If $<x,a> = 0$, then we have

$$\begin{aligned} \|x + \lambda a\|^2 &= <x + \lambda a, x + \lambda a> \\ &= <x,x> + 2\lambda <x,a> + \lambda^2<a,a> \\ &= \|x\|^2 + \lambda^2\|a\|^2 \\ &\geq \|x\|^2 \qquad (\text{as } \lambda^2\|a\|^2 \geq 0) \end{aligned}$$

and so, taking square roots

$$\|x + \lambda a\| \geq \|x\| \quad \text{as required.}$$

If, for all λ we have $\|x + \lambda a\| \geq \|x\|$, then squaring and using the definition of the norm we also have

$$<x + \lambda a, x + \lambda a> \geq <x,x>$$

or expanding the L.H.S. and cancelling

$$\lambda^2 <a,a> + 2\lambda <x,a> \geq 0.$$

Thus for $\lambda > 0$ we have

$$<x,a> \geq -\tfrac{1}{2}\lambda \|a\|^2$$

Letting $\lambda \to 0+$ gives

$$<x,a> \geq 0$$

while $\lambda \to 0-$ gives

$$<x,a> \leq 0,$$

and so $<x,a> = 0$ as required.

<u>EXERCISE:</u>

Let X be $C[-1,1]$ with inner-product

$$<f,g> = \int_{-1}^{1} f(x)g(t)dt.$$

Show that in X the best approximation to $f_0(t) = t^2$ by a function of the form $g(t) = at + b$ is given by $g_0(t) = a_0t + b_0$ where a_0 and b_0 are such that

$$\int_{-1}^{1} (t^3 - a_0t^2 = b_0t)dt = 0$$

and

$$\int_{-1}^{1} (t^2 - a_0t - b_0)dt = 0$$

Hence, find g_0.

From the example on page 102 and Theorem 8 we have:-

PROPOSITION 11: *If* A *is a closed subspace of the Hilbert space* X, *then* A *contains a unique best approximation to each point of* X.

Since finite dimensional subspaces are always closed (Exercise 1 on p.78) we have as a corollary.

PROPOSITION 12: *If* A *is a finite dimensional subspace of the Hilbert space* X, *then* A *contains a unique best approximation to each point of* X.

*<u>EXERCISE</u>: Since finite dimensional subspaces are always complete, by examining the proof of theorem 8 show that the assumption "X is a Hilbert space" in the above proposition may be replaced by "X is an inner-product space". That is: *every finite dimensional subspace of an inner-product space is a Tchebyscheff set.*

By using the characterization given in proposition 10 we can obtain an explicit expression for the best approximation in proposition 12.

PROPOSITION 13: *If* $\{e_1, e_2, \dots, e_n\}$ *is an orthonormal basis* of the finite dimensional subspace* A *in the Hilbert space* X, *then the best approximation to* x_0 *from* A *is*

$$a_0 = \sum_{i=1}^{n} \langle x_0, e_i \rangle \, e_i$$

Proof. Let a_0 be the best approximation to x_0 from A. Then, since $\{e_1, e_2, \dots, e_n\}$ is a basis for A we have

$$a_0 = \sum_{i=1}^{n} \alpha_i e_i \text{ for some scalars } \alpha_1, \alpha_2, \dots, \alpha_n.$$

Now, by proposition 10

$$\langle x_0 - a_0, a \rangle = 0 \text{ for all a A.}$$

* That is, the span of $e_1, e_2, \dots, e_n$ is A and $\langle e_i, e_j \rangle = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ if } i \neq j \end{cases}$

In particular then, for each j=1,2,...,n.

$$< x_0 - a_0, e_j > \; = 0$$

or

$$< x_0 - \sum_{i=1}^{n} \alpha_i e_i, e_j > = 0.$$

Expanding gives

$$< x_0, e_j > \; - \sum_{i=1}^{n} \alpha_i < e_i, e_j > = 0$$

Using the orthonormality of the basis $\left\{ < e_i, e_j > \; = \begin{cases} 1 \text{ if } i=j \\ 0 \text{ if } i \neq j \end{cases} \right\}$

we therefore have

$$< x_0, e_j > - \; \alpha_j = 0$$

or $\qquad \alpha_j = \; < x_0, e_j >$.

Substituting into the expression for a_0 gives the desired result. ■

EXERCISE: Give a direct proof of Proposition 13 by showing that

$$\| x_0 - \sum_{i=1}^{n} \alpha_i e_i \|^2$$

is minimized if and only if $\alpha_i = \; < x_0, e_i >$.

EXAMPLES

(I) In $C[-\pi,\pi]$ with inner product $<f,g> \; = \int_{-\pi}^{\pi} f(x)g(x)dx$ it is easily checked that the set $B = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \sin x, \frac{1}{\sqrt{\pi}} \cos x, \frac{1}{\sqrt{\pi}} \sin 2x, \frac{1}{\sqrt{\pi}} \cos 2x, \ldots \right\}$ forms an orthonormal set thus, the best approximation, f_0, to $f \in C[-\pi,\pi]$ form the finite dimensional subspace spanned by the first 2n+1 elements of B is

$$f(x) = \frac{\alpha_1}{\sqrt{2\pi}} + \frac{\alpha_2}{\sqrt{\pi}} \sin x + \frac{\alpha_3}{\sqrt{\pi}} \cos x + \ldots + \frac{\alpha_{2n-1}}{\sqrt{\pi}} \sin nx + \frac{\alpha_{2n}}{\sqrt{\pi}} \cos$$

where $\alpha_1 = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(x)dx$, $\alpha_{2m} = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(x)\sin mx \, dx$,

$$\alpha_{2m+1} = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(x) \cos mx \, dx \quad (m= 1,2,\ldots).$$

We ought to recognise this trigonometric approximation as the truncated Fourier series expansion of f.

(II) $\{1,x,x^2,\ldots\}$ is a linearly independent set of functions in $\mathcal{C}[-1,1]$ from which, using the Gram-Schmidt orthogonalization procedure with inner-product $\langle f,g\rangle = \int_{-1}^{1} f(x)g(x)dx$, we can arrive at the orthonormal sequence of polynomials (the Legendre Polynomials)

$$\left\{ 1/\sqrt{2}\ , \sqrt{\tfrac{3}{2}}\,x, \ldots \right\} .$$

Therefore, using the above results, we have for example, the best linear approsimation to e^x on $[-1,1]$, in the sense that $\int_{-1}^{1} (e^x - ax+b)^2 dx$ is a minimum, is $y = {}^{\alpha}/\sqrt{2} + \beta\left(\sqrt{\tfrac{3}{2}}\,x\right)$ where

$$\alpha = \frac{1}{\sqrt{2}} \int_{-1}^{1} e^x dx = \frac{e^2-1}{e\sqrt{2}}$$

and $$\beta = \sqrt{\frac{3}{2}} \int_{-1}^{1} xe^x dx = \frac{\sqrt{6}}{e}$$

so $$y = \left(\frac{e^2-1}{2} + 3x \right) /e \simeq e^x$$

*(III) Let A be an n×m matrix and consider the problem of finding $x \in \mathcal{R}^m$ such that $\|A\underset{\sim}{x} - \underset{\sim}{b}\|_2$ is a minimum, where $\underset{\sim}{b}$ is a given vector of $\mathcal{R}^n$. (c.f. example (V) of page 91). Regarding A as a linear mapping from $\mathcal{R}^m$ to $\mathcal{R}^n$ we know from our general theory that there exists a unique best approximation from the subspace $A(\mathcal{R}^m)$ to $\underset{\sim}{b}$. Thus there certainly exists an $\underset{\sim}{x}$ for which $\|A\underset{\sim}{x}-\underset{\sim}{b}\|_2$ is a minimum. Further provided A is 1-1 (which can only happen if m≤n, i.e. the system of equations is "over-specified") this $\underset{\sim}{x}$ is unique. To derive an expression for $\underset{\sim}{x}$ we proceed as follows.

