \geq N$ where N is to be specified in a moment, by the triangle inequality,

$$\begin{aligned} d(x_n, x_N) &= d(T^n x_0, T^N x_0) \\ &\leq d(T^n x_0, T^{n-1}x_0) + \cdots + d(T^{N+1}x_0, T^N x_0) \\ &= \sum_{j=0}^{n-N-1} d(T^{N+j+1}x_0, T^{N+j}x_0) \ . \end{aligned}$$

Using (3.1), we have

$$\begin{aligned} d(x_n, x_N) &\leq \sum_{j=0}^{n-N-1} \gamma^{N+j} d(Tx_0, x_0) \\ &\leq \gamma^N d(Tx_0, x_0) \sum_{j=0}^{n-N-1} \gamma^j \\ &< \gamma^N d(Tx_0, x_0) \sum_{j=0}^{\infty} \gamma^j \\ &= \frac{d(Tx_0, x_0)}{1-\gamma} \gamma^N \ . \end{aligned} \tag{3.2}$$

For $\varepsilon > 0$, choose N so large that $d(Tx_0, x_0)\gamma^N/(1-\gamma) < \varepsilon/2$. Then for $n, m \geq N$,

$$\begin{aligned} d(x_n, x_m) &\leq d(x_n, x_N) + d(x_N, x_m) \\ &< \frac{2d(Tx_0, x_0)}{1-\gamma} \gamma^N \\ &< \varepsilon, \end{aligned}$$

thus $\{x_n\}$ forms a Cauchy sequence. As X is complete, $x = \lim_{n\to\infty} x_n$ exists. By the continuity of T, $\lim_{n\to\infty} Tx_n = Tx$. But on the other hand, $\lim_{n\to\infty} Tx_n = \lim_{n\to\infty} x_{n+1} = x$. We conclude that $Tx = x$.

Suppose there is another fixed point $y \in X$. From

$$\begin{aligned} d(x, y) &= d(Tx, Ty) \\ &\leq \gamma d(x, y), \end{aligned}$$

and $\gamma \in (0, 1)$, we conclude that $d(x, y) = 0$, i.e., $x = y$. □

Incidentally, we point out that this proof is a constructive one. It tells you how to find the fixed point starting from an arbitrary point. In fact, letting $n \to \infty$ in (3.2) and then replacing N by n, we obtain an error estimate between the fixed point and the approximating sequence $\{x_n\}$:

$$d(x, x_n) \leq \frac{d(Tx_0, x_0)}{1-\gamma} \gamma^n, \quad n \geq 1.$$

The following two examples demonstrate the sharpness of the Contraction Mapping Principle.

Example 3.3. Consider the map $Tx = x/2$ which maps $(0, 1]$ to itself. It is clearly a contraction. If $Tx = x$, then $x = x/2$ which implies $x = 0$. Thus T does not have a fixed point in $(0, 1]$. This example shows that completeness of the underlying space cannot be removed from the assumption of the theorem.

Example 3.4. Consider the map S from $\mathbb{R}$ to itself defined by

$$Sx = x - \log(1 + e^x) \ .$$

We have

$$\frac{dS}{dx} = \frac{1}{1+e^x} \in (0,1) \ , \quad \forall x.$$

By the Mean-Value Theorem, for some z lying between x and y,

$$|Sx - Sy| = \frac{1}{1+e^z}|x-y| < |x-y| \ .$$

However, in view of $(1+e^z)^{-1} \to 1$ as $x, y \to -\infty$, it is impossible to find a single $\gamma \in (0,1)$ to satisfy

$$|Sx - Sy| \leq \gamma |x-y| \ , \quad \forall x, y \ .$$

It is easy to see that S admits no fixed points. Therefore, the contraction condition cannot be removed from the assumption of the theorem.

Example 3.5. Let $f : [0,1] \to [0,1]$ be a continuously differentiable function satisfying $|f'(x)| < 1$ on $[0,1]$. We claim that f admits a unique fixed point. For, by the Mean-Value Theorem, for $x, y \in [0,1]$ there exists some $z \in (0,1)$ such that $f(y) - f(x) = f'(z)(y-x)$. Therefore,

$$\begin{aligned} |f(y) - f(x)| &= |f'(z)||y-x| \\ &\leq \gamma |y-x|, \end{aligned}$$

where $\gamma = \sup_{t\in[0,1]} |f'(t)| < 1$ (Why?). We see that f is a contraction. By the Contraction Mapping Principle, it has a unique fixed point.

In fact, by using the mean-value theorem one can show that *every continuous function* from $[0,1]$ to itself admits *at least* one fixed point. This is a general fact. More generally, according to Brouwer's Fixed Point Theorem, every continuous maps from a compact convex set in $\mathbb{R}^n$ to itself admits at least one fixed point.

Our applications of the fixed point theorem are mainly on solving equations in certain form. Let us first recall what the meaning of solving an equation is. Here are some examples:

- Solve $2x - 5 = 0$.
- Solve $x^2 - 3x + 5 = 0$.
- Solve $x^{12} - 6x^3 + 6x - 127 = 0$.

- Solve $x - y + 12 = 0,\ 3x + 5y = 0$.
- Solve $x^2 - xy + y^3 - x - 16 = 0,\ x^3 - 13xy + y^2 - 12x = 12.$
- Solve $y' = x^2y^3 + \cos x,\ y(0) = 0.$

All these equations can be formulated as some mappings from a metric space to itself. For instance, in the second case we take $Tx = x^2 - 3x$ and $X = \mathbb{R}$. Then the equation becomes $Tx = -5$, thus solving the equation means to find the preimage of -5 under T. In the fourth case we take $S(x, y) = (x - y, 3x + 5y)$ which maps $\mathbb{R}^2$ to itself. Solving the equation means to determine $S^{-1}(-12, 0)$. There are other choices of the map, say, if we let $S_1(x, y) = (x - y + 12, 3x + 5y)$, then we need to find $S_1^{-1}(0, 0)$. In the sixth case, first observe that it is equivalent to solving the integral equation

$$y(x) = \int_0^x t^2y^3(t)\,dt + \sin x.$$

Hence, taking $\Phi(y) = y(x) - \int_0^x t^2y^3(t)\,dt$ and $X = C[a, b]$, solving the differential equation means to determine $\Phi^{-1}(\sin x)$. Here in addition to the requirement $a < 0 < b$, there are also some technical ones to ensure Φ really maps X to X.

All in all, we have seen that solving equations, algebraic or differential alike, means to determine the preimage of a given map on some metric space.

Example 3.6. Show that the equation

$$x = \frac{1}{2} + \frac{1}{8}\cos 5x\ ,$$

has a unique solution. Well, let us define

$$Tx = \frac{1}{2} + \frac{1}{8}\cos 5x\ .$$

We claim that it is a contraction on $\mathbb{R}$. Indeed,

$$\begin{aligned} |Tx - Tx'| &= \left|\frac{1}{2} + \frac{1}{8}\cos 5x - (\frac{1}{2} + \frac{1}{8}\cos 5x)\right| \\ &\le \frac{1}{8}|\cos 5x - \cos 5x'| \\ &\le \frac{1}{8}|-5\sin 5z(x - x')| \\ &\le \frac{5}{8}|x - x'|\ , \end{aligned}$$

where z lies between x and x'. By appealing to the contraction mapping principle, we conclude that T has a unique fixed point, which is the solution to the equation.

This example is a rather straightforward application of the fixed point theorem. Now we describe a common situation where the theorem can be applied. Let $(X, \|\cdot\|)$ be a normed space and $\Phi : X \to X$ satisfying $\Phi(x_0) = y_0$. We asked: Is it locally solvable? That is, for all y sufficiently near y_0, is there some x close to x_0 so that $\Phi(x) = y$ holds? We have the following result.

Theorem 3.4 (**Perturbation of Identity**). *Let $(X, \|\cdot\|)$ be a Banach space and $\Phi : \overline{B_r(x_0)} \to X$ satisfies $\Phi(x_0) = y_0$. Suppose that Φ is of the form $I + \Psi$ where I is the identity map and Ψ satisfies*

$$\|\Psi(x_2) - \Psi(x_1)\| \leq \gamma\|x_2 - x_1\| \ , \quad x_1, x_2 \in \overline{B_r(x_0)}, \quad \gamma \in (0,1) \ .$$

Then for $y \in \overline{B_R(y_0)}, R = (1-\gamma)r$, there is a unique $x \in \overline{B_r(x_0)}$ satisfying $\Phi(x) = y$.

The idea of the following proof can be explained in a few words. Taking $x_0 = y_0 = 0$ for simplicity, we would like to find x solving $x + \Psi(x) = y$. This is equivalent to finding a fixed point for the map T, $Tx + \Psi(x) = y$, that is, $Tx = y - \Psi(x)$. By our assumption, Ψ is a contraction, so is T.

Proof. We first shift the points x_0 and y_0 to 0 by redefining Φ. Indeed, for $x \in \overline{B_r(0)}$, let

$$\widetilde{\Phi}(x) = \Phi(x + x_0) - \Phi(x_0) = x + \Psi(x + x_0) - \Psi(x_0) \ .$$

Then $\widetilde{\Phi}(0) = 0$. Consider this map on $\overline{B_r(0)}$ given by

$$Tx = x - (\widetilde{\Phi}(x) - y), \quad y \in \overline{B_R(0)} \ .$$

We would like to verify that T is a well-defined contraction on $\overline{B_r(0)}$. First, we claim that T maps $\overline{B_r(0)}$ into itself. Indeed,

$$\begin{aligned} \|Tx\| &= \|x - (\widetilde{\Phi}(x) - y)\| \\ &= \|\Psi(x_0) - \Psi(x_0 + x) + y\| \\ &\leq \|\Psi(x_0 + x) - \Psi(x_0)\| + \|y\| \\ &\leq \gamma\|x\| + R \\ &\leq r \ . \end{aligned}$$

Next, we claim that T is a contraction. Indeed,

$$\begin{aligned} \|Tx_2 - Tx_1\| &= \|\Psi(x_1 + x_0) - \Psi(x_2 + x_0)\| \\ &\leq \gamma\|x_2 - x_1\| \ . \end{aligned}$$

As $\overline{B_r(0)}$ is a closed subset of the complete space X, it is also complete. The Contraction Mapping Principle can be applied to conclude that for each $y \in \overline{B_R(0)}$, there is a unique fixed point for T, $Tx = x$, in $\overline{B_r(0)}$. In other words, $\widetilde{\Phi}(x) = y$ for a unique $x \in \overline{B_r(0)}$. The desired conclusion follows after going back to Φ. □

Remark 3.1. (a) It suffices to assume Ψ is a contraction on $B_r(x_0)$ in the theorem. As a contraction is uniformly continuous, it extends to become a contraction with the same contraction constant in $\overline{B_r(x_0)}$, see exercise.

