A First Look at Numerical Functional Analysis Notes

Kyle J. Bunkers

December 16, 2016

Contents

Chapter 1

A first course in functional analysis

A nice introduction to what and why we are doing things.

Chapter 2

Old ideas in new contexts

Note for the first exercise given in the text (unnumbered) that the certain equation is $x(x^2-3)=0$ that yields

$$x_{n+1} = \frac{2x_n^3}{3(x_n^2-1)} \tag{2.1}$$

I used python to program this iteration

chapter2/iteration.py

```python
#!/usr/bin/env python2

import numpy as np

# Solve a certain equation iwth x_{n+1} = 2x_n^3/(3(x_n^2-1))
def iterate(xin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  xg1=xin+2*tol
  while ((np.abs(xg2-xg1)>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=2/3.*xg1**3/(xg1**2-1)
    itercount+=1
  return xin,xg2,itercount

print iterate(0.81)
print iterate(0.78)
print iterate(0.779)
print iterate(0.775)
print iterate(0.7747)
print iterate(0.77462)
print iterate(0.7746)
print iterate(0.77)
```

And the output is given by table 2.1

Where we note that $\sqrt{3} \approx 1.7320508075688772$.

2.1 Exercise 1

Prove that the ℓ_1 and ℓ_∞ definition in two dimensions do satisfy the distance axioms (1) and (5).

(1) The distance from A to B is measured (in some suitable unit) by a real number $d(A,B)$. (5) $d(A,C)+d(C,B) \geq d(A,B)$ (Triangle inequality)

input	output	iterations
0.81	-1.7320508075688772	12
0.78	1.7320508075688774	15
0.779	-1.7320508075688772	12
0.775	1.7320508075688772	8
0.7747	-1.7320508075688772	12
0.77462	1.7320508075688774	20
0.7746	-1.7320508075692274	15
0.77	-1.1526368786890824$\times 10^{-51}$	7

Table 2.1: Input and Output for `iteration.py`.

ℓ_∞ distance is given by $x_0, x_1, y_0, y_1 \in \mathbb{R}$, $d = \max\{|x_1 - x_0|, |y_1 - y_0|\}$ whereas ℓ_1 is given by $d = |x_1 - x_0| + |y_1 - y_0|$

Solution:

First, for (1), the operations given will never yield anything but real numbers given real numbers as inputs. That is max and $|\cdot|$ only yield real numbers as output when given real numbers as input.

Consider distinct points (if they aren't distinct, the proof is trivial) $A = (x_0, y_0)$, $B = (x_1, y_1)$ and a third point $C = (x_2, y_2)$. Then, for ℓ_∞ we have

$$\begin{aligned} d(A,C) + d(C,B) &= \max\{|x_0 - x_2|, |y_0 - y_2|\} + \max\{|x_2 - x_1|, |y_2 - y_1|\} && (2.2)\\ d(A,B) &= \max\{|x_0 - x_1|, |y_0 - y_1|\} && (2.3)\end{aligned}$$

We can use $|x + y| \leq |x| + |y|$. Rewrite

$$\begin{aligned} d(A,B) &= \max\{|x_0 - x_1|, |y_0 - y_1|\}\\ &= \max\{|x_0 - x_2 + x_2 - x_1|, |y_0 - y_2 + y_2 - y_1|\} \leq \max\{|x_0 - x_2| + |x_2 - x_1|, |y_0 - y_2| + |y_2 - y_1|\}\\ &\leq \max\{|x_0 - x_2|, |y_0 - y_2|\} + \max\{|x_2 - x_1|, |y_2 - y_1|\} = d(A,C) + d(C,B)\end{aligned} \tag{2.4}$$

where we have used (for $a, b, c, d > 0$)

$$\max\{a + b, c + d\} \leq \max\{a, c\} + \max\{b, d\} \tag{2.5}$$

For ℓ_1 we instead have

$$\begin{aligned} d(A,C) + d(C,B) &= |x_0 - x_2| + |y_0 - y_2| + |x_2 - x_1| + |y_2 - y_1| && (2.6)\\ d(A,B) &= |x_0 - x_1| + |y_0 - y_1| && (2.7)\end{aligned}$$

We now use that $|x + y| \leq |x| + |y|$ so that using $x_0 - x_1 = (x_0 - x_2) + (x_2 - x_1)$ we find

$$\begin{aligned} d(A,B) &= |(x_0 - x_2) + (x_2 - x_1)| + |(y_0 - y_2) + (y_2 - y_1)| \leq |x_0 - x_2| + |x_2 - x_1| + |y_0 - y_2| + |y_2 - y_1|\\ &= d(A,C) + d(C,B)\end{aligned} \tag{2.8}$$

proving (5).

2.2 Exercise 2

For the ℓ_2 definition of distance in two dimensions, properties (1) to (4) are easily seen to be true, and we know property (5) as a theorem in geometry. It must be possible to prove property (5) from the definition above purely by means of algebra, but it is not bovious how to do so. Conisder this problem.

Solution:

This is easy if we allow ourselves to define an inner product and use the Cauchy-Inequality, but otherwise we run into problems due to

$$(x+y)^2 \nleq x^2 + y^2 \tag{2.9}$$

Later in the book we will figure this out via Minkowski's inequality, which is certainly not a trivial proof.