For any $y \in \mathcal{R}^m$ we have $\|A\underset{\sim}{x} - \underset{\sim}{b}\|_2 \leq \|A(\underset{\sim}{x}+\underset{\sim}{y})-\underset{\sim}{b}\|_2$, as $\underset{\sim}{x}+\underset{\sim}{y}$ is an element of $\mathcal{R}^m$,

or $$\langle A\underset{\sim}{x}-\underset{\sim}{b},\ A\underset{\sim}{x}-\underset{\sim}{b}\rangle \leq \langle A(\underset{\sim}{x}+\underset{\sim}{y})-\underset{\sim}{b},\ A(\underset{\sim}{x}+\underset{\sim}{y})-\underset{\sim}{b}\rangle$$

expanding both sides we obtain

$$0 \le 2 \langle Ax-b,\ Ay\rangle \ + \ \|Ay\|^2$$

or $\quad \langle A^T A\underset{\sim}{x} - A^T\underset{\sim}{b},\ \underset{\sim}{y}\rangle \ \ge -\tfrac{1}{2}\|A\underset{\sim}{y}\|^2$ for all $\underset{\sim}{y}$ where A^T denotes the "transpose" of A.

Replacing $\underset{\sim}{y}$ by $\lambda\underset{\sim}{z}$ we obtain

$$\langle A^T A\underset{\sim}{x} - A^T\underset{\sim}{b},\underset{\sim}{z}\rangle \ge -\tfrac{1}{2}\lambda\|A\underset{\sim}{z}\|^2 \qquad \text{for all } \lambda \in R \text{ and } \underset{\sim}{z} \in R^m.$$

Letting $\lambda \to 0$ from above and below gives

$$\langle A^T A\underset{\sim}{x} - A^T\underset{\sim}{b},\underset{\sim}{z}\rangle = 0 \text{ for all } \underset{\sim}{z}.$$

But this happens if and only if

$$A^T A\underset{\sim}{x} - A^T\underset{\sim}{b} = 0$$

and so we see that $\underset{\sim}{x}$ is a solution as required if and only if

$$A^T A\underset{\sim}{x} = A^T\underset{\sim}{b}$$

[Note; $A^T A$ is an m×m matrix and so maybe invertible, in which case the unique solution is $\underset{\sim}{x} = (A^T A)^{-1} A^T \underset{\sim}{b}$]

Now consider the special case when

$$\underset{\sim}{x} = \begin{bmatrix} b \\ m \end{bmatrix}, \quad A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \underset{\sim}{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

which corresponds to finding m and b such that

$$\|A\underset{\sim}{x}-\underset{\sim}{b}\|_2^2 = \sum_{i=1}^{n} (y_i - mx_i - b_i)^2 \quad \text{is a minimum.}$$

This is the problem of finding the straight line y=mx+b *of least squares best fit to the* n *data points* $(x_1,y_1),(x_2,y_2),\dots,(x_n,y_n)$ *and amounts to the statistical problem of obtaining a linear regression on the data.*

Substituting into our general solution we obtain

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

or
$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

which, provided the system is non-degenerate, has as its unique solution

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{D}$$

$$m = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{D}$$

where $$D = n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2$$

[Note, this same solution could have been obtained by equating to zero the partial derivatives w r t b and m of $f(b,m) = \sum_{i=1}^{n} (y_i - mx_i - b)^2$. However this does not establish that f is a minimum at the solution point.]

EXERCISE: 1(a) (optional)

(Gram-Schmidt orthogonalization procedure)

Let $\{b_1, b_2, \ldots, b_n\}$ be any linearly independent subset of an inner-product space.

Show that the set of vectors $\{u_1, u_2, \ldots, u_n\}$ defined inductively by $u_1 = b_1/\|b_1\|$, $u_n = a_n/\|a_n\|$ where

$$a_n = b_n - \sum_{k=1}^{n-1} \langle b_n, u_k \rangle \, u_k,$$

is an orthonormal set of vectors whose span is the same as that of $\{b_k : k=1,2,\ldots,n\}$

(b) The set of vectors $\{1,x,x^2,\ldots,x^n\}$ is a basis of $P_n[0,1]$ in the inner-product space $C[0,1]$ with $<f,g> = \int_0^1 fg$. Use the procedure outlined in (a) to obtain an orthonormal basis for $P_2[0,1]$. Hence find the best quadratic approximation $p(x)$ to $y = x^3$ on $[0,1]$, in the sense that $\|x^3-p(x)\|_2$ is a minimum.

2. Let C be a closed convex subset of the Hilbert space H, and let $x \in H$. Let x_0 be the best approximation to x from C [which exists by Theorem 8, p.105 and the example on p.102].

For each $y \in C$ define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(t) = \|x - [tx_0 + (1 - t)y]\|^2.$$

By expanding this expression for $f(t)$ out in terms of the inner-product, show that $f(t)$ is a quadratic in t and hence a differential function.

From the observation that for $0 \le t \le 1$ we have $tx_0 + (1-t)y \in C$ and the definition of x_0 note that $f(t)$ attains its minimum value on $[0,1]$ at $t = 1$, hence conclude that $f'(1) \le 0$.

Deduce that

$$\langle x - x_0, \ y - x_0\rangle \le 0 \qquad \text{for all } y \in C.$$

Show that the converse is also true. That is, if $x_0 \in C$ is such that $\langle x - x_0 , y - x_0\rangle \le 0$ for all $y \in C$, then x_0 is the best approximation to x from C.

By noting that a closed subspace is a particular example of a closed convex set, rededuce propostion 10 from this last result.

SECTION 2.   FIXED POINT THEOREMS AND THEIR APPLICATIONS

§0. PRELIMINARIES

Recall: $x_0 \in X$ is a *fixed point* of the mapping $f: X \to X$ if $f(x_0) = x_0$.
Exercise: Let $f: [0,1] \to [0,1]$ be a continuous real valued function of a real variable. Show that f has a fixed point. [HINT: Apply the intermediate value theorem of Bolzano to the function $x \mapsto f(x) - x$.]

Many problems in Pure and Applied Mathematics have as their solutions the fixed "point" of some mapping f and so a number of the procedures in 'numerical' analysis and approximation theory amount to obtaining successive approximations to the fixed point of an appropriate mapping. (For example, *Newton's method* for finding the zero's of a function may be interpreted in this way.)

"Popular" accounts of fixed point theorems and their applications may be found in the following.
Courant and Robbins "What is Mathematics?", Ch. VIII.
and
Marvin Shinbrot "Fixed Point Theorems", *Scientific American*, January 1966, reprinted in: Readings from Scientific American "Mathematics in the Modern World".

For an interesting discussion of fixed point theorems and their applications consult:
Rosenlicht "Introduction to Analysis", Scott, Foresmand, 1968,
and
Hille "Methods in Classical and Functional Analysis", Addison-Wesley, 1972.

Because of their importance we will prove several fixed point theorems for mappings of a complete metric space into itself. We will then illustrate their use by giving one application, to the theory of ordinary differential equations.* A further application is given in an appendix (which you may treat as optional, but which you should at least look through).

We begin by briefly summarizing the necessary background material.

Throughout (X,d) will denote a <u>metric space</u>. That is, a set X for which a metric function $d: X \times X \to \mathbb{R}$ is defined and satisfies:

* The existence result established here will be assumed in the differential equations course.

i) $d(x,y) \geq 0$ for all $x, y \in X$;

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

iii) [Symmetry] $d(x,y) = d(y,x)$ for all $x, y \in X$.

iv) [Triangle inequality] $d(x,y) \leq d(x,z) + d(z,y)$ for all $x,y,z \in X$.