(b) By examining the proof above, one can see that the fixed point $x \in B_r(x_0)$ whenever $y \in B_R(y_0)$. Indeed, when $y \in B_R(0)$, that is, $\|y\| < R$,

$$\begin{aligned} \|Tx\| &= \|x - (\widetilde{\Phi}(x) - y)\| \\ &= \|\Psi(x_0) - \Psi(x_0 + x) + y\| \\ &\leq \|\Psi(x_0 + x) - \Psi(x_0)\| + \|y\| \\ &< \gamma\|x\| + R \\ &\leq r \ . \end{aligned}$$

It follows that the preimage x which satisfies $Tx = x$ belongs to $B_r(0)$.

(c) The inverse map that sends $y \in \overline{B_R(y_0)}$ back to $x \in \overline{B_r(x_0)}$, the fixed point of T, is well-defined. Denote it by Φ^{-1}. We claim that it is continuous. For, let $y_1, y_2 \in B_R(y_0)$. Then $x_i = \Phi^{-1}(y_i)$ satisfy $x_i = y_i - \Psi(x_i), i = 1, 2$, that is,

$$\begin{aligned} \|\Phi^{-1}(y_1) - \Phi^{-1}(y_2)\| &= \|y_1 - \Psi(x_1) - (y_2 - \Psi(x_2)\| \\ &\leq \|y_1 - y_2\| + \|\Psi(x_2) - \Psi(x_1)\| \\ &\leq \|y_1 - y_2\| + \gamma\|x_1 - x_2\| \\ &= \|y_1 - y_2\| + \gamma\|\Phi^{-1}(y_1) - \Phi^{-1}(y_2)\| \ , \end{aligned}$$

which implies

$$\|\Phi^{-1}(y_1) - \Phi^{-1}(y_2)\| \leq \frac{1}{1-\gamma}\|y_1 - y_2\| \ .$$

It follows that Φ^{-1} is uniformly continuous (in fact, "Lipschitz continuous") in $\overline{B_R(y_0)}$.

Obviously, the terminology "perturbation of identity" comes from the expression

$$\widetilde{\Phi}(x) = \Phi(x + x_0) - \Phi(x_0) = x + \Psi(x + x_0) - \Psi(x_0) \ ,$$

which is in form of the identity plus a term satisfying the "smallness condition"

$$|\Psi(x + x_0) - \Psi(x_0)| \leq \gamma|x| \ , \quad \gamma \in (0, 1) \ .$$

Example 3.7. Show that the equation $3x^4 - x^2 + x = -0.05$ has a real root. We look for a solution near 0. Let X be $\mathbb{R}$ and $\Phi(x) = x + \Psi(x)$ where $\Psi(x) = 3x^4 - x^2$ so that $\Phi(0) = 0$. According to the theorem, we need to find some r so that Ψ becomes a contraction. For $x_1, x_2 \in \overline{B_r(0)}$, that is, $x_1, x_2 \in [-r, r]$, , we have

$$\begin{aligned} |\Psi(x_1) - \Psi(x_2)| &= |(3x_2^4 - x_2^2) - (3x_1^4 - x_1^2)| \\ &\leq (3|x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3| + |x_2 + x_1|)|x_2 - x_1| \\ &\leq (12r^3 + 2r)|x_2 - x_1| \ , \end{aligned}$$

which is a contraction as long as $\gamma = (12r^3+2r) < 1$. Taking $r = 1/4$, then $\gamma = 11/16 < 1$ will do the job. Then $R = (1-\gamma)r = 5/64 \sim 0.078$. We conclude that for all numbers $b, |b| < 5/64$, the equation $3x^4 - x^2 + x = b$ has a unique root in $(-1/4, 1/4)$. Now, -0.05 falls into this range, so the equation has a root.

Example 3.8. Solve

$$x - 3\sin^2(x-1) = 1.01 \ .$$

Here we take $\Phi(x) = x - 3\sin^2(x-1)$ and $x_0 = 1, y_0 = 1$. Using $\sin^2(x-1) - \sin^2(x'-1) = 2\sin(z-1)\cos(z-1)(x-x')$ where z lies between x and x', we have

$$|-3\sin^2(x-1) + 3\sin^2(x'-1)| \le 3\sin 2(z-1)|x-x'| \ .$$

When $|x-1|, |x'-1| \le r$, $|\sin 2(z-1)| \le 2r$. Therefore,

$$|-3\sin^3(x-1) + 3\sin(x'-1)| \le 6r|x-x'| \ .$$

We take $r = 1/7$ so that $R = (1-6/7)1/7 = 1/49$. We conclude that the equation has a unique solution $x \in [1-1/7, 1+1/7]$ whenever $y \in [1-1/49, 1+1/49] \sim [1-0.02, 1+0.02]$, so it applies to $y = 1.01$.

The same method can be applied to solving systems of equations in $\mathbb{R}^n$. We formulate it as a general result.

Proposition 3.5. *Let $\Phi = x + \Psi(x) : U \to \mathbb{R}^n$ be C^1 where U is an open set in $\mathbb{R}^n$ containing $0, \Psi(0) = 0$ and $\nabla\Psi(0) = 0$. Then there is some $r > 0$ such that $\Psi(x) = y$ has a unique solution in $\overline{B_r(0)}$ for each y in $\overline{B_R(0)}, R = r/2$.*

Proof. It suffices to verify that Ψ is a contraction on $B_r(0)$ for sufficiently small r. Then we can apply the theorem on perturbation of identity to obtain the desired result. To this end, we fix $x_1, x_2 \in B_r(0)$ where r is to be determined and consider the function $\varphi(t) = \Psi_i(x_1 + t(x_2 - x_1))$. We have $\varphi(0) = \Psi_i(x_1)$ and $\varphi(1) = \Psi_i(x_2)$. By the mean value theorem, there is some $t^* \in (0,1)$ such that $\varphi(1) - \varphi(0) = \varphi'(t^*)(1-0) = \varphi'(t^*)$. By the Chain Rule,

$$\begin{aligned}
\varphi'(t) &= \frac{d}{dt}\Psi_i(x_1 + t(x_2 - x_1)) \\
&= \frac{\partial\Psi_i}{\partial x_1}(x_1 + t(x_2 - x_1))(x_{21} - x_{11}) + \cdots + \frac{\partial\Psi_i}{\partial x_n}(x_1 + t(x_2 - x_1))(x_{2n} - x_{1n}) \\
&= \sum_{j=1}^{n} \frac{\partial\Psi_i}{\partial x_j}(x_1 + t(x_2 - x_1))(x_{2j} - x_{1j}) \ .
\end{aligned}$$

Setting $z = x_1 + t^*(x_2 - x_1)$, we have

$$\Psi_i(x_2) - \Psi_i(x_1) = \sum_{j=1}^{n} \frac{\partial\Psi_i}{\partial x_j}(z)(x_{2j} - x_{1j}) \ .$$

Recalling that for the equation $y = Ax$ where $A = \{a_{ij}\}$ is an $n \times n$-matrix and $x, y \in \mathbb{R}^n$, by Cauchy-Schwarz inequality the following inequality holds:

$$|y| \leq \sqrt{\sum_{i,j} a_{ij}^2}\ |x| \quad .$$

Applying this inequality to our situation, we have

$$|\Psi(x_1) - \Psi(x_2)| \leq M|x_1 - x_2| \ ,$$

where

$$M = \sup_{|z| \leq r} \sqrt{\sum_{i,j} \left(\frac{\partial \Psi_i}{\partial x_j}(z) \right)^2} \ .$$

Using the assumptions $\nabla\Psi(0) = 0$ and Ψ is C^1, we can find some small r such that $M = 1/2$. Applying the theorem of perturbation of identity, the equation $\Phi(x) = y$ is uniquely solvable for $y \in \overline{B_R(0)}, R = (1 - 1/2)r = r/2$ with solution $x \in \overline{B_r(0)}$.

□

Theorem 3.4 is also applicable to function spaces. Let us example the following example.

Example 3.9. Consider the integral equation

$$y(x) = tg(x) + \int_0^1 K(x,t)y^2(t)dt \ ,$$

where $K(x,t) \in C([0,1]^2), g \in C[0,1]$ are given and t is a small parameter. We would like to show that it admits a solution y as long as t is small in some sense. Our first job is to formulate this problem as a problem of perturbation of identity. We work on the Banach space $C[0,1]$ and let

$$\Phi(y)(x) = y(x) - \int_0^1 K(x,t)y^2(t)dt \ .$$

That is,

$$\Psi(y)(x) = -\int_0^1 K(x,t)y^2(t)dt \ .$$

We further choose x_0 to be 0, the zero function, so $y_0 = \Phi(0) = 0$. Then, for $y_2, y_1 \in \overline{B_r(0)}$ (r to be specified later),

$$\begin{aligned} \|\Psi(y_2) - \Psi(y_1)\|_\infty &\leq \int_0^1 |K(x,t)||y_2^2 - y_1^2|(t)dt \\ &\leq M \times 2r\|y_2 - y_1\|_\infty, \quad M = \max\{|K(x,t) : \ (x,t) \in [0,1]^2\} \ , \end{aligned}$$

which shows that Ψ is a contraction as long as $\gamma = 2Mr < 1$. Under this condition, we may apply Theorem 3.4 to conclude that for all t such that $t\|g\| \leq R, R = (1-\gamma)r$, the integral equation

$$y(x) - \int_0^1 K(x,t)y^2(t)dt = tg(x) \ ,$$

has a unique solution $y \in \overline{B_r(0)}$. For instance, we fix $r = 1/(4M)$ so that $2Mr = 1/2$ and $R = 1/(8M)$. This integral equation is solvable for g as long as $|t| < 1/(8M\|g\|)$.

You should be aware that in these two examples, the first underlying space is the Euclidean space and the second one is the space of continuous functions under the supnorm. It shows the power of abstraction, that is, the fixed point theorem applies to all complete metric spaces.

3.3 The Inverse Function Theorem

We start by recalling two old results.

First, the general chain rule.