2.3 Exercise 3

A signal sent by telegraph or fed into a computer may be represented by a series of noughts and ones. 0 indicating *no pulse* and 1 denoting *a pulse*. In the study of errors that may arise in transmission the Hamming distance is defined as the number of errors in the first signal that would be required to turn it into the second signal. For instance the distance between the signals

$$\begin{array}{ccccc} * & & * & * & \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{array} \tag{2.10}$$

is 3; the signals differ at the points marked *. We will assume that all signals are of length 5 as in the example above.

2.3.1 Axioms of Distance

Does the Hamming definition satisfy the axioms of distance and so define a metric space of signals?

Solution:

(1) Clearly, we will only ever get non-negative integers out, which are real numbers so (1) is satisfied.

(2) We only get non-negative integers so (2) is satisfied.

(3) The only way to get a distance of 0 is when the two signals are the same, so (3) is satisfied.

(4) The order of evaluating the number of "errors" makes no difference for the distance as defined. Thus (4) is satisfied.

(5) If we consider three signals, a, b, c, that are distinct from each other then we need to show

$$d(a,b) \leq d(a,c) + d(c,b) \tag{2.11}$$

(if any of a, b, c are non-distinct then it clearly follows because one of the distances on the right-side will be zero, and we can replace the appropriate variable so that both sides are identical).

Consider this bit by bit. If $d(a_i, b_i) \leq d(a_i, c_i) + d(c_i, b_i)$ for every bit, then clearly $d(a, b) \leq d(a, c) + d(c, b)$ will be true. We have a few cases:

- $a_i = b_i$: $d(a_i, b_i) = 0$ and so clearly the right hand side will be greater than or equal to this as our distances must be positive.
- $a_i \neq b_i, a_i = c_i$ (so $b_i \neq c_i$): Now $d(a_i, b_i) = 1$, and $d(a_i, c_i) = 0$ but $d(b_i, c_i) = d(b_i, a_i) = 1$ and so the inequality holds.
- $a_i \neq b_i, a_i \neq c_i$ (so $b_i = c_i$): Now $d(a_i, b_i) = 1$. Note that as $a_i \neq c_i$ then $c_i = b_i$. Therefore $d(a_i, c_i) = 1$ and $d(b_i, c_i) = d(b_i, b_i) = 0$ and so the inequality holds.

This exhausts all cases, and because the bit-wise triangle inequality applies, we have the triangle inequality applying to the entire thing, because $d(a, b) = \sum_i d(a_i, b_i)$. Thus,

$$d(a, b) = \sum_i d(a_i, b_i) \leq \sum_i [d(a_i, c_i) + d(c_i, b_i)] = d(a, c) + d(c, b) \tag{2.12}$$

Note we could generalize the proof for signals with more than just 1's and 0's by extending the cases above, and come to the same conclusion. We would simply add the case $a_i \neq b_i, a_i \neq c_i, b_i \neq c_i$ where $d(a_i, b_i) = 1$ and $d(a_i, c_i) = 1$ and $d(b_i, c_i) = 1$ so that we would still have the triangle inequality satisfied bitwise.

2.3.2 Number of Signals

Let $a = 10101$. How many signals are there in $S(a, 2)$? in $\overline{B}(a, 2)$?

Solution:

(Note for such a small space we could actually just list the possibilities, but let's use some combinatorics to lessen the task.)

For $S(a, 2)$ these are the number of signals with exactly two differences. So choose 2 out of the 5, or (5 choose 2) in standard notation given by

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5(4)}{2} = 10 \tag{2.13}$$

where as $\overline{B}(a, 2)$ will be given by

$$\sum_{i=0}^{2} \binom{5}{i} = \binom{5}{0} + \binom{5}{1} + \binom{5}{1} = 1 + 5 + 10 = 16 \tag{2.14}$$

Summarizing, $S(a, 2) = 10$ and $\overline{B}(a, 2) = 16$.

We can note that there are $2^5 = 32$ total signals in our space. Thus, we check (using that $\overline{B}(a,5) = 32$ is the entire space)

$$\overline{B}(a,5) = \sum_{i=0}^{5} \binom{5}{i} = \binom{5}{0} + \binom{5}{1} + \binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5} \tag{2.15}$$

$$= 1 + 5 + 10 + 10 + 5 + 1 = 32 \tag{2.16}$$

2.3.3 Spheres

What is $S(a,5)$?

Solution:

This is every possible signal except the one which has an error in every single digit. There are 31 of them from the calculation above, and the only signal not in it would be $b = 01010$.

2.4 Exercise 4

What point (x,y) on the line with equation $2x + y = 5$ is nearest to the origin in the metric ℓ_∞? — in metric ℓ_2? — in metric ℓ_1? What are the distances in these three metrics? When these distances are arranged in order of magnitude does there appear to be any connection with the order of the numbers 1,2,∞?

Solution:

For ℓ_∞ we need the (x,y) that is the minimum of the maximum distance, i.e.,

$$d_{\ell_\infty} = \min\left\{\max\left(|x|, |y|\right)\right\} \tag{2.17}$$

for all possible values of x and y consistent with equation $2x + y = 5$ Note that $y = 5 - 2x$ so we would take the maximum of $|x|$ and $|5-2x|$. We see that the distance function has discontinuities. We can see that between $5/3 < x < 5$ that $|x|$ is larger than $|2x - 5|$. We want the place where the maximum of these two is smallest, so we look for the place where $|x| = |5 - 2x|$ (which we did above) and evaluate, since these are the points that have the smallest possible distance.