For example, if $(X,\|.\|)$ is a normed linear space, then $d(x,y) = \|x-y\|$ defines a metric on X - intuitively we think of $d(x,y)$ as the "distance" between the two points x and y.

A sequence (x_n) of points of X is convergent (to x) if $d(x_n,x) \to 0$ as $n \to \infty$ and is a Cauchy sequence if $d(x_n,x_m) \to 0$ as both n and $m \to \infty$. The metric space (X,d) in complete if every Cauchy sequence is convergent. For example, any (norm) closed subset of Banach space (X,d) is complete with respect to the metric $d(x,y) = \|x-y\|$.

A metric space (X,d) is (sequentially) compact if every sequence (x_n) of points of X has a subsequence (x_{n_k}) which is convergent (necessarily to a point of X). For example, any compact subset K of a normed linear space forms a compact metric space (K,d) where the metric is defined by $d(x,y) = \|x-y\|$ for all $x,y \in K$.

If (X,d) is a compact metric space, then it is complete, and every continuous function $f:(X,d) \to \mathbb{R}$ achieves its maximum and minimum on X.*

§1 FIXED POINT THEOREMS

We will be interested in mappings $T: (X,d) \to (X,d)$ which are of the following types.

i) *Non-expansive*, that is; $d(Tx,Ty) \leq d(x,y)$ for all x and $y \in X$.

ii) A *contraction*, that is, $d(Tx,Ty) \underset{(\neq)}{\leq} d(x,y)$ for all $x,y \in X$ with $x \neq y$.

iii) A *strict contraction*, that is, for some k with $0 \leq k \underset{(\neq)}{\leq} 1$,
$d(Tx,Ty) \leq k\, d(x,y)$ for all x and $y \in X$.

Clearly, T a strict contraction $\Rightarrow$ T a contraction $\Rightarrow$ T is non-expansive. EXERCISE, give examples of mappings from the closed interval $[-1,1]$ into itself which show that in general none of these implications can be reversed.

Before turning to the question of existence, we establish the "uniqueness" of a fixed point for a contraction (and hence also for a strict contraction). For many applications, uniqueness of the fixed point is almost as important as existence.

THEOREM 1. *Let* (X,d) *be a metric space and* $T: X \to X$ *a contraction mapping, then* T *can have, at most, one fixed point.*

Proof. Assume both x_1 and x_2 are fixed points of T, that is, $T(x_1) = x_1$ and $T(x_2) = x_2$. Then,

$$d(x_1,x_2) = d(Tx_1,Tx_2).$$

Further, by the definition of a contraction, either $x_1 = x_2$ or $d(Tx_1,Tx_2) \lneqq d(x_1,x_2)$, and so either

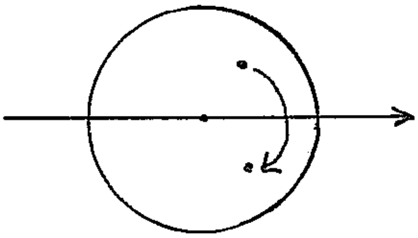
$$x_1 = x_2$$

or

$$d(x_1,x_2) = d(Tx_1,Tx_2) \lneqq d(x_1,x_2).$$

Since the second conclusion is impossible, we must have $x_1 = x_2$ and so T cannot have two distinct fixed points. ■

EXERCISES: (1) By means of an example, show that the conclusion of Theorem 1 will not, in general, be true for non-expansive mappings. [Hint: Let $X = \{z \in \mathbb{C}: |z| \leq 1\}$ and define $T: X \to X$ by $Tz = \bar{z}$. That is, T is the reflection of the unit disk about the diameter $\operatorname{Im} z = 0$.]



(2) Let $(X, \|\cdot\|)$ be a normed linear space and $T: X \to X$ be a linear mapping (that is, $T(x+\lambda y) = T(x) + \lambda T(y)$ for all $x, y \in X$ and $\lambda \in \mathbb{R}$). Find an "obvious" fixed point of T.

Define the "norm" of T to be

$$\|T\| = \inf\{m: \|Tx\| \leqslant m\|x\| \text{ for all } x \in X\}.^*$$

If $\|T\| < 1$ show that T has precisely one fixed point. Show that this need not be true if $\|T\| \geqslant 1$.

*You might like to try proving the important observation of Banach, that $\|T\|$ defined in this way is indeed a norm function for the space of all bounded linear mappings from X to X. Recall, a linear mapping T is bounded (or continuous) if $\|T\| < \infty$.

THEOREM 2 (Banach's Fixed point Theorem):

Let $T: X \to X$ *be a strict contraction of the complete metric space* (X,d) *into itself, then* T *has a unique fixed point in* X.

Proof. Let $k \in [0,1)$ be the "Lipschitz" constant such that $d(Tx,Ty) \leqslant kd(x,y)$ for all $x,y \in X$.

Take any point $x_1 \in X$ and inductively construct the sequence of points

$$\begin{aligned} x_2 &= T(x_1) \\ x_3 &= T(x_2) = T^2(x_1) \\ x_4 &= T(x_3) = T^3(x_1) \\ &\dots \\ x_{n+1} &= T(x_n) = T^n(x_1) \\ &\dots \end{aligned}$$

We first show that $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence in (X,d).

Thus, without loss of generality, take $m < n$ $(m,\ n \in \mathbb{N})$, then

$$\begin{aligned} d(x_m,\ x_n) &= d(T^{m-1}(x_1),\ T^{n-1}(x_1)) \\ &\leqslant kd(T^{m-2}(x_1),\ T^{n-2}(x_1)) \\ &\leqslant k^2d(T^{m-3}(x_1),\ T^{n-3}(x_1)) \\ &\leqslant \dots \\ &\leqslant k^{m-1}d(x_1,\ T^{n-m}(x_1)) \\ &\leqslant k^{m-1}\{d(x_1,\ T(x_1)) + d(T(x_1),\ T^2(x_1)) + \dots + d(T^{n-m-1}(x_1), \\ &\qquad T^{n-m}(x_1))\} \end{aligned}$$

(by extended application of the triangle inequality)

$$\begin{aligned} &\leqslant k^{m-1}\{d(x_1,\ T(x_1) + kd(x_1,\ T(x_1)) + k^2d(x_1,\ T(x_1)) + \dots + \\ &\qquad k^{n-m-1}\ d(x_1,\ T(x_1))\} \end{aligned}$$

(using T is a strict contraction)

$$\leqslant\ k^{m-1}(1 + k + k^2 + \dots + k^{n-m-1})\ d(x_1,\ T(x_1)).$$

Since $0 \leqslant k < 1$,

$$1 + k + k^2 + \dots + k^{n-m-1} < \sum_{j=0}^{\infty} k^j = \frac{1}{1 - k}$$

(sum of an infinite geometric progression),

whence

$$d(x_m, x_n) \leqslant \frac{k^{m-1}}{1-k} d(x_1, T(x_1)) \qquad \ldots (*)$$

$$\to 0 \text{ as } m \text{ (and hence } n) \to \infty.$$

Thus $\{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence, and so, by the completeness of (X, d). there exists $x_0 \in X$ with $x_n \to x_0$.

We now show x_0 is a fixed point of T.

Now $d(x_0, T(x_0)) \leqslant d(x_0, x_n) + d(x_n, T(x_0))$

$$\leqslant d(x_0, x_n) + kd(x_{n-1}, x_0) \quad (\text{as } x_n = T(x_{n-1}))$$

$$\to 0 \text{ as } x_n, \ x_{n-1} \to x_0$$

whence

$$d(x_0, T(x_0)) = 0 \text{ or}$$

$$T(x_0) = x_0,$$

and so, x_0 is a fixed point of T.