Let $F: U \to \mathbb{R}^m$ and $G: V \to \mathbb{R}^l$ where U is open in $\mathbb{R}^n$ and V open in $\mathbb{R}^m$ and $F(U) \subset V$. Assume the partial derivatives of F and G exist in U and V respectively. The Chain Rule asserts that their composition $H = G \circ F: U \to \mathbb{R}^l$ also has partial derivatives in U. Moreover, letting $F = (F_1, \cdots, F_m)$, $G = (G_1, \cdots, G_l)$ and $H = (H_1, \cdots, H_l)$. From

$$H_k(x_1, \cdots, x_n) = H_k(F_1(x), \cdots, F_m(x)), \quad k = 1, \cdots, l,$$

we have

$$\frac{\partial H_k}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial G_k}{\partial y_i}\frac{\partial F_i}{\partial x_j} \ .$$

It is handy to write things in matrix form. Let DF be the Jacobian matrix of F, that is,

$$DF = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}$$

and similarly for DG and DH. Then the formula above becomes, in matrix product,

$$DH(x) = DG(F(x))DF(x) \ .$$

Next, the mean-value theorem in one-dimensional case reads as $f(y) - f(x) = f'(c)(y - x)$ for some value c lying between x and y. To remove the uncertainty of c, we note the alternative formula

$$f(y) = f(x) + \int_0^1 f'(x + t(y - x))\, dt\ (y - x)\ ,$$

which is obtained from

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \frac{d}{dt} f(x + t(y - x))\, dt \quad \text{(Fundamental Theorem of Calculus)} \\ &= f(x) + \int_0^1 f'(x + t(y - x))\, dt\ (y - x) \quad \text{(Chain Rule)}\ . \end{aligned}$$

We will need its n-dimensional version.

Proposition 3.6. *Let $F : B \to \mathbb{R}^n$ be C^1 where B is a ball in $\mathbb{R}^n$. For $x_1, x_2 \in B$,*

$$F(x_2) - F(x_1) = \int_0^1 DF(x_1 + t(x_2 - x_1))\ dt \cdot (x_2 - x_1)\ .$$

Here $F(x_2) - F(x_1)$ and $x_2 - x_1$ are viewed as column vectors. Componentwise this means

$$F_i(x_2) - F_i(x_1) = \sum_{j=1}^{n} \int_0^1 \frac{\partial F_i}{\partial x_j}(x_1 + t(x_2 - x_1))dt\ (x_{2j} - x_{1j}), \quad i = 1, \cdots, n\ .$$

Proof. Applying Chain Rule to each function F_i, we have

$$\begin{aligned} F_i(x_2) - F_i(x_1) &= \int_0^1 \frac{d}{dt} F_i(x_1 + t(x_2 - x_1))dt \\ &= \int_0^1 \sum_j \frac{\partial F_i}{\partial x_j}(x_1 + t(x_2 - x_1))(x_{2j} - x_{1j})dt \\ &= \int_0^1 DF(x_1 + t(x_2 - x_1))\ dt \cdot (x_2 - x_1)\ . \end{aligned}$$

□

The Inverse Function Theorem and Implicit Function Theorem play a fundamental role in analysis and geometry. They illustrate the principle of linearization which is ubiquitous in mathematics. We learned these theorems in advanced calculus but the proofs were not emphasized. Now we fill out the gap.

All is about linearization. Recall that a real-valued function on an open interval I is differentiable at some $x_0 \in I$ if there exists some $a \in \mathbb{R}$ such that

$$\lim_{x \to x_0} \left| \frac{f(x) - f(x_0) - a(x - x_0)}{x - x_0} \right| = 0.$$

In fact, the value a is equal to $f'(x_0)$, the derivative of f at x_0. We can rewrite the limit above using the little o notation:

$$f(x_0 + z) - f(x_0) = f'(x_0)z + \circ(z), \quad \text{as } z \to 0.$$

Here $\circ(z)$ denotes a quantity satisfying $\lim_{z\to 0} \circ(z)/|z| = 0$. The same situation carries over to a real-valued function f in some open set in $\mathbb{R}^n$. A function f is called differentiable at x_0 in this open set if there exists a vector $a = (a_1, \cdots, a_n)$ such that

$$f(x_0 + x) - f(x_0) = \sum_{j=1}^{n} a_j x_j + \circ(|x|) \quad \text{as } x \to 0.$$

Note that here $x_0 = (x_0^1, \cdots, x_0^n)$ is a vector. Again one can show that the vector a is uniquely given by the gradient vector of f at x_0

$$\nabla f(x_0) = \left(\frac{\partial f}{\partial x_1}(x_0), \cdots, \frac{\partial f}{\partial x_n}(x_0) \right).$$

More generally, a map F from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ is called differentiable at a point x_0 in this open set if each component of $F = (f^1, \cdots, f^m)$ is differentiable. We can write the differentiability condition collectively in the following form

$$F(x_0 + x) - F(x_0) = DF(x_0)x + o(x), \tag{3.3}$$

where $DF(x_0)$ is the linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ given by

$$(DF(x_0)z)_i = \sum_{j=1}^{n} a_{ij}(x_0)x_j, \quad i = 1, \cdots, m,$$

where $\left(a_{ij}\right) = \left(\partial f^i/\partial x_j\right)$ is the Jabocian matrix of f. (3.4) shows near x_0, that is, when x is small, the function F is well-approximated by the linear map plus a constant $F(x_0) + DF(x_0)x$ as long as $DF(x_0)$ is invertible (i.e., nonsingular). It suggests that the local information of a map at a differentiable point could be retrieved from its a linear map, which is much easier to analyse. This principle, called linearization, is widely used in analysis. The Inverse Function Theorem is a typical result of linearization. It asserts that a map is locally invertible if its linearization is invertible. Therefore, local bijectivity of the map is ensured by the invertibility of its linearization. When $DF(x_0)$ is not invertible, the first term on the right hand side of (3.4) may degenerate in some or even all direction so that $DF(x_0)x$ cannot control the error term $\circ(z)$. In this case the local behavior of F may be different from its linearization.

Theorem 3.7 (**Inverse Function Theorem**). *Let $F : U \to \mathbb{R}^n$ be a C^1-map where U is open in $\mathbb{R}^n$ and $x_0 \in U$. Suppose that $DF(x_0)$ is invertible.*

(a) There exist open sets V and W containing x_0 and $F(x_0)$ respectively such that the restriction of F on V is a bijection onto W with a C^1-inverse.

(b) The inverse is C^k when F is $C^k, 1 \le k \le \infty$, in V.

A map from some open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ is C^k, $1 \le k \le \infty$, if all its components belong to C^k. It is called a C^∞-map or a smooth map if its components are C^∞. Similarly, a matrix is C^k or smooth if its entries are C^k or smooth accordingly.

The condition that $DF(x_0)$ is invertible, or equivalently the non-vanishing of the determinant of the Jacobian matrix, is called the **nondegeneracy condition**. Without this condition, the map may or may not be local invertible, see the examples below.

When the inverse is differentiable, we may apply this chain rule to differentiate the relation $F^{-1}(F(x)) = x$ to obtain

$$DF^{-1}(y_0)\ DF(x_0) = I\ , \quad y_0 = F(x_0),$$

where I is the identity map. We conclude that

$$DF^{-1}(y_0) = \big(DF(x_0)\big)^{-1}.$$

In other words, the matrix of the derivative of the inverse map is precisely the inverse matrix of the derivative of the map. We conclude that although the inverse may exist without the non-degeneracy condition. This condition is necessary in order to have a *differentiable* inverse. We single it out in the following proposition.

Proposition 3.8. *Let $F : U \to \mathbb{R}^n$ be a C^1-map and $x_0 \in U$. Suppose for some open V in U containing x_0, F is invertible in V with a differentiable inverse. Then $DF(x_0)$ is non-singular.*

Now we prove Theorem 3.7. At first sight it is not clear how to link this theorem to the Theorem of Perturbation of Identity. The ideas of the proof is as follows. Taking $x_0, y_0 = 0$, formally we have $F(x) = F(0) + DF(0)(x-0) + \frac{1}{2}D^2F(0)(x-0)^2 + \cdots$. Hence solving $F(x) = y$ is the same as solving $DF(0)x + \frac{1}{2}D^2F(0)x^2 + \cdots = y$. Since $DF(0)$ is invertible, it is equivalent to solving $x + DF(0)^{-1}(\frac{1}{2}D^2F(0)x^2 + \cdots) = DF(0)^{-1}y$, and this is in the form of $x + \Psi(x) = y$.

Now let us turn to the proof. First assume that $x_0 = y_0 = 0$ and $DF(0) = I$, the identity matrix. We write $F(x) = y$ as $x + \Psi(x) = y$ where $\Psi(x) = F(x) - x$ and apply Theorem 3.4. For this purpose we need to verify Ψ is a contraction. First fix a ball $B_{r_0}(0)$ satisfying $\overline{B_{r_0}(0)} \subset U$. As U is open and $0 \in U$, this is always possible. For $x_1, x_2 \in \overline{B_{r_0}(0)}$, we have, by Proposition 3.6,

$$\begin{aligned} |\Psi(x_1) - \Psi(x_2)| &= |F(x_1) - x_1 - (F(x_2) - x_2)| \\ &= |(F(x_1) - F(x_2) - (x_1 - x_2))| \\ &= |B \cdot (x_1 - x_2)|\ , \end{aligned}$$

where the matrix B is given by

$$B = \int_0^1 (DF(x_2 + t(x_1 - x_2)) - DF(0))\, dt \ ,$$

where we have used the assumption $DF(0) = I$. The ij-th entry of $B = (b_{ij})$ is given by

$$b_{ij} = \int_0^1 \left(\frac{\partial F_i}{\partial x_j}(x_2 + t(x_1 - x_2)) - \frac{\partial F_i}{\partial x_j}(0) \right) dt \ .$$

By the continuity of $\partial F_i/\partial x_j$ at 0, given $\varepsilon > 0$, there is some $r \leq r_0$ such that

$$\left| \frac{\partial F_i}{\partial x_j}(x) - \frac{\partial F_i}{\partial x_j}(0) \right| < \varepsilon \ , \quad \forall x \in \overline{B_r(0)} \ .$$

As $x_2 + t(x_1 - x_2) \in \overline{B_r(0)}$ whenever $x_1, x_2 \in \overline{B_r(0)}$,

$$\left| \frac{\partial F_i}{\partial x_j}(x_2 + t(x_1 - x_2)) - \frac{\partial F_i}{\partial x_j}(0) \right| < \varepsilon \ , \quad x_1, x_2 \in \overline{B_r(0)} \ .$$

It follows that

$$\begin{aligned} |\Psi(x_1) - \Psi(x_2)| &= |B \cdot (x_1 - x_2)| \\ &\leq \sqrt{\sum_{i,j} b_{ij}^2}\ |x_1 - x_2| \\ &\leq \sqrt{n^2 \varepsilon^2} |x_1 - x_2| \\ &= n\varepsilon |x_1 - x_2| \ . \end{aligned}$$

Now, by choosing ε to be $1/2n$, we find some $r \leq r_0$ such that

$$|\Psi(x_1) - \Psi(x_2)| \leq \frac{1}{2} |x_2 - x_1| \ , \quad \forall x_1, x_2 \in \overline{B_r(0)} \ ,$$

that is, Ψ is a contraction with $\gamma = 1/2$.