Clearly 5/3 is smaller than 5, so that $d_{\ell_\infty} = 5/3 \approx 1.66$ is the minimum value, obtained at $(x,y) = (5/3, 5/3)$.

For ℓ_2 we can use that the distance formula (from the origin) is given by

$$d_{\infty_2} = \sqrt{x^2 + y^2} = \sqrt{x^2 + (5 - 2x)^2} \tag{2.18}$$

We can minimize this (as it is continuous and nice) by taking the derivative and setting it to zero. We can also minimize $d^2_{\infty_2}$ because that will clearly minimize d_{∞_2}

$$\frac{\mathrm{d}d^2_{\infty_2}}{\mathrm{d}x} = 2x + 4(2x - 5) = 10x - 20 = 0 \tag{2.19}$$

$$x = 2y = 1 \tag{2.20}$$

So that the distance is $\sqrt{2^2+1^2}=\sqrt{5}\approx 2.23$ at $(x,y)=(2,1)$.

For ℓ_1 we can use the distance formula and minimize again.

$$d_{\infty_1}=|x|+|5-2x| \tag{2.21}$$

We note that this has discontinuities at $x=0$ and $x=5/2$. Thus, the minima will have to be at one of these two points. WE see $d_{\infty_1}(x=0)=3$ while $d_{\infty_1}(x=5/2)=5/2$. Thus, at the point $(x,y)=(5/2,0)$ we have the minimum distance of $5/2\approx 2.5$.

Altogether,

$$d_{\ell_\infty}=5/3\approx 1.667 \tag{2.22}$$
$$d_{\ell_2}=\sqrt{5}\approx 2.236 \tag{2.23}$$
$$d_{\ell_1}=5/2\approx 2.5 \tag{2.24}$$

So we see that ℓ_∞ gives the smallest distance, with ℓ_1 giving the largest.

2.5 Exercise 5

If the surface of the earth is regarded as a perfect sphere of radius R, distances are measured by the shoretst routes on the *surface* (not by chords through the earth), and N represents the North Pole, what are the usual geographic names of the following?

2.5.1 a

$S(N,\pi R/2)$

Solution:

This is two dimensional, so S is a circle. Thus at $\pi R/2$ we are at the equator, as we are a quarter of a circle around the earth (from the North pole).

2.5.2 b

$B(N,\pi R/2)$

Solution:

This is the Northern Hemisphere, not including the equator. This is because it is the "ball" for the top half of the Earth, but is not closed.

2.5.3 c

$\overline{B}(N,\pi R)$

Solution:

This is the entire surface of the Earth, because it includes every point since πR is where the South pole is located.

2.5.4 d

$B(N, \pi R)$

Solution:

This is everywhere on Earth but the South pole via the same reasoning as above.

2.5.5 e

$S(N, \pi R)$

Solution:

This is the South pole, as this is the only point at this distance from the North pole. It can be viewed as a degenerate circle.

2.6 Exercises Set II

State which of the following are vector spaces, the operations of addition and multiplication being defined in the natural way. For those which are not vector spaces, state an axiom that fails to be satisfied.

2.6.1 Exercise 1

The set of all quadratic expressions, $Q(x) = ax^2 + bx + c$.

Solution:

This is a vector space, when you add quadratics they remain in the quadratics, and when you multiply by scalars and add they remain in the quadratics.

2.6.2 Exercise 2

The set of all quadratic expressions, $Q(x)$, having $Q(0) = 0$.

Solution:

This is also a vector space, because adding two elements from the space will still have $Q(0) = 0$. Note that this just sets $c = 0$ so that $Q(x) = ax^2 + bx$.

2.6.3 Exercise 3

The set of all quadratic expressions, $Q(x)$, having $Q(0) = 0$ and $Q(1) = 0$.

Solution:

This is still a vector space as it simply requires $a = b$, and so we have $Q(x) = ax^2 - ax$ as our space. When we add and multiply by scalars, this will still always be true.

2.6.4 Exercise 4

The set of all quadratic expressions, $Q(x)$, having $Q(0) = 0$ and $Q(1) = 0$ and $Q(2) = 0$.

Solution:

Now we have restricted $Q(x) = a(x^2 - x) = ax(x - 1)$ such that $a = 0$ only. That is there is only one polynomial that satisfies all of these, and that is the zero polynomial. This is technically a vector space given our axioms, but it is a trivial vector space consisting of a single element.

2.6.5 Exercise 5

All the expressions $x^2 + bx + c$.

Solution:

This is not a vector space, consider adding two elements of this space together. $(x^2 + bx + c) + (x^2 + dx + e) = 2x^2 + (b + d)x + (c + e)$. This is not in the vector space despite both elements being in the vector space.

2.6.6 Exercise 6

All polynomials with degree not exceeding 5.

Solution:

For the same reason as the quadratics were a vector space, this is also a vector space.

2.6.7 Exercise 7

All polynomials.

Solution:

This system satisfies all the vector space properties we have required.

2.6.8 Exercise 8

All functions defined on $[0, 1]$ with real values.

Solution:

This should be a vector space. By saying they have real values, I am assuming that they are not bounded, though, given the next exercise. If you add any two real-valued functions in the interval. We also clearly have a scalar identity and function that corresponds to zero.