That x_0 is the unique fixed point of T has already been established in Theorem 1. ■

Almost as important as the result itself is the "constructive" nature of Banach's proof. Starting with any point x_1 of the space the successive iterates of x_1 under T; Tx_1, T^2x_1, T^3x_1, ..., converge to the fixed point x_0 of T. Further from the step (*) of the proof we can derive an easily evaluated estimate for the error in the m'th such approximation:

$$\begin{aligned} d(T^m(x_1), x_0) &= d(x_{m+1}, x_0) \\ &= \operatorname*{limit}_{n\to\infty} d(x_{m+1}, x_n) \\ &\leqslant \frac{k^m}{1-k} d(x_1, T(x_1)) \quad (\text{by } (*)). \end{aligned}$$

That is, $$d(T^m(x_1), x_0) \leqslant k^m\left(\frac{d(x_1, T(x_1))}{1-k}\right).$$

At each iteration the error is decreased by a factor of k (<1, by assumption). The constant in the error estimate $\frac{d(x_1,T(x_1))}{1-k}$ is, of course, easily evaluated for any particular choice of x_1.

EXERCISES:

(1) Let $f: [a,b] \to \mathbb{R}$ be such that $f(a) < 0 < f(b)$, f' exists and is continuous on $[a,b]$ and . there exists constants m, M with $0 < m \leqslant f'(x) \leqslant M$.

(1) (cont'd)

Show that for a suitable choice of constant k the mapping g defined by $g(x) = x - k\,f(x)$ is a contraction on $([a, b], d_1)$.
[Hint: Apply the mean-value theorem to show
$g(x) - g(y) = (x - y)(1 - kf'(z))$, for some $z \in (x, y)$.]
Hence, conclude that f has a unique zero in (a, b).

(2) Let $a \in \mathbb{R}$ be such that $|a - 1| < 1$. Show that the mapping f defined by

$$f(x) = \tfrac{1}{2}\big((1 + x)^2 - (1 + a)\big)$$

is a contraction on $X = \{x\colon |1 + x| \leqslant |1 - a|\}$ with respect to the usual metric. Hence conclude that a has a unique square root in X.
[REMARK. This result remains valid when a is an element of a Banach algebra (that is, a Banach space $(X, \|\cdot\|)$ in which a multiplication xy is defined and satisfies $\|xy\| \leqslant \|x\|\|y\|$). In such spaces it plays an important role by establishing the existence of square roots for certain elements.]

The remaining results indicate that by strengthening the assumptions on the domain of T it is possible to relax the requirements on T.

THEOREM 3: *Let* (X, d) *be a compact metric space and* $T\colon X \to X$ *a contraction* (not necessarily strict) *from* X *into itself, then* T *has a unique fixed point in* X.

Proof. Uniqueness has already been proved in Theorem 1, we therefore need only establish existence.

Define $\phi\colon X \to \mathbb{R}$ by

$$\phi(x) = d(x,Tx).$$

We begin by showing ϕ is continuous.
To see this, give $\varepsilon > 0$ we note that

$$\begin{aligned}
|\phi(x) - \phi(y)| &= |d(x,Tx) - d(y,Ty)| \\
&= |d(x,Tx) - d(Tx,y) + d(Tx,y) - d(y,Ty)| \\
&\leqslant |d(x,Tx) - d(Tx,y)| + |d(Tx,y) - d(y,Ty)| \\
&\leqslant d(x,y) + d(Tx,Ty)
\end{aligned}$$

~~$\leqslant 2d(x,y)$ (as T is a contraction)~~

$$< \varepsilon \quad \text{provided} \quad d(x,y) < \delta = \frac{\varepsilon}{2}\,.$$

Now since ϕ is continuous and X is compact there exists $x_0 \in X$ such that $\phi(x_0) \leq \phi(x)$ for all $x \in X$ (that is, ϕ attains its minimum at x_0). We show that x_0 is a fixed point of T. Since $\phi(x_0) \leq \phi(x)$ for all $x \in X$, taking $x = Tx_0$ we have

$$\phi(x_0) \leq \phi(Tx_0)$$

or
$$d(x_0, Tx_0) \leq d(Tx_0, T(Tx_0)).$$

Now if x_0 is not a fixed point of T, that is $Tx_0 \neq x_0$, then using the definition of a contraction we have

$$d(x_0, Tx_0) \leq d(Tx_0, T(Tx_0))$$

$$\underset{\neq}{\leq} d(x_0, Tx_0)$$

which is impossible, and so we conclude that $Tx_0 = x_0$ as required. ∎

REMARK: The above proof does not furnish an explicit procedure for approximating the fixed point in the same way that Banach's Theorem did. Nonetheless the successive iterates of any point $x_1 \in X$; Tx_1, T^2x_1, T^2x_1, ..., can be shown to converge to the unique fixed point x_0. What one cannot do is obtain the same type of precise error estimate possible for a strict contraction.

<u>THEOREM 4</u>: *Let* (X,d) *be a compact metric space and* T *a contraction of* X *into itself, then, for any* $x_1 \in X$, *the successive iterates* $Tx_1, T^2x_1, \ldots, T^nx_1, \ldots$ *converge to the unique fixed point of* T.

Proof (<u>Optional</u>).

Let x_0 be the unique fixed point of T.

First Note
$$\begin{aligned} d(T^{n+1}(x_1), x_0) &= d(T^{n+1}(x_1), T^{n+1}(x_0)) \\ &\leq d(T^n(x_1), T(x_0)) \\ &= d(T^n(x_1), x_0) \qquad \text{for all } n \in \mathbb{N}. \end{aligned}$$

Thus, $(d(T^n(x_1), x_0))_{n=1}^{\infty}$ is a decreasing sequence of positive real numbers and so converges to some limit $\alpha \geq 0$. It suffices to show $\alpha = 0$, for then $d(T^n(x_1), x_0) \to \alpha = 0$ <u>or</u> $T^n(x_1) \to x_0$. Now, $(T^n(x_1))_{n=1}^{\infty}$ is a sequence of points of x, so by compactness, there exists a subsequence $(T^{n_k}(x_1))_{k=1}^{\infty}$ which converges to some point $y \in X$, and

$$d(y, x_0) = \lim_{k\to\infty} d(T^{n_k}(x_1), x_0) = \alpha.$$

If $\alpha \neq 0$ then $y \neq x$ so

$$\begin{aligned}\alpha = d(y, x_0) &\gneq d(Ty, Tx_0)\\ &= d(Ty, x_0)\\ &= \lim_{k\to\infty} d(T(T^{n_k}x_1), x_0)\\ &= \lim_{k\to\infty} d(T^{n_k+1}(x_1), x_0)\\ &= \alpha.\end{aligned}$$

Thus $\alpha > \alpha$ which is impossible. So $\alpha = 0$ and $T^n(x_1) \to x_0$. ■

In the next theorem we further restrict the domain of T to be a metric space of the form (K,d) where K is a compact convex subset of a normed linear space $(X, \|..\|)$ and $d(x,y) = \|x - y\|$.

RECALL: $K \subset X$ is convex if the line segment joining any two points in K lies entirely in K, that is, if $x, y \in K$ and $\lambda \in [0,1]$, then $\lambda x + (1 - \lambda) y \in K$.

THEOREM 5: *Let K be a compact convex subset of the normed linear space $(X,\|.\|)$ and let $T: K \to K$ be a non-expansive mapping of K into itself* (that is, $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in K$), *then T has a fixed point in K.*

Of course, the uniqueness of the fixed point can no longer be asserted – see Exercise on page 3, or take T to be the identity mapping.

PROOF: Choose $y_0 \in K$ and for each $n \in \mathbb{N}$ let

$$G_n(x) = \left(\frac{1}{n+1}\right)y_0 + \left(\frac{n}{n+1}\right) Tx \quad \text{for all } x \in K.$$

Firstly note that $G_n(x)$ is a convex combination of the two points y_0 and Tx of K and so, since K is convex $G_n(x) \in K$. Thus G_n is a mapping from K into itself.