By Theorem 3.4 and Remark 3.1, we conclude that $F(x) = y$ is uniquely solvable for $y \in B_R(0), R = (1 - 1/2)r = r/2$, with $x \in B_r(0)$. Moreover, the inverse of F, G, is continuous from $B_R(0)$ back to $B_r(0)$ whose image $G(B_R(0))$ is an open set in $B_r(0)$. Indeed, from Remark 3.1 (c) we have, since $1/(1 - \gamma) = 2$,

$$|G(y_1) - G(y_2)| \leq 2|y_1 - y_2| \ , \quad y_1, y_2 \in B_R(0) \ . \tag{3.4}$$

It remains to establish the differentiability of the inverse map. As by assumption, DF is invertible at 0, we may further restrict r so that DF is invertible in $B_r(0)$. We take $W = B_R(0)$ and $V = G(B_R(0))$ and claim that the partial derivatives of G exist in $B_R(0)$.

To this end we recall the following fact: The partial derivatives of a function Φ exist at x_0 if there is an $n \times n$-matrix A such that

$$\Phi(x_0 + x) - \Phi(x_0) = Ax + R \ ,$$

where $R = \circ(|x|)$. Moreover, when this happens, $D\Phi(x_0) = A$. (see Remark 3.2 below.) Here we are concerned with $\Phi = G$.

Let $y \in B_R(0)$ and $y' \in B_R(0)$ close to y. Let x and x' be their respective preimages in $B_r(0)$ under F. We have

$$F(x') = F(x) + DF(x)(x' - x) + \circ(|x' - x|) \ .$$

Writing it in terms of y,

$$y' = y + DF(G(y))(G(y') - G(y)) + \circ(|x' - x|) \ .$$

Here $\circ(|x' - x|)$ is a quantity which satisfies $\circ(|x' - x|)/|x' - x|$ as $x' \to x$. Now, since G is continuous, as $y' \to y$, $x' = G(y') \to x = G(y)$, and, in view of (3.4),

$$\frac{\circ(|x' - x|)}{|y' - y|} = \frac{\circ(|x' - x|)}{|x' - x|} \times \frac{|G(y') - G(y)|}{|y' - y|} \to 0 \ , \quad \text{as } y' \to y \ .$$

So we can write

$$y' - y = DF(G(y))(G(y') - G(y)) + \circ(|y' - y|) \ .$$

Since $DF(x)$ is invertible in $B_r(0)$,

$$(DF(G(y))^{-1}(y' - y) = G(y') - G(y) + \circ(|y' - y|) \ ,$$

that is,

$$G(y') - G(y) = (DF(G(y))^{-1}(y' - y) + \circ(|y' - y|) \ .$$

We conclude that G is differentiable in $B_R(0)$ and $DG(y) = (DF(G(y))^{-1}$.

From linear algebra we know that each entry of $DG(y)$ can be expressed as a rational function of the entries of the matrix of $DF(G(y))$. Consequently, $DG(y)$ is C^k in y if $DF(G(y))$ is C^k for $1 \leq k \leq \infty$.

So far we have been assuming $x_0 = y_0 = 0$ and $DF(0) = I$. For a general F and x_0, y_0, set

$$\tilde{F}(x) = A(F(x + x_0) - y_0) \ ,$$

where $A = (DF)^{-1}(x_0)$. Then $\tilde{F}$ is a C^1-map in the open set $\tilde{U} \equiv U - x_0$ and it satisfies $\tilde{F}(0) = 0$ and $(D\tilde{F})^{-1}(0) = I$. By what has been done, $\tilde{F}$ admits an inverse $\tilde{G}$ from some open set $\tilde{W}$ containing 0 to an open set $\tilde{V}$ containing 0 in $\tilde{U}$. Letting $V = \tilde{V} + x_0$ and $W = A^{-1}\tilde{W} + y_0$, then V and W are open sets containing x_0 and y_0 respectively. Define

$$G(y) = \tilde{G}(A(y - y_0)) + x_0,$$

where maps W bijectively onto V. We claim that $F(G(y)) = y$ for For $y \in W$. For, observe that

$$F(x) = A^{-1}\tilde{F}(x - x_0) + y_0 \ , \quad x \in V \ .$$

We have

$$\begin{aligned} F(G(y)) &= A^{-1}\tilde{F}(G(y) - x_0) + y_0 \\ &= A^{-1}\tilde{F}(\tilde{G}(A(y - y_0)) + x_0 - x_0) + y_0 \\ &= y \ . \end{aligned}$$

Finally, observe that G is C^k in W as long as $\tilde{G}$ is C^k in $\tilde{W}$. The proof of the Inverse Function Theorem is completed.

Remark 3.2. Recall that given a function $\varphi : U \to \mathbb{R}$ where $U \subset \mathbb{R}^n$ is open and $x_0 \in U$. The partial derivative of φ at x_0 exists if there is some $\alpha \in \mathbb{R}^n$ such that

$$\varphi(x_0 + x) - \varphi(x_0) = \sum_{j=1}^{n} \alpha_j x_j + \circ(|x|) \ .$$

When this happens, $\partial\varphi/\partial x_j(x_0) = \alpha_j$. For $\Phi : U \to \mathbb{R}^n$, applying this fact to each component of Φ, Φ_i, we see that the Jacobian matrix $D\Phi$ at x_0 exists if there is a matrix $A = \{\alpha_{ij}\}$ such that

$$\Phi(x_0 + x) - \Phi(x_0) = Ax + \circ(|x|) \ .$$

When this happens, $\partial\Phi_i/\partial x_j(x_0) = \alpha_{ij}$.

Example 3.10. The Inverse Function Theorem asserts a local invertibility. Even if the linearization is non-singular everywhere, we cannot assert global invertibility. Consider

$$x = e^t \cos\theta, \quad y = e^t \sin\theta \ .$$

The function $F : \mathbb{R}^2 \to \mathbb{R}^2$ given by $F(t, \theta) = (x, y)$ is a continuously differentiable function whose Jacobian matrix is non-singular everywhere. However, it is clear that F is not bijective, for instance, all points $(t, \theta + 2n\pi), n \in \mathbb{Z}$, have the same image under F.

Example 3.11. An exceptional case is dimension one where a global result is available. Indeed, in 2060 we learned that if f is continuously differentiable on (a, b) with non-vanishing f', it is either strictly increasing or decreasing so that its global inverse exists and is again continuously differentiable.

Example 3.12. Consider the map $F : \mathbb{R}^2 \to \mathbb{R}^2$ given by $F(x, y) = (x^2, y)$. Its Jacobian matrix is singular at $(0, 0)$. In fact, for any point $(a, b), a > 0$, $F(\pm\sqrt{a}, b) = (a, b)$. We cannot find any open set, no matter how small is, at $(0, 0)$ so that F is injective. On the other hand, the map $H(x, y) = (x^3, y)$ is bijective with inverse given by $J(x, y) = (x^{1/3}, y)$. However, as the non-degeneracy condition does not hold at $(0, 0)$ so it is not differentiable there. In these cases the Jacobian matrix is singular, so the nondegeneracy condition does not hold.

Next we discuss the Implicit Function Theorem. The simplest situation of this general theorem concerns the zero set (or locus) of a single function $f(x, y)$ in the plane. Namely, when is the zero set $\{(x, y) : \ f(x, y) = 0\}$ a curve? Consider the function $x^2 + y^2 - 1$ where the zero set is the unit circle. It is easy to see that it is a curve. Moreover, when $(x, y) \neq (\pm 1, 0)$, it is a graph over the x-axis, and, when $(x, y) \neq (0, \pm 1)$, it is a graph over the y-axis.

Theorem 3.9. *Let f be a C^1-function in some open U in the plane and $f(x_0, y_0) = 0$. Suppose that $f_y(x_0, y_0) \neq 0$, then there is some open interval $I, x_0 \in I$, an open set V containing (x_0, y_0) in U and a C^1-function φ on I whose graph lies in V such that $\{(x, y) \in V : \ f(x, y) = 0\} = \{(x, \varphi(x)) : x \in I\}$.*

In other words, the zero set of $f(x, y) = 0$ near (x_0, y_0) is given by the graph $(x, \varphi(x))$, hence it is a curve. Likewise, when $f_x(x_0, y_0) \neq 0$, the locus is locally given the graph $\{\psi(y), y) : \ y \in J\}$ for some interval J containing y_0 and a C^1-function ψ on J satisfying $\psi(y_0) = x_0$.

Proof. Define $\Phi(x, y) = (x, f(x, y))$. Then $\Phi(x_0, y_0) = (x_0, 0)$ and $\det D\Phi(x_0, y_0) = f_y(x_0, y_0) \neq 0$. By Inverse Function Theorem, Φ has a C^1-inverse Ψ from some open set W containing $(x_0, 0)$ satisfying $\Phi(\Psi(x, z)) = (x, z)$ on W. By shrinking W a bit, we may assume $W = I \times J$ for two intervals. Writing $\Psi(x, z) = (\Psi_1(x, z), \Psi_2(x, z))$, we have

$$\Phi(\Psi_1(x, z), \Psi_2(x, z)) = (x, z).$$

On the other hand, from the definition of Φ,

$$\Phi(\Psi_1(x, z), \Psi_2(x, z)) = (\Psi_1(x, z), f(\Psi_1(x, z), \Psi_2(x, z))).$$

By comparing the two components, we have $\Psi_1(x, z) = x$ and $f(\Psi_1(x, z), \Psi_2(x, z)) = z$. It follows that $f(x, \Psi_2(x, z)) = z$. Thus each horizontal line $I \times \{z\}, z \in J = (c, d)$ is mapped to a curve $(x, \Psi_2(x, z))$. By restricting U we may assume $f_y(x, y) \neq 0$ for all $(x, y) \in U$. When $f_y > 0$, the horizontal line $I \times \{c\}$ and $I \times \{d\}$ are mapped to the curves $(x, \Psi_2(x, c))$ and $(x, \Psi_2(x, d))$ respectively with $\Psi_2(x, c) < \Psi_2(x, d)$. (When $f_y(x_0, y_0) < 0, \Psi_2(x, c) > \Psi_2(x, d)$.) Thus the image of $I \times J$ under Ψ is precisely the set bounded by $x = a, b$ and the two curves $(x, \Psi_2(x, c))$ and $(x, \Psi_2(x, d))$. In particular, at $z = 0$, we have $f(x, \Psi_2(x, 0)) = 0$. Our desired conclusion follows by taking $\varphi(x) = \Psi_2(x, 0)$. □

Let us look at some examples.