2.6.9 Exercise 9

All bounded functions $[0, 1] \to \mathbb{R}$. This means that for each function f, there is a number M such that $|f(x)| \leq M$ for $0 \leq x \leq 1$.

Solution:

This is a vector space. If we add any two functions with some bounds, M and N, then that function will be bounded by at least $M + N$.

2.6.10 Exercise 10

All functions $[0, 1] \to \mathbb{R}$ with the bound $M = 100$.

Solution:

Take $f(x) = 99$ and $g(x) = 2$ then $f(x) + g(x) = 101 > 100$ so this is not a vector space.

2.6.11 Exercise 11

All continuous functions $[0, 1] \to \mathbb{R}$.

Solution:

This is a vector space. When you add two continuous functions, the result will be continuous.

2.6.12 Exercise 12

All continuous functions $[0, 1] \to \mathbb{R}$ with $f(0.5) = 0$.

Solution:

This will be a vector space because when you add any two they remain in the space and we have the zero vector.

2.6.13 Exercise 13

All continuous functions $[0, 1] \to \mathbb{R}$ with $f(0.5) = 2$.

Solution:

This is not. If you add the function $f(x) = 2$ to $g(x) = 2(1-x)$ then $f(0.5)+g(0.5) = 2+2-2(0.5) = 3 \neq 2$. Also there is no zero vector.

2.6.14 Exercise 14

All functions, f, with f, f', f'' continuous, $[0, \pi] \to \mathbb{R}$ with f satisfying the differential equation $f''(x) + f(x) = 0$ and the end conditions $f(0) = f(\pi) = 0$.

Solution:

The differential equation's solution is

$$f(x) = A\cos(x) + B\sin(x) \tag{2.25}$$

and the boundary conditions impose

$$f(x) = A\cos(\pi) = 0 \Rightarrow A = 0 \tag{2.26}$$

So we then have (we also of course have $f(x) = 0$ as a solution)

$$f(x) = B\sin(x) \tag{2.27}$$

This should be a vector space as if we add any two functions satisfying the above conditions, they will still be in the same space.

2.7 Question

: We assumed $M(0) = 0$. Does this follow from M being additive, or does it require a separate assumption?

Solution:

This follows from additivity. Consider the zero vector 0 and some other vector v in the space. We require

$$M(v) = M(v + 0) = M(v) + M(0) \tag{2.28}$$

The first equality comes from the definition $v = v + 0$. Thus, if $M(v) = M(v)$, then $M(0) = 0$.

Chapter 3

Iteration and contraction mappings

3.1 Exercise

Let $g_n(x) = (n^2x)^n e^{-n^2x}$. Make a table to show how g_n behaves in $[0,1]$ for $n = 10$. Estimate $\int_0^1 g_{10}(x)\,\mathrm{d}x$. Do the same two things for g_{20}.

Solution:

Here is the code for generating the graphs.

chapter3/gn.py

```
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0,1,1001)

def gn(x,n):
  gn = (n**2*x)**n*np.e**(-n**2*x)
  return gn

y=gn(x,20)

fig = plt.figure()
ax = fig.add_subplot(111)

ax.plot(x,y)
plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
ax.set_ylabel('$g_{20}(x)$',fontsize=30)
#ax.set_ylabel('$g_{10}(x)$',fontsize=30,rotation='horizontal')
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('g20plot.png',bbox_inches='tight')
```

Let's just graph it, this gives a better indication of behavior than a table.

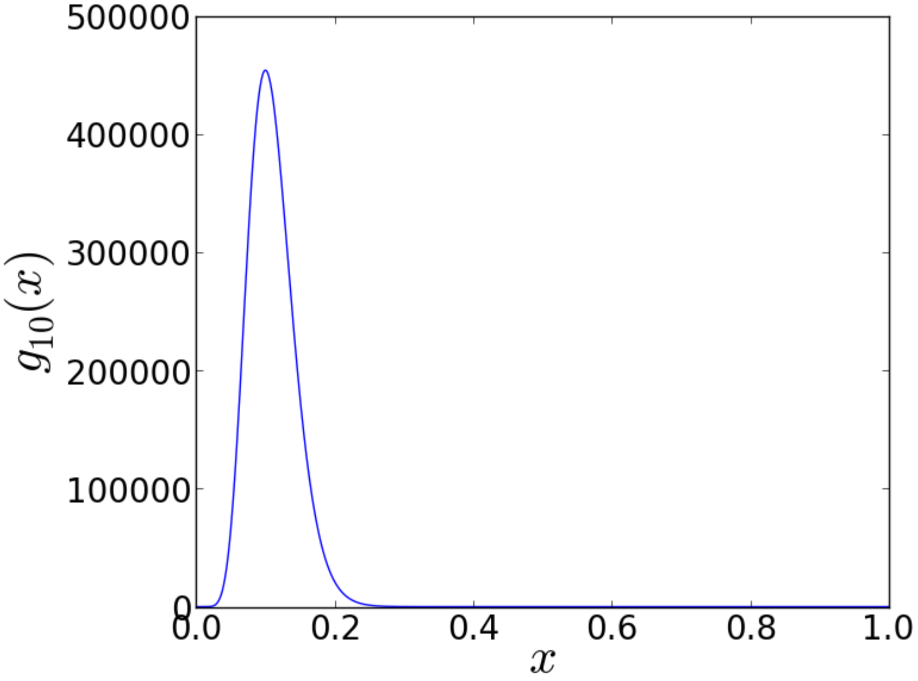

Figure 3.1: A graph of $g_{10}(x)$ on $[0, 1]$.