Further, for $x,y \in K$ we have

$$\begin{aligned}\|G_n(y) - G_n(x)\| &= \|y_0 + nTy - (y_0 + nTx)\|/(n+1)\\ &= \frac{n}{n+1}\| Ty - Tx \|\\ &\leq \frac{n}{n+1} \|y - x\| \text{ (as T is non-expansive)}\end{aligned}$$

Thus G_n is a strict contraction $\left(\frac{n}{n+1} < 1\right)$ and so, by Theorem 2 has a unique fixed point x_n in K.

That is

$$x_n = G_n(x_n) = \frac{1}{1+n} y_0 + \frac{n}{n+1} Tx_n$$

and so

$$x_n - Tx_n = \frac{1}{n+1} y_0 + \frac{n}{n+1} Tx_n - \frac{n+1}{n+1} Tx_n$$

$$= \frac{1}{n+1} (y_0 - Tx_n).$$

Now, by the compactness of K the sequence (x_n) has a subsequence (x_{n_k}) convergent to some point x_0 of K. Since T is continuous we have

$$\begin{aligned} \|x_0 - Tx_0\| &= \operatorname*{limit}_{k\to\infty} \|x_{n_k} - Tx_{n_k}\| \\ &\leqslant \operatorname*{limit}_{k\to\infty} \frac{1}{n_k+1} \|y_0 - Tx_{n_k}\| \\ &= \left(\operatorname*{limit}_{k\to\infty} \frac{1}{n_k+1}\right) \|y_0 - Tx_0\| \\ &= 0 \end{aligned}$$

or $x_0 = Tx_0$ and so x_0 is a fixed point of T. ■

The above proof is taken from an article by Dotson and Mann in the American Mathematical Monthly. It establishes a special case of the much more general (and more difficult to prove) Schauder fixed point theorem which states:

A <u>*continuous*</u> *mapping of a compact convex subset of a normed linear space into itself has a fixed point.*

§2. AN APPLICATION -

the local existence and uniqueness of solutions for initial value problems of systems of first order ordinary differential equations.

THE PROBLEM: For the simultaneous system of first order ordinary differential equations

$$\left.\begin{aligned}
&\frac{du_1}{dt} = F_1(u_1(t),\ u_2(t),\ \ldots,\ u_n(t),t)\\
&\frac{du_2}{dt} = F_2(u_1(t),\ u_2(t),\ \ldots,\ u_n(t),t)\\
&\ldots\\
&\frac{du_n}{dt} = F_n(u_1(t),\ u_2(t),\ \ldots,\ u_n(t),t)\\
&\text{with the initial conditions}\\
&u_1(t_0) = u_{10}\\
&u_2(t_0) = u_{20}\\
&\ldots\\
&u_n(t_0) = u_{n0}
\end{aligned}\right\}\quad \ldots\ldots\ (1)$$

We are interested in conditions which will ensure the existence of a unique solution

$$\begin{aligned}
u_1(t) &= \phi_1(t)\\
u_2(t) &= \phi_2(t)\\
&\ldots\\
u_n(t) &= \phi_n(t)
\end{aligned}$$

"locally" in some neighbourhood of the initial point t_0; that is, for $t \in (t_0 - h,\ t_0 + h)$, where h is some, sufficiently small, strictly positive number

REMARK: In the special case $n = 1\cdot$ we are of course considering the uniqueness-existence question for the initial value problem of a single first order ordinary differential equation - for a discussion of this simplest case, which might help you to understand the general case considered here, you might look at

Boyce and DiPrima, "Elementary Differential Equations and Boundary Value Problems" (Wiley), sections 2.11 and 2.12.

The consideration of a system rather than a single equation introduces little in the way of extra work or difficulty and is worthwhile for at least two reasons:-

(i) Very frequently in applications, systems of equations arise naturally. This would be true for example in: electrical circuit theory, engineering control theory, the rates of chemical reactions and ecological or physiological models;

(ii) An n'th order ordinary differential equation can always be replaced by a system of n simultaneous first order equations.

Given the n'th order equation

$$\left.\begin{aligned}&\frac{d^n u}{dt^n} = f\left(u, \frac{du}{dt}, \frac{d^2u}{dt^2}, \ldots, \frac{d^{(n-1)}u}{dt^{(n-1)}}, t\right)\\ &\text{with initial conditions}\\ &u(t_0) = u_0, \quad \frac{du}{dt}(t_0) = u_0', \ldots, \frac{d^{n-1}u}{dt^{(n-1)}}(t_0) = u_0^{(n-1)}\end{aligned}\right\} \quad \ldots\ldots (2)$$

let $u_1 = u, \quad u_2 = \frac{du}{dt}, \quad u_3 = \frac{d^2u}{dt^2}, \quad \ldots, \quad u_n = \frac{d^{(n-1)}u}{dt^{(n-1)}}$,

then (2) is equivalent to the system

$$\begin{aligned}\frac{du_1}{dt} &= u_2\\ \frac{du_2}{dt} &= u_3\\ &\ldots\\ \frac{d\,u_n}{dt} &= f(u_1, u_2, \ldots, u_n, t)\end{aligned}$$

with the initial conditions

$$\begin{aligned}u_1(t_0) &= u_0\\ u_2(t_0) &= u_0'\\ &\ldots\\ u_n(t_0) &= u_0^{(n-1)} .\end{aligned}$$

So by establishing existence-uniqueness for a system of equations we are, as a special case, establishing the result for n'th order equations.

<u>NOTATION</u>: It is convenient to rewrite (1) in vector notation. Let

$$\underset{\sim}{u} \equiv \underset{\sim}{u}(t) = \begin{bmatrix}u_1(t)\\ u_2(t)\\ \vdots\\ u_n(t)\end{bmatrix}$$

Thus, the k'th "component" of $\underset{\sim}{u}$ is the function $u_k(t)$.

Let us also agree to write

$$\frac{d\underset{\sim}{u}}{dt} \quad \text{for} \quad \begin{bmatrix} \frac{du_1}{dt} \\ \frac{du_2}{dt} \\ \vdots \\ \frac{du_n}{dt} \end{bmatrix}, \qquad \int_0^x \underset{\sim}{u}(t)dt \quad \text{for} \quad \begin{bmatrix} \int_0^x u_1(t)dt \\ \int_0^x u^2(t)dt \\ \vdots \\ \int_0^x u_n(t)dt \end{bmatrix}$$

etc.

Then (1) may be rewritten as

$$\frac{d\underset{\sim}{u}}{dt} = F(\underset{\sim}{u}(t),t), \quad \underset{\sim}{u}(t_0) = \underset{\sim}{u}_0 \qquad \ldots\ldots \ (1)$$

A SIMPLIFICATION: Observe that under the transformations

$$x = t - t_0, \quad \underset{\sim}{y}(x) = \underset{\sim}{u}(x + t_0) - \underset{\sim}{u}_0{}^{*}$$

and

$$\underset{\sim}{f}(\underset{\sim}{y}(x),\ x) = \underset{\sim}{F}(\underset{\sim}{y}(x) + \underset{\sim}{u}_0,\ x + t_0)$$

the system (1') becomes

$$\frac{d\underset{\sim}{y}}{dx} = \underset{\sim}{f}(\underset{\sim}{y}(x),\ x), \quad \underset{\sim}{y}(0) = \underset{\sim}{0} \qquad \ldots\ldots \ (1").$$

Henceforth, we will assume that our system has been reduced to this form with initial point 0 and all initial values also 0.

INTRODUCTION OF AN APPROPRIATE SPACE:

For any real number $h > 0$, let us denote by X_h the set of vectors

$$\underset{\sim}{y}(x) = \begin{bmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{bmatrix}$$

all of whose "component" functions $y_k(x)$ are real valued continuous mappings on $[-h,h]$.