Example 3.13. First, consider the function $f_1(x, y) = x - y^2 + 3$. We have $f_1(-3, 0) = 0$ and $f_{1x}(-3, 0) = 1 \neq 0$. By Implicit Function Theorem, the zero set of F_1 can be described near $(-3, 0)$ by a function $x = \varphi(y)$ near $y = 0$. Indeed, by solving the equation $f_1(x, y) = 0$, $\varphi(y) = y^2 - 3$. On the other hand, $f_{1y}(-3, 0) = 0$ and from the formula $y = \pm\sqrt{x + 3}$ we see that the zero set is not a graph over an open interval containing -3.

Next we consider the function $f_2(x,y) = x^2 - y^2$ at $(0,0)$. We have $f_{2x}(0,0) = F_{2y}(0,0) = 0$. Indeed, the zero set of f_2 consists of the two straight lines $x = y$ and $x = -y$ intersecting at the origin. It is impossible to express it as the graph of a single function near the origin.

Finally, consider the function $f_3(x,y) = x^2 + y^2$ at $(0,0)$. We have $f_{3x}(0,0) = F_{3y}(0,0) = 0$. Indeed, the zero set of f_3 degenerates into a single point $\{(0,0)\}$ which cannot be the graph of any function.

Next we state the general Implicit Function Theorem.

Theorem 3.10 (**Implicit Function Theorem**). *Consider C^1-map $F : U \to \mathbb{R}^m$ where U is an open set in $\mathbb{R}^n \times \mathbb{R}^m$. Suppose that $(x_0, y_0) \in U$ satisfies $F(x_0, y_0) = 0$ and $D_y F(x_0, y_0)$ is invertible in $\mathbb{R}^m$. There is an open set G in $\mathbb{R}^n$ containing x_0, an open set V containing (x_0, y_0) in U, and a C^1-map φ on G whose graph lies in V such that*

$$\{(x,y) \in V : \ F(x,y) = 0\} = \{(x, \varphi(x)) : \ x \in G\} \ .$$

The map φ belongs to C^k on G when F is $C^k, 1 \le k \le \infty$, in U.

In words, this theorem asserts that near (x_0, y_0), the locust of F is given by the graph of φ.

The notation $D_y F(x_0, y_0)$ stands for the Jocabian matrix $(\partial F_i / \partial y_j(x_0, y_0))_{i,j=1,\cdots,m}$ where x_0 is fixed. In general, a version of Implicit Function Theorem holds when the rank of DF at a point is m. In this case, we can rearrange the independent variables to make $D_y F$ non-singular at this point.

The proof of the general case is essentially the same as the proof of the simplest case.

Proof. Consider $\Phi : U \to \mathbb{R}^n \times R^m$ given by

$$\Phi(x,y) = (x, F(x,y)).$$

One readily checks that $\det D\Phi(x,y) = \det D_y F(x,y)$, so $\det D\Phi(x_0, y_0) \neq 0$. By the Inverse Function Theorem, there exists a C^1-inverse $\Psi = (\Psi_1, \Psi_2)$ from some open set W in $\mathbb{R}^n \times \mathbb{R}^m$ containing $\Phi(x_0, y_0) = (x_0, 0)$ to an open subset of U. By restricting W further we may assume W is of the form $V_1 \times V_2$ where V_1 and V_2 are rectangles centered at x_0 and 0 respectively. We have

$$\Phi(\Psi_1(x,z), \Psi_2(x,z)) = (x,z), \quad (x,z) \in V_1 \times V_2 \ .$$

On the other hand, the definition of Φ gives $\Phi(\Psi_1(x,z), \Psi_2(x,z)) = (\Psi_1(x,z), F(\Psi_1(x,z), \Psi(x,z)))$. Therefore,

$$\Psi_1(x,z) = x, \text{ and } F((\Psi_1(x,z), \Psi_2(x,z)) = z.$$

In other words, $F(x, \Psi_2(x, z)) = z$ holds for $(x, z) \in V_1 \times V_2$. In particular, taking $z = 0$ gives

$$F(x, \Psi_2(x, 0)) = 0, \quad \forall x \in V_1 \ ,$$

so the function $\varphi(x) \equiv \Psi_2(x, 0)$ satisfies our requirement. Here we take $G = V_1$ and $V = \Psi(V_1 \times V_2)$.

□

A basic knowledge we pick up from the implicit function theorem is, keeping the notations in Theorem 3.10, whenever DF is of full rank on the locus of $F(x, y) = 0$, the locus is an "n-dimensional surface" in $\mathbb{R}^{n+m}$. Thinking of $n + m$ many free variables are constrained by m many equations $F(x, y) = 0$ and thus leaving with n many free variables, the terminology of n-dimensional surface is easily understood. In general, for a given smooth F, the level set $\{(x, y) : \ F(x, y) = c\}$ may or may not be an n-dimensional surface. We call those values c such that $DF(x, y)$ is of rank m at every (x, y) satisfying $F(x, y) = c$ a *regular value* of F. A theorem of Sard asserts that for a *smooth* F, regular values are of full measure. It implies that, in case $F(x, y) = c$ is not regular, we can always find some regular value c' arbitrarily close to c. For instance, 0 is not a regular value for the function $x^2 - y^2$. However, $x^2 - y^2 = a, \ a \neq 0$ is a regular value.

It is interesting to note that the Inverse Function Theorem can be deduced from Implicit Function Theorem. Thus they are equivalent. To see this, keeping the notations used in Theorem 3.7. Define a map $\Phi : U \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Phi(x, y) = F(x) - y.$$

Then $\Phi(x_0, y_0) = 0, \Phi(x_0, y_0) = 0$, and $D_x\Phi(x_0, y_0) = DF(x_0)$ is invertible. By Theorem 3.10 (exchange x and y), there exists a C^1-function φ near y_0 satisfying $\varphi(y_0) = x_0$ and $\Phi(\varphi(y), y) = F(\varphi(y)) - y = 0$, hence φ is the local inverse of F.

We end this section by providing a justification to the method of Lagrange multipliers in optimization.

Theorem 3.11. *Let f and $g_1, \cdots, g_m, 1 \leq m < n$, be C^1-functions in some open set U in $\mathbb{R}^n$. Suppose that p_0 is a local minimum of f subject to the constraint $g_j(p) = 0, j = 1, \cdots, m$. Assuming $DG(p_0), G = (g_1, \cdots, g_m)$ if of rank m, there is some $\lambda = (\lambda_1, \cdots, \lambda_m)$ such that*

$$\nabla f(p_0) + \sum_{j=1}^{m} \lambda_j \nabla g_j(p_0) = 0 \ .$$

Proof. Here we take $n = 3, m = 1$ and write $p_0 = (x_0, y_0, z_0)$. Since $\nabla g(x_0, y_0, z_0) \neq (0, 0, 0)$, WLOG we may assume $g_z(p_0) \neq 0$. by the Implicit Function Theorem, there is

some open set $V \subset \mathbb{R}^2$ containing (x_0, y_0) and a C^1-function φ on V such that the locust of $g = 0$ is given by the graph of φ, that is, $g(x, y, \varphi(x, y)) = 0$ and $\varphi(x_0, y_0) = z_0$ for $(x, y) \in V$. It follows that (x_0, y_0) is a local minimum for the function $h(x, y) \equiv f(x, y, \varphi(x, y))$ over V. We have

$$h_x(x_0, y_0) = f_x(x_0, y_0, \varphi(x_0, y_0)) + f_z(x_0, y_0, \varphi(x_0, y_0))\varphi_x(x_0, y_0) = \nabla f \cdot (1, 0, \varphi_x) = 0,$$

and

$$h_y(x_0, y_0) = f_y(x_0, y_0, \varphi(x_0, y_0)) + f_z(x_0, y_0, \varphi(x_0, y_0))\varphi_x(x_0, y_0) = \nabla f \cdot (0, 1, \varphi_y) = 0,$$

That is, ∇f is perpendicular to the two dimensional subspace spanned by $(1, 0, \varphi_x)$ and $(0, 1, \varphi_y)$ at p_0. (In fact, this subspace is the tangent space of $g = 0$ at p_0.) On the other hand, by differentiating $g(x, y, \varphi(x, y)) = 0$, we also have

$$g_x(x, y, \varphi(x, y)) + g_z(x, y, \varphi(x, y))\varphi_x(x, y) = \nabla g \cdot (1, 0, \varphi_x) = 0,$$

and

$$g_y(x, y, \varphi(x, y)) + g_z(x, y, \varphi(x, y))\varphi_x(x, y) = \nabla g \cdot (0, 1, \varphi_y) = 0.$$

It shows that the three vectors $\nabla g, (1, 0, \varphi_x), (0, 1, \varphi_y)$ forms a basis at p_0. Therefore, $\nabla f(p_0)$ either points to the same or the opposite direction of $\nabla g(p_0$, that is, or $\nabla f + \lambda \nabla g$ at p_0 for some λ. □

3.4 Picard-Lindelöf Theorem for Differential Equations

In this section we discuss the fundamental existence and uniqueness theorem for differential equations. I assume that you learned the skills of solving ordinary differential equations already so we will focus on the theoretical aspects.

Most differential equations cannot be solved explicitly, in other words, they cannot be expressed as the composition of elementary functions. Nevertheless, there are two exceptional classes which come up very often. Let us review them before going into the theory.

Example 3.14. Consider the equation

$$\frac{dx}{dt} = a(t)x + b(t),$$

where a and b are continuous functions defined on some open interval I. This differential equation is called a linear differential equation because it is linear in x (with coefficients functions of t). The general solution of this linear equation is given by the formula

$$x(t) = e^{\alpha(t)} \left(x_0 + \int_{t_0}^{t} e^{-\alpha(s)} b(s) ds \right), \quad \alpha(t) = \int_{t_0}^{t} a(s) ds,$$

where $t_0 \in I, x_0 \in \mathbb{R}$, are arbitrary.