To estimate the integral, use $y = n^2 x$ so $\mathrm{d}y = n^2\,\mathrm{d}x$ and we find

$$\begin{aligned}\int_0^1 (n^2x)^n e^{-n^2x}\,\mathrm{d}x &= \frac{1}{n^2}\int_0^{n^2} y^n e^{-y}\,\mathrm{d}y = \frac{1}{n^2}\left[-ye^{-y}|_0^{n^2} - n\int_0^{n^2} -y^{n-1}e^{-y}\,\mathrm{d}y\right] \\ &= \frac{1}{n^2}\left[\left\{-y^n e^{-y} - ny^{n-1}e^{-y}\right\} + n(n-1)\int_0^{n^2} y^{n-2}e^{-y}\,\mathrm{d}y\right]\end{aligned} \tag{3.1}$$

We can clearly repeat this process until we get to $n - i = 0$. Thus,

$$\begin{aligned}\int_0^1 (n^2x)^n e^{-n^2x}\,\mathrm{d}x = \frac{1}{n^2}\left[-\sum_{i=0}^{n}\frac{n!}{(n-i)!}y^{n-i}e^{-y}\right]_{y=0}^{n^2} &= -\frac{(n-1)!}{n}\left[e^{-y}\sum_{i=0}^{n}\frac{y^{n-i}}{(n-i)!}\right]_{y=0}^{n^2} \\ = -\frac{1}{n^2}\left[(n)!e^{-y}\sum_{i=0}^{n}\frac{y^i}{i!}\right]_{y=0}^{n^2}&\end{aligned} \tag{3.2}$$

where in the last line I have simply rearranged the sum. We can use the identity

$$\Gamma(s,x) = (s-1)!e^{-x}\sum_{i=0}^{s-1}\frac{x^i}{i!} \tag{3.3}$$

for the incomplete Gamma function and say

$$\int_0^1 (n^2x)^n e^{-n^2x}\,\mathrm{d}x = -\frac{1}{n^2}\left[\Gamma(n+1,y)\right]_{y=0}^{n^2} = \frac{\Gamma(n+1,0) - \Gamma(n+1,n^2)}{n^2} \tag{3.4}$$

Thus, we estimate $\int_0^1 \mathrm{d}x \; g_{10}(x) \approx 36288$.

Similarly, for $g_{20}(x)$ we find

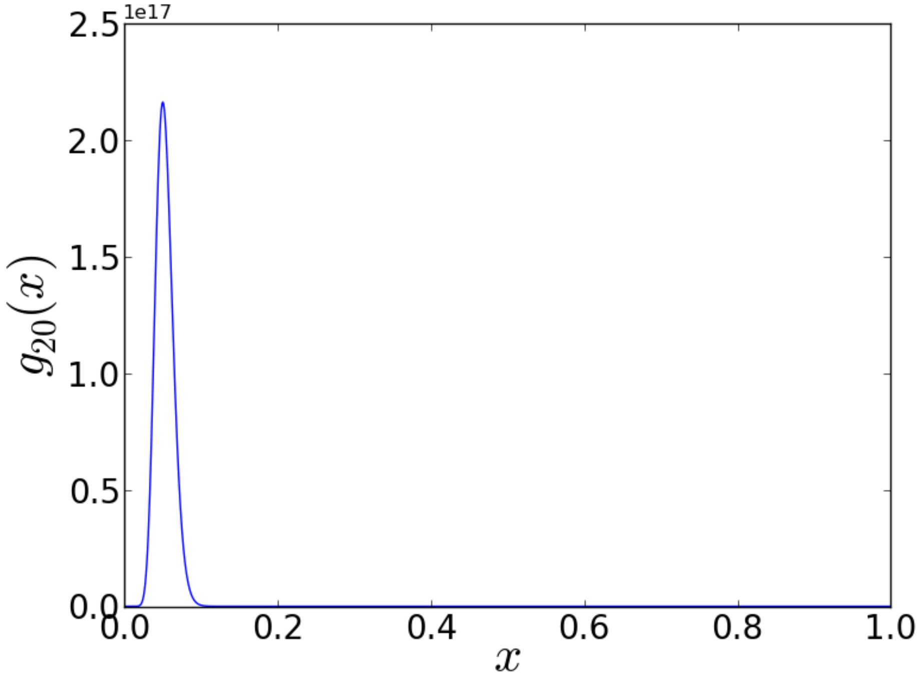

Figure 3.2: A graph of $g_{20}(x)$ on $[0, 1]$.

and $\int_0^1 \mathrm{d}x \; g_{20}(x) \approx 6.08 \times 10^{15}$.

WE can see that as $n \to \infty$ that the integral grows without bound, despite a narrower and narrower region becoming nonzero, so that $g_n(x) \to 0$ as $n \to \infty$.

3.2 Exercise 1

Find the norm of the function f in $\mathcal{C}[0, 1]$ corresponding to each of the following expressions for $f(x)$.

(Note that all the functions given are continuous, so we need only test the extrema and the endpoints to find the largest excursion from 0.

3.2.1 (a)

$3x + 4$

Solution:

There are no extrema, so we need only check $|3(0) + 4| = 4$ and $|3(1) + 4| = 7$. Clearly 7 is larger, so $\|3x + 4\| = 7$.