* here, subtraction has the obvious meaning.

That is

$$X_h = \underbrace{C[-h,h] \times C[-h,h] \times \ldots \times C[-h,h]}_{n \text{ "factors"}}$$

It is readily seen that X_h is a vector (or linear) space with addition and scalar multiplication being defined in the obvious way, and that

$$\|\underset{\sim}{y}\| = \max_{k=1,2,\ldots,n} \|y_k\|_\infty = \max_{k=1,2,\ldots,n} \max_{|x|\leqslant h} |y_k(x)|$$

defines a norm on X_h with respect to which it is a Banach (or complete) space.

FORMULATION AS A SYSTEM OF INTEGRAL EQUATIONS:

Henceforth, we will <u>assume</u> that each of the functions $f_1, f_2, \ldots, f_n$ is a continuous mapping from (A,d_2) into $\mathbb{R}$, where A is the "(n+1)-dimensional cube" of "side length" $2a > 0$:

$$A = \{(y_1,y_2,\ldots,y_n,x) \in \mathbb{R}^{n+1}: |x| \leqslant a \text{ and } |y_k| \leqslant a \text{ for } k = 1,2,\ldots,n\}.$$

For each $h \in (0,a]$ let $M_h = \{\underset{\sim}{y} \in X_h: \|\underset{\sim}{y}\| \leqslant a\}$,

then it is readily verified that M_h is a closed subset of X_h and so M_h is a complete metric space with respect to the metric

$$d(\underset{\sim}{y},\underset{\sim}{z}) = \|\underset{\sim}{y}-\underset{\sim}{z}\|.$$

Using the assumptions on each f_k, motivation for the definition of M_h comes from the <u>observation</u> that for $\underset{\sim}{y} \in M_h$ the composite functions $f_k(\underset{\sim}{y}(x),x)$ are continuous functions of x for $|x| \leqslant h$. That is for $\underset{\sim}{y} \in M_h$, $\underset{\sim}{f}(\underset{\sim}{y}(x),x) \in X_h$.

LEMMA: $\underset{\sim}{\phi} \in M_h$ <u>is a solution of (1") if and only if</u>

$$\underset{\sim}{\phi}(x) = \int_0^x \underset{\sim}{f}(\underset{\sim}{\phi}(t),t)dt \qquad \ldots (3)$$

Proof. (⇐)

Since, by the above observation, for $k = 1,2,\ldots,n$, $f_k(\underset{\sim}{\phi}(x),x)$ is a continuous function, the fundamental theorem of calculus applies to give that the R.H.S. (and hence, also the L.H.S.) of

$$\phi_k(x) = \int_0^x f_k(\underset{\sim}{\phi}(t),t)dt$$

is differentiable and

$$\frac{d\phi_k}{dx} = f_k(\underset{\sim}{\phi}(x),x).$$

That is,

$$\frac{d\underset{\sim}{\phi}}{dx} = \underset{\sim}{f}(\phi(x),x).$$

It is also clear that

$$\underset{\sim}{\phi}(0) = \int_0^0 \underset{\sim}{\phi}(t)dt = \underset{\sim}{0},$$

and so (1") is satisfied.

(⇒)

Again the above observation shows that, if

$$\frac{d\phi_k}{dx} = f_k(\phi(\underset{\sim}{x}),x)$$

and $\phi_k(0) = 0$ for $k = 1,2,\ldots,n,$ then both sides are continuous and so by the fundamental theory of calculus

$$\phi_k(x) = \phi_k(x) - \phi_k(0) = \int_0^x \frac{d\phi_k}{dt}\,dt$$
$$= \int_0^x f_k(\underset{\sim}{\phi}(t),t)dt.$$

Thus, $\underset{\sim}{\phi}(x) = \int_0^x \underset{\sim}{f}(\underset{\sim}{\phi}(t),t)dt.$ ■

OPERATOR REFORMULATION:

For any $h \in (0,a]$ define the operator (mapping) $T: M_h \to X_h$ by

$$T(\underset{\sim}{\psi})(x) = \int_0^x \underset{\sim}{f}(\underset{\sim}{\psi}(t),t)dt.$$

Then the above lemma may be restated as $\underset{\sim}{\phi} \in M_h$ <u>is a solution of (1") if and only if</u>

$$\underset{\sim}{\phi} = T\underset{\sim}{\phi},$$

that is, if and only if $\underset{\sim}{\phi}$ is a <u>fixed point</u> of T.

Thus, to establish a local existence uniqueness theorem for (1"), and hence (1), it suffices to show that for a sufficiently small h, T has a unique fixed point in M_h. In turn, this will follow from Banach's fixed point theorem provided we can find a value of h such that T is a strict contraction of M_h into itself.

We first show,

FOR h SUFFICIENTLY SMALL T MAPS M_h INTO ITSELF:
Now, since for $k = 1,2,\ldots,n$, $f_k(\underset{\sim}{y},x)$ is by assumption continuous on the compact set A (Heine-Borel Theorem), there exists $m_k > 0$ such that $|f_k(\underset{\sim}{y}\ ,x)| \leqslant m_k$ for all $(\underset{\sim}{y},\ x) \in A$.
Let $m = \text{Max}\{m_1,m_2,\ldots,m_n\}$, then for $\underset{\sim}{\psi} \in M_h$

$$\|T\underset{\sim}{\psi}\| = \|\int_0^x \underset{\sim}{f}(\underset{\sim}{\psi}(t),t)dt\|$$

$$\leqslant \left|\int_0^x m\ dt\right| \quad \text{(by the definition of } \|\cdot\|)$$

$$= m|x|$$

$$\leqslant mh.$$

Thus, provided we choose $h \leqslant \frac{a}{m}$ we have

$$\|T\underset{\sim}{\psi}\| \leqslant a \quad \text{or} \quad T\underset{\sim}{\psi} \in M_h.$$

The final step is to choose h so that T is a strict contraction. Regretably this is not, in general, possible without further restricting $\underset{\sim}{f}$. We will assume that $\underset{\sim}{f}$ satisfies a Lipschitz's condition in the first variable, that is, for some $K > 0$

$$|f_k(\underset{\sim}{y}_1,x) - f_k(\underset{\sim}{y}_2,x)| \leqslant K\|\underset{\sim}{y}_1-\underset{\sim}{y}_2\|_\infty$$

for $(\underset{\sim}{y}_1,x),(\underset{\sim}{y}_2,x) \in A$ and $k = 1,2,\ldots,n.$*

*This is equivalent to requiring each of the functions f_k satisfy a Lipschitz condition (K is the maximum of the Lipschitz constants for the individual f_k). A sufficient (and often used) condition for this to happen is that $\frac{\partial f_k}{\partial y_i}$ be continuous in A, as may easily be seen upon application of the mean value theorem.

Under this assumption we have for any $\underset{\sim}{\phi}, \underset{\sim}{\psi} \in M_h$

$$\|T_{\underset{\sim}{\phi}} - T_{\underset{\sim}{\psi}}\| = \|\int_0^x \underset{\sim}{f}(\underset{\sim}{\phi}(t),t)dt - \int_0^x \underset{\sim}{f}(\underset{\sim}{\psi}(t),t)dt\|$$

$$= \underset{k=1,2,\ldots,n}{\text{Max}} \ \underset{|x|\leqslant h}{\text{Max}} \left|\int_0^x (f_k(\underset{\sim}{\phi}(t),t) - f_k(\underset{\sim}{\psi}(t),t))dt\right|$$

$$\leqslant \underset{k=1,2,\ldots,n}{\text{Max}} \ \underset{|x|\leqslant h}{\text{Max}} \left|\int_0^x |f_k(\underset{\sim}{\phi}(t),t) - f_k(\underset{\sim}{\psi}(t),t)|dt\right|$$

$$\leqslant \underset{|x|\leqslant h}{\text{Max}} \left|\int_0^x K\|\underset{\sim}{\phi} - \underset{\sim}{\psi}\|dt\right|$$

$$= Kh\|\underset{\sim}{\phi} - \underset{\sim}{\psi}\|$$

and so, for $h < \frac{1}{K}$, T is a strict contraction.