Example 3.15. The second class is the so-called separable equation

$$\frac{dx}{dt} = \frac{f(t)}{g(x)},$$

where f and $g \neq 0$ are continuous functions on intervals I and J respectively. The solution can be obtained by an integration

$$\int_{x_0}^{x} g(z)dz = \int_{t_0}^{t} f(s)ds, \qquad t_0 \in I,\ x_0 \in J.$$

The resulting relation, written as $G(x) = F(t)$, can be converted formally into $x = G^{-1}F(t)$, a solution to the equation as immediately verified by the chain rule. For instance, consider the equation

$$\frac{dx}{dt} = \frac{t+3}{x}\ .$$

The solution is given by integrating

$$\int_{x_0}^{x} xdx = \int_{t_0}^{t} (t+3)dt\ ,$$

to get

$$x^2 = t^2 + 6t + c\ , \quad c \in \mathbb{R}.$$

We have

$$x(t) = \pm\sqrt{t^2 + 6t + c}\ .$$

When $x(0) = -2$ is specified, we find the constant $c = 4$, so the solution is given by

$$x(t) = -\sqrt{t^2 + 6t + 4}\ .$$

More interesting explicitly solvable equations can be found in texts on ODE's.

Well, let us consider the general situation. Numerous problems in natural sciences and engineering led to the initial value problem of differential equations. Let f be a function defined in some set E in $\mathbb{R}^2$ and (t_0, x_0) an interior point in E. We ask: Is there a solution $x = x(t)$ defined in some interval I containing $t_0, (t, x(t)) \in E, \forall t \in I$, satisfying the differentiable equation $x' = f(t, x)$ as well as $x(t_0) = x_0$? Since we are looking for a *local* solution, we may formulate the problem restricting to a rectangle centered at (t_0, x_0). In the following we will take E to be the rectangle $R = [t_0 - a, t_0 + a] \times [x_0 - b, x_0 + b]$ for some $a, b > 0$ and consider the **initial value problem** (IVP) (also called the **Cauchy Problem**)

$$\begin{cases} \dfrac{dx}{dt} = f(t, x), \\ \\ x(t_0) = x_0. \end{cases} \tag{IVP}$$

(In some books the independent variable t is replaced by x and the dependent variable x is replaced by y. We prefer to use t instead of x as the independent variable in many cases is the time.) To solve the initial value problem it means to find a function $x(t)$ defined in a perhaps smaller rectangle, that is, $x : [t_0 - a', t_0 + a'] \to [x_0 - b, x_0 + b]$, which is differentiable and satisfies $x(t_0) = x_0$ and $x'(t) = f(t, x(t)),\ \forall t \in [t_0 - a', t_0 + a']$, for some $0 < a' \leq a$. In general, no matter how nice f is, we do not expect there is always a solution on the entire $[t_0 - a, t_0 + a]$. Let us look at the following example.

Example 3.16. Consider the initial value problem

$$\begin{cases} \dfrac{dx}{dt} = 1 + x^2, \\ \\ x(0) = 0. \end{cases}$$

The function $f(t, x) = 1 + x^2$ is smooth on $[-a, a] \times [-b, b]$ for every $a, b > 0$. However, the solution, as one can verify immediately, is given by $x(t) = \tan t$ which is only defined on $(-\pi/2, \pi/2)$. It shows that even when f is very nice, a' could be strictly less than a.

Furthermore, replace the equation by $x' = \alpha(1+x^2)$. Accordingly the solution becomes $\tan \alpha t$ which exists in $(-\pi/2\alpha, \pi/2\alpha)$. It indicates that the interval of existence depending on f.

The Picard-Lindelöf theorem, sometimes referred to as the fundamental theorem of existence and uniqueness of differential equations, gives a clean condition on f ensuring the unique solvability of the initial value problem (IVP). This condition imposes a further regularity condition on f reminding what we did in the convergence of Fourier series. Specifically, a function f defined in R satisfies the **Lipschitz condition** (uniform in t) if there exists some $L > 0$ such that $\forall (t, x_i) \in R \equiv [t_0 - a, t_0 + a] \times [x_0 - b, x_0 + b],\ i = 1, 2,$

$$|f(t, x_1) - f(t, x_2)| \leq L\,|x_1 - x_2|\,.$$

Note that in particular means for each fixed t, f is Lipschitz continuous in x. The constant L is called a **Lipschitz constant.** Obviously if L is a Lipschitz constant for f, any number greater than L is also a Lipschitz constant. Not all continuous functions satisfy the Lipschitz condition. An example is given by the function $f(t, x) = tx^{1/2}$ is continuous. I let you verify that it does not satisfy the Lipschitz condition on any rectangle containing the origin.

In application, most functions satisfying the Lipschitz condition arise in the following manner. A C^1-function $f(t, x)$ in a closed rectangle automatically satisfies the Lipschitz condition. For, by the mean-value theorem, for some z lying on the segment between x_1 and x_2,

$$f(t, x_2) - f(t, x_1) = \frac{\partial f}{\partial x}(t, z)(x_2 - x_1).$$

Letting

$$L = \max\left\{ \left| \frac{\partial f}{\partial x}(t,x) \right| : \ (t,x) \in R \right\},$$

(L is a finite number because $\partial f/\partial y$ is continuous on R and hence bounded), we have

$$|f(t,x_2) - f(t,x_1)| \le L|x_2 - x_1|, \quad \forall (t,x_i) \in R, \ i = 1,2.$$

Theorem 3.12 (Picard-Lindelöf Theorem). *Consider (IVP) where $f \in C(R)$ satisfies the Lipschitz condition on $R = [t_0 - a, t_0 + a] \times [x_0 - b, x_0 + b]$. There exist $a' \in (0, a)$ and $x \in C^1[t_0 - a', t_0 + a']$, $x_0 - b \le x(t) \le x_0 + b$ for all $t \in [t_0 - a', t_0 + a']$, solving (IVP). Furthermore, x is the unique solution in $[t_0 - a', t_0 + a']$.*

From the proof one will see that a' can be taken to be any number satisfying

$$0 < a' < \min\left\{ a, \frac{b}{M}, \frac{1}{L} \right\},$$

where $M = \sup\{|f(t,x)| : \ (t,x) \in R\}$.

To prove Picard-Lindelöf Theorem, we first convert (IVP) into a single integral equation.

Proposition 3.13. *Setting as Picard-Lindelöf Theorem, every solution x of (IVP) from $[t_0 - a', t_0 + a']$ to $[x_0 - b, x_0 + b]$ satisfies the equation*

$$x(t) = x_0 + \int_{t_0}^{t} f(t, x(t))\, dt. \tag{3.7}$$

Conversely, every continuous function $x(t), t \in [t_0 - a', t_0 + a']$, satisfying (3.7) is continuously differentiable and solves (IVP).

Proof. When x satisfies $x'(t) = f(t, x(t))$ and $x(t_0) = x_0$, (3.7) is a direct consequence of the Fundamental Theorem of Calculus (first form). Conversely, when $x(t)$ is continuous on $[t_0 - a', t_0 + a']$, $f(t, x(t))$ is also continuous on the same interval. By the Fundamental Theorem of Calculus (second form), the left hand side of (3.7) is continuously differentiable on $[t_0 - a', t_0 + a']$ and solves (IVP). □

Note that in this proposition we do not need the Lipschitz condition; only the continuity of f is needed.

Proof of Picard-Lindelöf Theorem. Instead of solving (IVP) directly, we look for a solution of (3.7). We will work on the metric space

$$X = \{x \in C[t_0 - a', t_0 + a'] : \ x(t) \in [x_0 - b, x_0 + b], x(t_0) = x_0\} \ ,$$

with the uniform metric (the metric induced by the supnorm). It is easily verified that it is a closed subset in the complete metric space $C[t_0 - a', t_0 + a']$ and hence complete. Recall that every closed subset of a complete metric space is complete. The number a' will be specified below.

We are going to define a contraction on X. Indeed, for $x \in X$, define $\mathcal{T}$ by

$$(\mathcal{T}x)(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds.$$

First of all, for every $x \in X$, it is clear that $f(t, x(t))$ is well-defined and $\mathcal{T}x \in C[t_0 - a', t_0 + a']$. To show that it is in X, we need to verify $x_0 - b \leq (\mathcal{T}x)(t) \leq x_0 + b$ for all $t \in [t_0 - a', t_0 + a']$. We claim that this holds if we choose a' satisfying $a' \leq b/M$, $M = \sup\{|f(t, x)| : (t, x) \in R\}$. For,

$$\begin{aligned} |(\mathcal{T}x)(t) - x_0| &= \left| \int_{t_0}^{t} f(t, x(t))\, dt \right| \\ &\leq M\, |t - t_0| \\ &\leq Ma' \\ &\leq b. \end{aligned}$$

Next, we claim $\mathcal{T}$ is a contraction on X when a' is further restricted to $a' < 1/L$ where and L is the Lipschitz constant for f. For,

$$\begin{aligned} |(\mathcal{T}x_2 - \mathcal{T}x_1)(t)| &= \left| \int_{t_0}^{t} f(t, x_2(t)) - f(t, x_1(t))\, dt \right| \\ &\leq \left| \int_{t_0}^{t} \big| f(t, x_2(t)) - f(t, x_1(t)) \big|\, dt \right| \\ &\leq L \left| \int_{t_0}^{t} |x_2(t) - x_1(t)|\ dt \right| \\ &\leq L \sup_{t \in I} |x_2(t) - x_1(t)|\, |t - t_0| \\ &\leq La' \|x_2 - x_1\|_\infty, \end{aligned}$$

where $I = [t_0 - a', t_0 + a']$. It follows that

$$\|\mathcal{T}x_2 - \mathcal{T}x_1\|_\infty \leq \gamma \|x_2 - x_1\|_\infty\ , \quad \gamma = a'L < 1\ .$$

Now we can apply the Contraction Mapping Principle to conclude that $\mathcal{T}x = x$ for some x, and x solves (IVP). We have shown that (IVP) admits a solution in $[t_0 - a', t_0 + a']$ where a' can be chosen to be any number less than $\min\{a, b/M, 1/L\}$.

Finally, any solution to the IVP is a fixed point of the map $\mathcal{T}$, so the IVP has a unique solution on $[t_0 - a', t_0 + a']$.

□

We point out that the existence part of Picard-Lindelöf Theorem still holds without the Lipschitz condition. We will prove this in the next chapter. However, the solution may not be unique.