3.2.2 (b)

$x^2 - x$

Solution:

Here we have an extrema at $x = 1/2$, so

$$|(0)^2 - (-0)| = 0 \tag{3.5}$$
$$|(1)^2 - (1)| = 0 \tag{3.6}$$
$$|(0.5)^2 - (0.5)| = |\frac{-1}{4}| = \frac{1}{4} \tag{3.7}$$

Thus $\|x^2 - x\| = \frac{1}{4}$.

3.2.3 (c)

$5x - 3$

Solution:

There are no extrema, so we need only check $|5(0) - 3| = 3$ and $|5(1) - 3| = 2$. Clearly 3 is larger, so $\|5x - 3\| = 3$.

3.2.4 (d)

$x^2 + x$

Solution:

Now we have an extrema at $x = \frac{-1}{2}$ which is outside of the interval of interest. Clearly, $|1^2 + 1| = 2$ is greater than $|0^2 + 0| = 0$ so $\|x^2 + x\| = 2$

3.2.5 (e)

$\sin(\pi x)$

Solution:

We have $\frac{\mathrm{d}\sin(\pi x}{\mathrm{d}x} = \pi\cos(\pi x) = 0$ implying $x = \frac{1}{2}$ as the only extremum in our interval. $\sin(0) = \sin(\pi) = 0$ whereas $\sin(\pi/2) = 1$ so $\|\sin(\pi x)\| = 1$.

3.2.6 (f)

$-x^2 + x - 0.2$

Solution:

We have an extremum at $-2x+1=0$ or $x=1/2$. Thus

$$\begin{aligned}
|-0^2+0-0.2| &= 0.2 \tag{3.8}\\
|-1^2+1-0.2| &= 0.2 \tag{3.9}\\
|-(0.5)^2+0.5-0.2| &= |-\frac{5}{20}+\frac{10}{20}-\frac{4}{20}|\frac{1}{20} = 0.05 \tag{3.10}
\end{aligned}$$

Thus, $\|-x^2+x-0.2\| = 0.2$.

3.2.7 (g)

$-x^2+x-0.1$

Solution:

This is the same as before, but with 0.1 instead of 0.2, so

$$\begin{aligned}
|-0^2+0-0.1| &= 0.1 \tag{3.11}\\
|-1^2+1-0.1| &= 0.1 \tag{3.12}\\
|-(0.5)^2+0.5-0.1| &= |-\frac{5}{20}+\frac{10}{20}-\frac{2}{20}|\frac{3}{20} = 0.15 \tag{3.13}
\end{aligned}$$

So that we find $\|-x^2+x-0.1\| = 0.15$.

3.2.8 (h)

$(x^2-x)^{10}$

Solution:

Finding the extrema, we see

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 0 = 10(x^2-x)^9(2x-1) = 10x^9(x-1)^9(2x-1) \tag{3.14}$$

Thus, we have extrema at $x=0$, $x=1$, and $x=1/2$. We need to check the first two because they are the boundaries of the interval anyway.

$$\begin{aligned}
|(0^2-0)^{10}| &= 0 \tag{3.15}\\
|(1^2-1)^{10}| &= 0 \tag{3.16}\\
|((0.5)^2-0.5)^{10}| &= \left|\left(\frac{1}{4}-\frac{1}{2}\right)^{10}\right| = \left|\left(\frac{-1}{4}\right)^{10}\right| = \frac{1}{4^{10}} = \frac{1}{1048576} \approx 9.53\times 10^{-7} \tag{3.17}
\end{aligned}$$

Thus, $\|(x^2-x)^{10}\| = 1/4^{10} \approx 9.53\times 10^{-7}$.

3.3 Exercise 2

In $\mathcal{C}[0,1]$ let p denote the function $x \to x$ and q the function $x \to 1-x$. A function is known to belong to $\overline{B}(p, 0.5)$ and $\overline{B}(q, 0.5)$. Sketch the region in which the graph of f must lie. Can any function belong both to $B(p, 0.5)$ and $B(q, 0.5)$?

Solution:

We see that there are no functions that lie in $B(p, 0.5)$ and $B(q, 0.5)$ because at the boundaries the shaded regions just touch each other, so there is no way for a function to be guaranteed not to lie at this boundary which neither open ball contains. The function lies in the below graph where the two shaded regions overlap.

Below is the code for the graph

chapter3/pqshade.py

```python
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0,1,1001)
p = x+0
q = 1-x

pu=x+0.5
pl=x-0.5

qu=1-x+0.5
ql=1-x-0.5

fig = plt.figure()
ax = fig.add_subplot(111)

ax.plot(x,p,'b',label=r'$p=x$')
ax.plot(x,q,'r',label=r'$q=1-x$')
ax.fill_between(x,pl,pu,facecolor='blue',alpha=0.5)
ax.fill_between(x,ql,qu,facecolor='red',alpha=0.5)

plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$g_{20}(x)$',fontsize=30)
#ax.set_ylabel('$g_{10}(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc=2,prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('qpplot.png',bbox_inches='tight')
```