Thus, taking $0 < h < \text{Min}\left\{a, \frac{a}{m}, \frac{1}{K}\right\}$ we have proved:

THEOREM: *If* $f: A \subseteq \mathbb{R}^{n+1} \to \mathbb{R}^n$ *is continuous on the "cube"* $A = \{(\underset{\sim}{y}, x): |x| \leqslant a, \|\underset{\sim}{y}\|_\infty \leqslant a\}$ *and satisfies the Lipschitz condition*

$$\|\underset{\sim}{f}(\underset{\sim}{y}_1, x) - \underset{\sim}{f}(\underset{\sim}{y}_2, x)\|_\infty \leqslant K\|\underset{\sim}{y}_1 - \underset{\sim}{y}_2\|$$

on A, *then there exist* $h > 0$ *such that the initial value problem*

$$\frac{d\underset{\sim}{y}}{dx} = \underset{\sim}{f}(\underset{\sim}{y}(x), x), \quad \underset{\sim}{y}(0) = \underset{\sim}{0}$$

has a unique solution for $-h \leqslant x \leqslant h$.

REMARK: From the proof of Banach's fixed point theorem and the remarks following it, for T defined by

$$T\underset{\sim}{\phi}(x) = \int_0^x \underset{\sim}{f}(\underset{\sim}{\phi}(t),t)dt$$

and starting with any initial "guess" $\underset{\sim}{\phi}_0$, the sequence of iterates

$$\underset{\sim}{\phi}_0, T\underset{\sim}{\phi}_0, T^2\underset{\sim}{\phi}_0, \ldots, T^n\underset{\sim}{\phi}_0, \ldots$$

form successively better approximations (known as Picard's approximations) to the solution of $\frac{d\underset{\sim}{y}}{dx} = \underset{\sim}{f}(\underset{\sim}{y}(x),x)$, $\underset{\sim}{y}(0) = \underset{\sim}{0}$.

EXERCISES:

(i) Transform the initial value problem $u' = u$, $u(0) = 1$ into the form $y' = f(y)$ $y(0) = 0$.

(ii) Obtain the first five successive Picard approximations to the solution of $y' = f(y)$, $y(0) = 0$.. Start with the initial approximation $\phi_0(x) \equiv 0$.

(iii) Show that f is continuous and satisfies a Lipschitz condition. Determine a range of x-values for which the above theory guarantees the Picard approximations converge to a unique solution of $y' = f(y)$, $y(0) = 0$.

(iv) Using the error estimates obtained in the discussion following the proof of Banach's fixed point theorem, determine the maximum number of Picard approximations which need to be computed if the solution is to be approximated with an error of no more than 0.01 throughout the interval determined in (iii).

APPENDIX

IMPLICIT FUNCTIONS

Our aim is to illustrate the power of fixed point Theorems by proving, via the Banach Fixed point Theorem, the following simple

IMPLICIT FUNCTION THEOREM.

Let $x, y \in \mathbb{R}$ *be related by*

(1) $y = ax + R(y)$ *where* $R(0) = 0$ *and for* $|y| < r$

R *satisfies the Lipschitz Condition*

$$|R(y_1) - R(y_2)| \leqslant k|y_1 - y_2|$$

where k *is a fixed constant with* $0 < k < 1$.

Then there exists a unique continuous function f *with* $f(0) = 0$ *and domain* $D = \{x: |x| \leqslant \rho < \frac{1-k}{|a|} r\}$ *such that* $y = f(x)$, *all* $x \in D$.

i.e. the relation implies y is functionally related to x at least in a neighbourhood of 0.

Proof. If a solution exists it will belong to the subspace X of $C\,[-\rho, \rho]$ consisting of those functions g with $g(0) = 0$ and $\|g\|_\infty = \underset{|x|\leqslant\rho}{\text{Max}} |g(x)| \leqslant r$.

with the induced metric, $d_\infty(g, h) = \underset{|x|\leqslant\rho}{\text{Max}} |g(x) - h(x)|$ for all $g, h \in X$.

It is easily verified that (X, d_∞) is a *complete* metric space.
Further observe that f is a solution if and only if

$T(f) = f$ where T is the operator on X defined by
$T(g)(x) = ax + R(g(x))$ for all $|x| < \rho$, $g \in X$.

Thus, provided we can show T is a *strict contraction* mapping into X, the desired result will follow upon invoking the Banach fixed point Theorem.
But $\|T(g)\|_\infty \leqslant |a|\,|x| + k\|g\|_\infty \leqslant |a|\,|\rho| + kr \leqslant r$ by the choice of $\rho\left(< \frac{1-k}{|a|} r\right)$ so $T(g) \in X$.

Further $d_\infty(T(g), T(h)) = \underset{|x|\leqslant\rho}{\text{Max}} |R(g(x)) - R(h(x))|$

$$\leqslant k \underset{|x|\leqslant\rho}{\text{Max}} |g(x) - h(x)| = kd_\infty(g, h),$$

~~so T is a strict contraction and the result follows.~~ ■

Application: INVERSE FUNCTION THEOREM

Take $f \in C^1\,(x_0 - r_1, x_0 + r_1)$ *with* $f'(x_0) \neq 0$. *If* $f(x_0) = y_0$ *we aim to show there exists a unique function* g, *domain* $D \equiv (y_0 - r_2, y_0 + r_2)$ *for some* $r_2 > 0$, *such that if* $y = f(x)$ *then* $x = g(y)$ *all* $y \in D$.

i.e. f is invertible on a neighbourhood of x_0.

It suffices to show that such a g must satisfy a relation of the form (1). The following reasoning is due to Édouard Goursat (1858 - 1936) in 1903.

We rewrite $y = f(x)$ as

$$x - x_0 = f'(x_0)^{-1}(y - y_0) - R(x)$$

where $$R(x) = f'(x_0)^{-1}[f(x) - f(x_0)] - (x - x_0).$$

This is of precisely the right form, and further, from the continuity of f' there exists $r_3 \in (0, r_1]$ such that

$$|f'(x) - f'(x_0)| < \tfrac{1}{2}|f'(x_0)| \text{ all } x \text{ with } |x - x_0| < r_3$$

whence, for x_1, $x_2 \in (x_0 - r_3, x_0 + r_3)$ we have

$$\begin{aligned}|R(x_1) - R(x_2)| &= |f(x_1) - f(x_2) - (x_1 - x_2)\, f'(x_0)|\,|f'(x_0)|^{-1} \\ &= \left|\int_{x_2}^{x_1} [f'(x) - f'(x_0)]dx\right|\left|f'(x_0)\right|^{-1} \\ &\leqslant \int_{x_2}^{x} |f'(x) - f'(x_0)|dx|f'(x_0)|^{-1} \\ &\leqslant \tfrac{1}{2}|f'(x_0)|\,|x_1 - x_2|\,|f'(x_0)|^{-1}\end{aligned}$$

by choice of r_3

and so R satisfies the required Lipschitz condition

$$|R(x_1) - R(x_2) \leqslant \tfrac{1}{2}|x_1 - x_2|.$$

Thus an application of our simple implicit function Theorem gives a unique h such that

$x - x_0 = h(y - y_0)$ for all x with $|x - x_0| \leqslant r_2 = \max\{r_3, \tfrac{1}{2}|f'(x_0)|r_3\}$

whence $x = g(y) \equiv x_0 + h(y - y_0)$ as required. ■

REMARKS: (1) The proof of our simple implicit function theorem may be trivially extended to cover the case where $R \equiv R(x, y)$ provided the Lipschitz condition

$$|R(x_1, y_1) - R(x_2, y_2) \leqslant k[|x_1 - x_2| + |y_1 - y_2|], \; 0 < k < 1,$$

is satisfied for $|x_1|$, $|x_2| < r_1$ and $|y_1|$, $|y_2| < r_2$ some $r_1, r_2 > 0$.