The uniqueness assertion in this theorem is restricted to the interval $[t_0 - a', t_0 + a']$. In fact, uniqueness holds regardless of the size of the interval of existence. We have

Proposition 3.14. *Consider the IVP where $f \in C(D), D \subset \mathbb{R}^2$ an open set satisfying the Lipschitz condition. Suppose x_1 and x_2 are two solutions of this IVP over an interval I such that their graphs lying inside D. Suppose that $x_1(t_0) = x_2(t_0)$ at some $t_0 \in I$, then x_1 coincides with x_2 on I.*

Proof. For $i = 1, 2$, we have

$$x_i(t) = x_i(t_0) + \int_{t_0}^{t} f(s, x(s))\, ds \ , \quad t \in I \ .$$

By subtracting, as $x_1(t_0) = x_2(t_0)$,

$$\begin{aligned} |x_1(t) - x_2(t)| &= \left| \int_{t_0}^{t} |f(s, x_1(s)) - f(s, x_2)|\, ds \right| \\ &\leq L \left| \int_{t_0}^{t} |x_1(s) - x_2(s)|\, ds \right| \ . \end{aligned}$$

Let us take $t > t_0$. (The case $t < t_0$ can be handled similarly.) The function

$$H(t) \equiv \int_{t_0}^{t} |x_1(s) - x_2(s)|\, ds$$

satisfies the differential inequality

$$H'(t) \leq LH(t) \ , \quad t \in I^+, \quad I^+ = I \cap \{t > t_0\}.$$

It satisfies $H(t_0) = 0$ and is always increasing. Moreover, it vanishes on I^+ if and only if x_1 coincides with x_2 on I^+. To show that H vanishes, we add an $\varepsilon > 0$ to the right hand side of this differential inequality to get $H' \leq L(H + \varepsilon)$. (The adding of ε makes $H + \varepsilon$ always positive.) Writing it as $(\log(H + \varepsilon))' \leq L$, and integrating it to get

$$\log(H(t) + \varepsilon) - \log \varepsilon \leq L(t - t_0) \ ,$$

or

$$H(t) + \varepsilon \leq \varepsilon e^{L(t-t_0)}, \quad t \in I^+ \ .$$

Now the desired conclusion follows by letting $\varepsilon \to 0$.

□

Under the assumption of this proposition, let $\mathcal{S}$ be the collection of all pairs $(x(t), I)$ where $x(t)$ is a solution over the open interval I and whose graph passing $(t_0, x_0), t_0 \in I$. Letting $I^* = \cup I_\alpha$ where I_α ranges over all I's in $\mathcal{S}$, the function x^* on I^* defines by $x^*(t) = x_\alpha(t)$ whenever $t \in I_\alpha$ is a well-defined function which solves the (IVP) over I^*. It is called the **maximal solution** to the (IVP).

Picard-Lindelöf Theorem remains valid for systems of differential equations. Consider the system

$$\begin{cases} \dfrac{dx_j}{dt} = f_j(t, x_1, x_2, \cdots, x_n), \\ \\ x_j(t_0) = x_{0j}, \end{cases}$$

where $j = 1, 2, \cdots, n$. By setting $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $f = (f_1, f_2, \cdots, f_n)$, we can express it as in (IVP) but now both $\mathbf{x}$ and $\mathbf{f}$ are vectors.

Essentially following the same arguments as the case of a single equation, we have

Theorem 3.15 (**Picard-Lindelöf Theorem for Systems**). *Consider (IVP) where* $\boldsymbol{f} = (f_1, \cdots, f_n)$, $f_j \in C(R)$ *satisfies the Lipschitz condition*

$$\|\boldsymbol{f}(t, \boldsymbol{x}) - \boldsymbol{f}(t, \boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\| \ ,$$

for all $(t, \boldsymbol{x}) \in R = [t_0 - a, t_0 + a] \times [x_{01} - b, x_{01} + b] \times \cdots \times [x_{0n} - b, x_{0n} + b]$. *There exists a unique solution*

$$\boldsymbol{x} \in C^1[t_0 - a', t_0 + a'], \ \boldsymbol{x}(t) \in [x_{01} - b, x_{01} + b] \times \cdots \times [x_{0n} - b, x_{0n} + b],$$

to (IVP) where

$$0 < a' < \min\left\{a, \frac{b}{M}, \frac{1}{L}\right\} \ , \quad M \geq |f_j(t, \boldsymbol{x})| : \ (t, \boldsymbol{x}) \in R, \ j = 1, \cdots, n \ .$$

Here for $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|$ is the Euclidean norm.

Finally, let me remind you that there is a standard way to convert the initial value problem for higher order differential equation ($m \geq 2$)

$$\begin{cases} x^{(m)} = f(t, x, x', \cdots, x^{(m-1)}), \\ \\ x(t_0) = x_0, \ x'(t_0) = x_1, \cdots, x^{(m-1)}(t) = x_{m-1}, \end{cases}$$

into a system of first order differential equations. As a result, we also have a corresponding Picard-Lindelöf theorem for higher order differential equations. I let you formulate it.

3.5 Appendix I: Completion of a Metric Space

A metric space (X, d) is called **isometrically embedded** in (Y, ρ) if there is a mapping $\Phi : X \to Y$ such that $d(x, y) = \rho(\Phi(x), \Phi(y))$. Note that this condition implies that Φ is 1-1 and continuous. We call the metric space (Y, ρ) a **completion** of (X, d) if it is complete, (X, d) is embedded in (Y, ρ) and $\overline{\Phi(X)} = Y$. The latter condition is a minimality condition; (X, d) is enlarged merely to accommodate those ideal points to make the space complete. When X is isometrically embedded in Y, we may identify X with its image $\Phi(X)$ and d with ρ. Or, we can image X being enlarged to a larger set Y where d is also extended to some ρ on Y which makes Y complete.

Before the proof of the Completion Theorem we briefly describe the ideas behind. When (X, d) is not complete, we need to invent ideal points and add them to X to make it complete. The idea goes back to Cantor's construction of the real numbers from rational numbers. Suppose now we have only rational numbers and we want to add irrationals. First we identify $\mathbb{Q}$ with a proper subset in a larger set as follows. Let $\mathcal{C}$ be the collection of all Cauchy sequences of rational numbers. Every point in $\mathcal{C}$ is of the form $(x_1, x_2, \cdots)$ where $\{x_n\}, x_n \in \mathbb{Q}$, forms a Cauchy sequence. A rational number x is identified with the constant sequence $\{x, x, x, \dots\}$ or any Cauchy sequence which converges to x. For instance, 1 is identified with $\{1, 1, 1, \dots\}$, $\{0.9, 0.99, 0.999, \dots\}$ or $\{1.01, 1.001, 1.0001, \dots\}$. Clearly, there are Cauchy sequences which cannot be identified with rational numbers. For instance, there is no rational number corresponding to $\{3, 3.1, 3.14, 3.141, 3.1415, \dots\}$, as we know, its correspondent should be the irrational number π. Similar situation holds for the sequence $\{1, 1.4, 1.41, 1.414, \cdots\}$ which should correspond to $\sqrt{2}$. Since the correspondence is not injective, we make it into one by introducing an equivalence relation on $\mathcal{C}$ Indeed, $\{x_n\}$ and $\{y_n\}$ are said to be equivalent if $|x_n - y_n| \to 0$ as $n \to \infty$. The equivalence relation $\sim$ forms the quotient $\mathcal{C}/\sim$ which is denoted by $\widetilde{\mathcal{C}}$. Then $x \mapsto \widetilde{x}$ sends $\mathbb{Q}$ injectively into $\widetilde{\mathcal{C}}$. It can be shown that $\widetilde{\mathcal{C}}$ carries the structure of the real numbers. In particular, those points not in the image of $\mathbb{Q}$ are exactly irrational numbers. Now, for a metric space the situation is similar. We let $\widetilde{\mathcal{C}}$ be the quotient space of all Cauchy sequence in X under the relation $\{x_n\} \sim \{y_n\}$ if and only if $d(x_n, y_n) \to 0$. Define $\widetilde{d}(\widetilde{x}, \widetilde{y}) = \lim_{n\to\infty} d(x_n, y_n)$, for $x \in \widetilde{x}$, $y \in \widetilde{y}$. We have the embedding $(X, d) \to (\widetilde{X}, \widetilde{d})$, and we can further show that it is a completion of (X, d).

The following proof is for optional reading. In the exercise we will present a simpler but less instructive proof.

Proof of Theorem 3.2. Let $\mathcal{C}$ be the collection of all Cauchy sequences in (X, d). We introduce a relation $\sim$ on $\mathcal{C}$ by $x \sim y$ if and only if $d(x_n, y_n) \to 0$ as $n \to \infty$. It is routine to verify that $\sim$ is an equivalence relation on $\mathcal{C}$. Let $\widetilde{X} = \mathcal{C}/\sim$ and define a map:

$\widetilde{X} \times \widetilde{X} \mapsto [0, \infty)$ by

$$\widetilde{d}(\widetilde{x}, \widetilde{y}) = \lim_{n\to\infty} d(x_n, y_n)$$

where $x = (x_1, x_2, x_3, \cdots)$ and $y = (y_1, y_2, y_3, \cdots)$ are respective representatives of $\widetilde{x}$ and $\widetilde{y}$. We note that the limit in the definition always exists: For

$$d(x_n, y_n) \leq d(x_n, x_m) + d(x_m, y_m) + d(y_m, y_n)$$

and, after switching m and n,

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

As x and y are Cauchy sequences, $d(x_n, x_m)$ and $d(y_m, y_n) \to 0$ as $n, m \to \infty$, and so $\{d(x_n, y_n)\}$ is a Cauchy sequence of real numbers.

Step 1. (well-definedness of $\widetilde{d}$) To show that $\widetilde{d}(\widetilde{x}, \widetilde{y})$ is independent of their representatives, let $x \sim x'$ and $y \sim y'$. We have

$$d(x_n, y_n) \leq d(x_n, x'_n) + d(x'_n, y'_n) + d(y'_n, y_n).$$

After switching x and x', and y and y',

$$|d(x_n, y_n) - d(x'_n, y'_n)| \leq d(x_n, x'_n) + d(y_n, y'_n).$$

As $x \sim x'$ and $y \sim y'$, the right hand side of this inequality tends to 0 as $n \to \infty$. Hence $\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x'_n, y'_n)$.