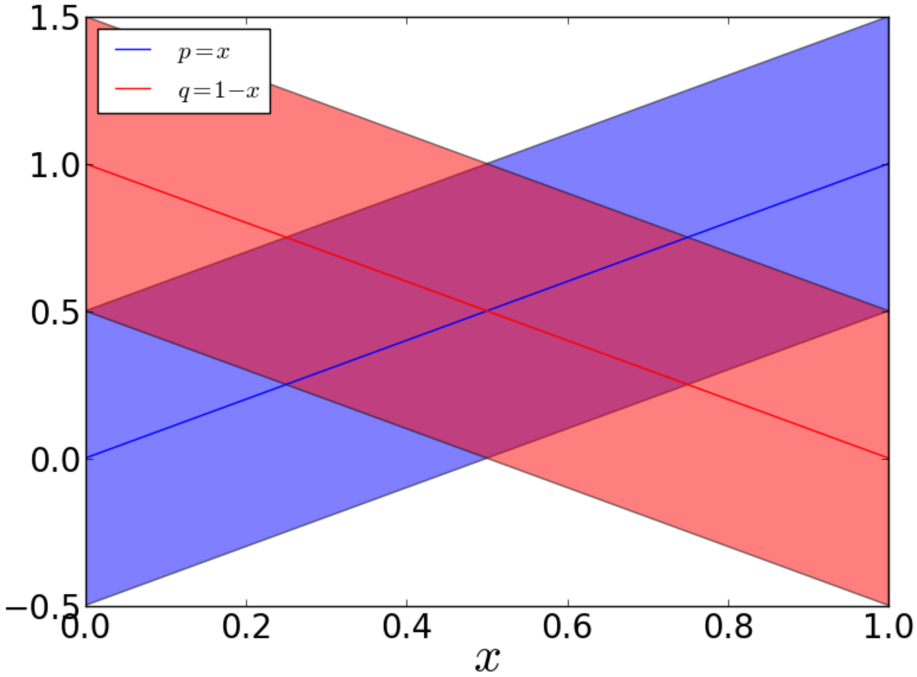

Figure 3.3: A graph of p and q on $[0,1]$.

3.4 Exercise 3

The functions f and g in $\mathcal{C}[0,1]$ are defined by $f(x) = a$ and $g(x) = x^9(1-x)^{11}$. For what value of a is the distance of f from g the least? What is then the value of this distance? Can we reduce this distance by allowing f to be given by $f(x) = a + bx$ and choosing the most suitable values for a and b?

Solution:

Take $h(x) = f(x) - g(x)$. Then $h(x) = a - x^9(1-x)^{11}$ and

$$\frac{\mathrm{d}h(x)}{\mathrm{d}x} = -9x^8(1-x)^{11} - 11x^9(1-x)^{10}(-1) = 0 \tag{3.18}$$
$$11x^9(1-x)^{10} = 9x^8(1-x)^{11} \tag{3.19}$$
$$11x = 9(1-x) \tag{3.20}$$
$$20x = 9 \tag{3.21}$$
$$x = 9/20 \tag{3.22}$$

We note that $g(x) = 0$ at its endpoints, so if we place a halfway between $g(x)$ at its most extreme point and 0, we minimize the distance.

For $x = 9/20$ we see that $g(x) \approx 1.054 \times 10^{-6}$ so if we place $a = g(9/20)/2 \approx 5.27 \times 10^{-7}$ we get the minimum distance, equal to $g(9/20)/2$.

If we allow $f(x) = a + bx$ we can not improve on this, because $g(x)$ is symmetric about $x = 0.5$ on this interval and goes to zero at the endpoints. So it looks very much like a parabola. The best we can do with a linear function is have it tie our straight line across approximation (by subsuming it).

3.5 Final Exercise of Chapter

Let $K(x, y)$ give a continuous function, defined for $0 \leq x \leq b$, $0 \leq y \leq b$, with $|K(x, y)| \leq M$. T is defined by $Tg = h$ where

$$h(x) = \int_0^x K(x, y) g(y) \, \mathrm{d}y \tag{3.23}$$

and $v_n = T^n v_0$ where $v_0 = \mathcal{C}[0, b]$. Show that

$$v_1(x) \leq M \|v_0\| \, x \,, \quad v_2(x) \leq M^2 \|v_0\| \frac{x^2}{2}$$

and generally that

$$v_n(x) \leq M^n \|v_0\| \frac{x^n}{n!}$$

Deduce that, however large b may be, the iteration $g_{n+1} = v_0 + Tg_n$ with $g_0 = 0$ is convergent. Does it make any difference if some other continuous function is chosen for the initial g_0?

Solution:

Let's look at v_n for the first couple.

$$v_1 = \int_0^x \mathrm{d}y \; K(x, y) v_0 \leq \int_0^x \mathrm{d}y \; M \|v_0\| = M \|v_0\| \, x \tag{3.24}$$

$$v_2 = \int_0^x \mathrm{d}y \; K(x, y) v_1 \leq \int_0^x \mathrm{d}y \; K(x, y) \left[M \|v_0\| \, x \right] \leq \int_0^x M^2 \|v_0\| \, x = M^2 \frac{x^2}{2} \tag{3.25}$$

Let's prove this by induction, now. Assume it is true for kth case so that $v_k(x) \leq M^k \|v_0\| \frac{x^k}{k!}$, then

$$v_{k+1} = \int_0^x \mathrm{d}y \; K(x, y) v_k \leq \int_0^x \mathrm{d}y \; M M^k \|v_0\| \frac{x^k}{k!} = M^{k+1} \|v_0\| \frac{x^{k+1}}{(k+1)k!} = M^{k+1} \|v_0\| \frac{x^{k+1}}{(k+1)!} \tag{3.26}$$

and so the $k + 1$ case is true, and we have proved this by induction.