(2) Under appropriate assumptions on F, Goursat's arguments can be combined with this extended implicit function Theorem to obtain a version of the usual implicit function Theorem:

$$F(x, y) \equiv 0 \Rightarrow y = f(x) \text{ some function } f.$$

The calculations are however considerably more involved.

(3) Both versions of the Implicit function Theorem considered remain valid if x and y are allowed to be elements of a Banach algebra, the only difference in the proof being the replacing of $|\cdot|$ by the norm in the appropriate places, while the Inverse function Theorem extends to cover complex valued functions of a complex variable.

Collateral Reading.

Essentially, the above considerations were extracted from the two books of Einar Hille, *"Analytic Function Theory"* Vol. I Gin, Boston, 1959 and *"Methods in Classical and Functional Analysis"*, Addison-Wesley, Massachusetts, 1972.

EXERCISE.

Let x and y be related implicitly by

$$x^3 + y^3 + x - y = 0.$$

Establish the existence of an $r > 0$ and function $f \in C(-r, r)$ with $f(0) = 0$, such that

$$y = f(x) \text{ for all } x \in (-r, r).$$

[Hint: Consider $T: X \to X$ where

$$T(g)(t) = t + t^3 + g^3(t) \text{ for all } g \in X,$$

a suitable subspace of $C(-r, r)$.]

SECTION 3. BAIRE'S CATEGORIZATION OF METRIC SPACES

0. Heuristic Outline

In 1899 the French mathematician, René Baire, developed a method for classifying the "size", in an appropriate sense, of subsets in a metric space. Intuitively we may consider subsets divided into three categories (for this reason, the work is frequently referred to as "Baire's Category* theory"):

"minute" - (these are formally defined below and referred to as *nowhere dense sets*)

"small" - (these are those sets which can be constructed in a suitable way from minute sets. They will be subsequently termed *meagre sets*. In much of the literature they are also known as sets of the first Baire category, or just first category.)

"large" - (a set which is not "small" is "large". These sets are often said to be of the second (Baire) category. We will refer to such sets simply as *non-meagre sets*.).

Baire's category theory has proved to be a valuable tool for the establishment of "pure existence results". That is, the existence of certain objects is established by methods which give no hint as to how these objects may be constructed and so provide no specific example of such an object. (This should be contrasted with the existence theorem for differential equations derived from Banach's fixed point theorem. There, not only is existence established but a method for constructing the solution, as a limit of certain iterates, is provided - this is a "constructive existence proof".)

To illustrate this use of Baire's category theory we will give one typical application, originally due to Banach (1931); establishing the existence of continuous functions which are not differentiable at any point. In essence the proof goes as follows.

Using Baire's Category Theorem: *Every complete metric space is non-meagre*, which is established below, we deduce that the space of continuous functions C[a,b] with the uniform metric is a non-meagre set. We then proceed to show that the subset D consisting of all continuous functions which are differentiable at at least one point of the interval [a,b] is a meagre subset. Since it follows from the formal definition of meagre sets that the union of two meagre sets is again a meagre set, we conclude that the

*Not to be confused with the more recently developed algebraic notion of *categories*.

compliment of D in C[a,b], C[a,b]\D, is non-meagre (otherwise, C[a,b] = (C[a,b]\D) ∪ D being the union of two meagre sets would itself be meagre, a contradiction, as we have already found it to be non-meagre). *But*, this complement comprises precisely those continuous functions which have no point of differentiability in [a,b], and so the continuous nowhere-differentiable functions constitute a "large" set. Certainly then such functions exist. [Since a large percentage of function theory is concerned with differentiable functions, we may say, "the majority of mathematics deals with a minority of functions." This came as somewhat of a jolt to mathematicians, though the ground had been somewhat softened by Cantor's researches into infinite sets which lead to analogous and at the time even more revolutionary conclusions. The mere existence of continuous nowhere-differentiable functions did not come as a shock, earlier the German mathematician Karl Weierstrass [1815-1897] had produced such a function, for example the function defined by

$$f(x) = \sum_{n=1}^{\infty} \frac{1}{10^n} \{10^n x\},$$

where $\{r\}$ denotes the "distance" from r to the nearest integer, thus $\{1\frac{3}{4}\} = \{5\frac{1}{4}\} = \frac{1}{4}$, $\{\frac{1}{2}\} = \frac{1}{2}$ etc.

<u>EXERCISE</u>: (a) Draw graphs of $\frac{1}{3^n}\{3^n x\}$ for $n = 1$, 2 and 3. Hence, construct the graph of

$$\sum_{n=1}^{3} \frac{1}{3^n}\{3^n x\}.$$

Can you see why $f(x)$, defined above, might prove to be nowhere-differentiable?

*(b) Prove that the function $f(x)$, defined above, is continuous on [0,1]

[Hint: Note that $\frac{1}{10^n}\{10^n x\}$ is continuous for each n, and so deduce that $f_N(x) = \sum_{n=1}^{N} \frac{1}{10^n}\{10^n x\}$ is a continuous function for each N. Finally conclude that f is continuous by showing that $f_N \xrightarrow{\text{uniformly}} f$ as $N \to \infty$. To do this, note that $\sup_{0\leqslant x\leqslant 1} |f(x) - f_N(x)| \leqslant \sum_{n=N+1}^{\infty} \frac{1}{10^n}$.]

**(c) (Optional) Establish that the function $f(x)$, defined above, is not differentiable at any point of [0,1].

**(c) (cont'd)

[Hint: For any $r \in [0,1]$ find a sequence $x_n \to r$ such that the appropriate difference quotients

$$\frac{f(x_n) - f(r)}{x_n - r}$$

do not converge. You may find it useful to consider numbers represented as decimals. A proof may be found in Spivak's "Calculus".]

1. Baire's Category Theory

Throughout (X,d) will denote a metric space.

RECALL, a subset B of (X,d) is *dense* if its closure $\overline{B}$ is the whole of X - intuitively this means that the points of B are distributed "thickly" throughout the space, B contains points as near as we like to every point of X.

Since an open ball consists of all points of X nearer to the centre than some given amount, we may think of an open ball (and hence any set with non-empty interior and so containing an open ball) as representing a "solid or substantial lump of the space". A set C whose closure contains an open ball (that is, $\operatorname{int} \overline{C} \neq \emptyset$) therefore has points which are "thickly" distributed throughout some "substantial lump" of the space. Such a set might be thought of as being *somewhere dense*. This motivates the following definition.

DEFINITION 1: A subset E of the metric space (X,d) is *nowhere dense* if its closure has empty interior, that is, if $\operatorname{int} \overline{E} = \emptyset$.

EXAMPLES: 1. In $\mathbb{R}$ with the usual metric the following subsets are nowhere dense.

(a) Any set consisting of a single point;

(b) Any finite set;

(c) $\mathbb{N}$ the set of natural numbers;

2. In $\mathbb{R}^2$ with the euclidean metric the set of points on a line is a nowhere dense set.

[As an EXERCISE you should verify these assertions.]

In our subsequent work we will make use of the following consequence of the above definition.

LEMMA 2: *Let* E *be a nowhere dense subset of the metric space* (X,d), *then if* G *is any non-empty open subset of* (X,d) *there exists a point* $x \in G$ *and* $r > 0$ *such that* $B_r(x) \subseteq G$ *and* $B_r(x) \cap E = \emptyset$.