Step 2. ($\widetilde{d}$ is a metric). Let $\{x_n\}, \{y_n\}$ and $\{z_n\}$ represent $\widetilde{x}, \widetilde{y}$ and $\widetilde{z}$ respectively. We have

$$\begin{aligned}
\widetilde{d}(\widetilde{x}, \widetilde{z}) &= \lim_{n\to\infty} \big(d(x_n, z_n) \\
&\leq \lim_{n\to\infty} \big(d(x_n, y_n) + d(y_n, z_n)\big) \\
&= \lim_{n\to\infty} d(x_n, y_n) + \lim_{n\to\infty} d(y_n, z_n) \\
&= \widetilde{d}(\widetilde{x}, \widetilde{y}) + \widetilde{d}(\widetilde{y}, \widetilde{z})
\end{aligned}$$

Step 3. We claim that there is a metric preserving map $\Phi : X \mapsto \widetilde{X}$ satisfying $\overline{\Phi(X)} = \widetilde{X}$.

Given any x in X, the "constant sequence" $(x, x, x, \cdots)$ is clearly a Cauchy sequence. Let $\widetilde{x}$ be its equivalence class in $\mathcal{C}$. Then $\Phi x = \widetilde{x}$ defines a map from X to $\widetilde{X}$. Clearly

$$\widetilde{d}(\Phi(x), \Phi(y)) = \lim_{n\to\infty} d(x_n, y_n) = d(x, y)$$

since $x_n = x$ and $y_n = y$ for all n, so Φ is metric preserving and it is injective in particular.

To show that the closure of $\Phi(X)$ is $\widetilde{X}$, we observe that any $\widetilde{x}$ in $\widetilde{X}$ is represented by a Cauchy sequence $x = (x_1, x_2, x_3, \cdots)$. Consider the constant sequence $x^n = (x_n, x_n, x_n, \cdots)$ in $\Phi(X)$. We have

$$\widetilde{d}(\widetilde{x}, \widetilde{x_n}) = \lim_{m\to\infty} d(x_m, x_n).$$

Given $\varepsilon > 0$, there exists an n_0 such that $d(x_m, x_n) < \varepsilon/2$ for all $m, n \geq n_0$. Hence $\widetilde{d}(\widetilde{x}, \widetilde{x_n}) = \lim_{m\to\infty} d(x_m, x_n) < \varepsilon$ for $n \geq n_0$. That is $\widetilde{x^n} \to \widetilde{x}$ as $n \to \infty$, so the closure of $\Phi(X)$ is precisely $\widetilde{X}$.

Step 4. We claim that $(\widetilde{X}, \widetilde{d})$ is a complete metric space. Let $\{\widetilde{x^n}\}$ be a Cauchy sequence in $\widetilde{X}$. As $\overline{\Phi(X)}$ is equal to $\widetilde{X}$, for each n we can find a $\widetilde{y}$ in $\Phi(X)$ such that

$$\widetilde{d}(\widetilde{x^n}, \widetilde{y^n}) < \frac{1}{n}\ .$$

So $\{\widetilde{y^n}\}$ is also a Cauchy sequence in $\widetilde{d}$. Let y_n be the point in X so that $y^n = (y_n, y_n, y_n, \cdots)$ represents $\widetilde{y^n}$. Since Φ is metric preserving, and $\{\widetilde{y^n}\}$ is a Cauchy sequence in $\widetilde{d}$, $\{y_n\}$ is a Cauchy sequence in X. Let $(y_1, y_2, y_3, \cdots) \in \widetilde{y}$ in $\widetilde{X}$. We claim that $\widetilde{y} = \lim_{n\to\infty} \widetilde{x^n}$ in $\widetilde{X}$. For, we have

$$\begin{aligned} \widetilde{d}(\widetilde{x^n}, \widetilde{y}) &\leq \widetilde{d}(\widetilde{x^n}, \widetilde{y^n}) + \widetilde{d}(\widetilde{y^n}, \widetilde{y}) \\ &\leq \frac{1}{n} + \lim_{m\to\infty} d(y_n, y_m) \to 0 \end{aligned}$$

as $n \to \infty$. We have shown that $\widetilde{d}$ is a complete metric on $\widetilde{X}$. □

Completion of a metric space is unique once we have clarified the meaning of uniqueness. Indeed, call two metric spaces (X, d) and (X', d') **isometric** if there exists a bijective embedding from (X, d) onto (X', d'). Since a metric preserving map is always one-to-one, the inverse of of this mapping exists and is a metric preserving mapping from (X', d') to (X, d). So two spaces are isometric provided there is a metric preserving map from one onto the other. Two metric spaces will be regarded as the same if they are isometric, since then they cannot be distinguish after identifying a point in X with its image in X' under the metric preserving mapping. With this understanding, the completion of a metric space is unique in the following sense: If (Y, ρ) and (Y', ρ') are two completions of (X, d), then (Y, ρ) and (Y', ρ') are isometric. We will not go into the proof of this fact, but instead leave it to the interested reader. In any case, now it makes sense to use "the completion" of X to replace "a completion" of X.

3.6 Appendix II: Construction of Real Numbers

After the discovery of calculus by Newton and Leibniz, mathematics developed in an incredibly fast speed. However, how to make it rigorous had not been a concern for

mathematicians in this period. Toward the end of the nineteenth century, people began to feel the need to clarify various things such as convergence of series. They soon realized mathematics should be built upon the new theory of sets and the number systems. By the effort of many people, nowadays the paramount building of mathematics stands on relatively solid ground.

Mathematics is all about deduction. Proceeding from a few axioms, together with insightful definitions, people deduce results from simple to sophisticated. Set theory is the first step. Consider, for instance, the axioms proposed by Zermelo-Fraenkel, which carefully tell us how a set is constructed. A remarkable axiom in this theory is the axiom of choice which has many equivalent versions including the Zorn's lemma commonly used in analysis. With the notion of a set, one introduce ordered pairs and relations. Among many relations, the equivalence relation and mappings are most useful.

Next it comes to numbers. The construction of the number system follows the order: Natural numbers, integers, rational numbers, real numbers and finally complex numbers. Natural numbers are introduced by the five axioms of Pearo:

A1. *Zero is a natural number.*

A2. *Every natural number has a successor in the natural numbers.*

A3. *Zero is not the successor of any natural number.*

A4. *If the successor of two natural numbers is the same, then the two original numbers are the same.*

A5. *If a set contains zero and the successor of every number is in the set, then the set contains the natural numbers.* With these five axioms one establishes unique factorization property of natural numbers, introducing prime numbers, thus classical number theory is born. After defining integers and its arithmetic, one introduces rational numbers as the order pairs (p, q) where p, q are integers. Rational numbers consist of the equivalence class of (p, q) under the relation $(p, q) \sim (r, s)$ if and only if $ps = qr$. The arithmetic and ordering of integers are easily extended to all rational numbers.

There are two popular construction of the real numbers from rational numbers. Dedekind's cuts and Cantor's Cauchy sequences. The latter was briefed in class. You may search the internet to learn more. (This is not in the scope of MATH3060.)

Recall in Bartle-Sherbert's book, the construction of real numbers is replaced by a few additional axioms. For instance, it is assumed that $\mathbb{R}$ is a field satisfying certain well-ordering property. A crucial assumption is the supremum property : Every nonempty subset in $\mathbb{R}$ which is bounded from above has a supremum. It is this axiom which enables us to deduce Nested-Interval, Theorem, Bolzano-Weierstrass Theorem, Completeness Theorem, etc.

All these postulations become superfluous after the construction of $\mathbb{R}$ from $\mathbb{Q}$. One can prove the supremum property and them deduce all the other theorems, look up Wiki

"Cantor construction of real numbers" for details.

Comments on Chapter 3. There are two popular constructions of the real number system, Dedekind cuts and Cantor's Cauchy sequences. Although the number system is fundamental in mathematics, we did not pay much attention to its rigorous construction. It is too dry and lengthy to be included in Mathematical Analysis I. Indeed, there are two sophisticate steps in the construction of real numbers from nothing, namely, the construction of the natural numbers by Peano's axioms and the construction of real numbers from rational numbers. Other steps are much easier. Cantor's construction of the irrationals from the rationals is adapted to construct the completion for a metric space in Theorem 3.2. You may google under the key words "Peano's axioms, Cantor's construction of the real numbers, Dedekind cuts" for more.

Contraction Mapping Principle, or Banach Fixed Point Theorem, was found by the Polish mathematician S. Banach (1892-1945) in his 1922 doctoral thesis. He is the founder of functional analysis and operator theory. According to P. Lax, "During the Second World War, Banach was one of a group of people whose bodies were used by the Nazi occupiers of Poland to breed lice, in an attempt to extract an anti-typhoid serum. He died shortly after the conclusion of the war." The interested reader should look up his biography at Wiki.

An equally famous fixed point theorem is Brouwer's Fixed Point Theorem. It states that every continuous map from a closed ball in $\mathbb{R}^n$ to itself admits at least one fixed point. Here it is not the map but the geometry, or more precisely, the topology of the ball matters. You will learn it in a course on topology.

Inverse and Implicit Function Theorems, which reduce complicated structure to simpler ones via linearization, are the most frequently used tool in the study of the local behavior of maps. We learned these theorems and some of its applications in Advanced Calculus I already. In view of this, we basically provide detailed proofs here but leave out many standard applications. You may look up Fitzpatrick, "Advance Calculus", to refresh your memory. By the way, the proof in this book does not use Contraction Mapping Principle.

The case of polar coordinates (see Example 3.8) shows that a local invertible map may not be globally invertible. A theorem of Hadamard asserts that a continuous, locally bijective map F is globally bijective under an additional condition, namely, $|F(x)| \to \infty$ whenever $|x| \to \infty$. Incidentally, let us mention the celebrated Jacobian conjecture. Consider a map $F : \mathbb{R}^n \to \mathbb{R}^n$ whose components F_j's are polynomials in $x_1, \cdots, x_n$. Assume

that its Jacobian determinant is a nonzero constant. The conjecture asserts that this map is globally bijective whose inverse is also a polynomial map. Except for some special cases, this conjecture is still open.

Picard-Lindelöf Theorem or the fundamental existence and uniqueness theorem of differential equations was mentioned in Ordinary Differential Equations and now its proof is discussed in details. Of course, the contributors also include Cauchy and Lipschitz. Further results without the Lipschitz condition can be found in Chapter 4. A classic text on ordinary differential equations is "Theory of Ordinary Differential Equations" by E.A. Coddington and N. Levinson. V.I. Arnold's "Ordinary Differential Equations" is also a popular text.