Thus, for $g_{n+1} = v_0 + Tg_n$ with $g_0 = 0$ we see that

$$g_{n+1} = v_0 + Tg_n = v_0 + Tv_0 + T^2 g_{n-1} = v_0 + \sum_{i=1}^{n} T^i v_0 \tag{3.27}$$

$$\|g_{n+1}\| \leq \|v_0\| \left(1 + \sum_{i=1}^{n} M^i \frac{x^i}{i!} \right) \tag{3.28}$$

We note that as $n \to \infty$ that the sum becomes an exponential, and so

$$\lim_{n\to\infty} \|g_{n+1}\| \leq \|v_0\| \, e^{Mx} \tag{3.29}$$

Thus, we see that no matter how large b becomes g_{n+1} will be less than or equal to a constant, which can be quite large, but is bounded.

We can also take

$$\begin{aligned}
\|g_{n+1} - g_{n+1+p}\| = \|\cancel{v_0} + Tg_n - \cancel{v_0} - Tg_{n+p}\| = \left\|\sum_{i=1}^{n} T^i v_0 - \sum_{i=1}^{n+p} T^i v_0\right\| = \left\|\sum_{i=n+1}^{n+p} T^i v_0\right\| \qquad (3.30)\\
\leq \sum_{i=n+1}^{n+p} \left\|T^i v_0\right\| \leq \|v_0\| \sum_{i=n+1}^{n+p} \frac{(Mx)^i}{i!} \leq \|v_0\| \frac{(Mx)^n}{n!} \sum_{i=1}^{p} \frac{(Mx)^i}{i!} \qquad (3.31)\\
\leq \|v_0\| \frac{(Mx)^n}{n!} e^{Mx} \leq \|v_0\| \frac{(Mb)^n}{n!} e^{Mb} \qquad (3.32)
\end{aligned}$$

Now, no matter how large b becomes, because of $n!$ we can get this to be as small as we please because $x^n/n! \to 0$ as $n \to \infty$. Thus this will converge. If $g_0 \neq 0$, then we simply modify the argument, and we get

$$\begin{aligned}
\|g_{n+1} - g_{n+1+p}\| &= \left\|\cancel{v_0} + \sum_{i=1}^{n} T^i v_0 + T^{n+1} g_0 - \cancel{v_0} - \sum_{i=1}^{n} T^i v_0 - T^{n+p+1} g_0\right\| && (3.33)\\
&= \left\|\sum_{i=1}^{n} T^i v_0 - \sum_{i=1}^{n+p} T^i v_0 + T^{n+1} g_0 - T^{n+1+p} g_0\right\| && (3.34)\\
&= \left\|T^{n+1} g_0 - T^{n+1+p} g_0 - \sum_{i=n+1}^{n+p} T^i v_0\right\| && (3.35)\\
&\leq \left\|T^{n+1+p} g_0 - T^{n+1} g_0\right\| + \sum_{i=n+1}^{n+p} \left\|T^i v_0\right\| && (3.36)
\end{aligned}$$

We can note that via the same argument we find $T^n g_0 \leq M \|g_0\| x^n/n!$ so that (with $\kappa(x) = \max\{(Mx)^{n+1}/(n+1)!, (Mx)^{n+p+1}/(n+p+1)!\}$)

$$\begin{aligned}
\|g_{n+1} - g_{n+1+p}\| &\leq \left[\frac{(Mx)^{n+1}}{(n+1)!} + \frac{(Mx)^{n+p+1}}{(n+p+1)!}\right] \|g_0\| + \sum_{i=n+1}^{n+p} \left\|T^i v_0\right\| && (3.37)\\
&\leq 2e^{Mx} \kappa \|g_0\| + \|v_0\| \sum_{i=n+1}^{n+p} \frac{(Mx)^i}{i!} && (3.38)\\
&\leq 2e^{Mx} \kappa \|g_0\| + \|v_0\| \frac{(Mx)^n}{n!} \sum_{i=1}^{p} \frac{(Mx)^i}{i!} && (3.39)\\
&\leq 2e^{Mx} \kappa \|g_0\| + \|v_0\| \frac{(Mx)^n}{n!} e^{Mx} && (3.40)\\
&\leq 2e^{Mb} \kappa(b) \|g_0\| + \|v_0\| \frac{(Mb)^n}{n!} e^{Mb} && (3.41)
\end{aligned}$$

and so both terms will get arbitrarily small because κ acts the same as $(Mb)^n/n!$ in the limit $n \to \infty$. It doesn't affect the fact that we convergence, it just changes how quickly.

Chapter 4

Minkowski Spaces

A chapter explaining the proof that confuses so many.

Chapter 5

Linear operators and their norms

5.1 Exercises on linear functions

In the following situations say whether the function specified is linear or not.

5.1.1 1

Projection $\mathbb{R}^3 \to \mathbb{R}; (x, y, z) \to x$:

Solution:

Take two objects $u = (x_1, y_1, z_1)$ and $v = (x_2, y_2, z_2)$ in the space. Then

$$L(u + v) = L(x_1 + x_2, y_1 + y_2, z_1 + z_2) = x_1 + x_2 = L(u) + L(v) \tag{5.1}$$
$$L(ku) = L(kx_1, ky_1, kz_1) = kx_1 = kL(u) \tag{5.2}$$

and so it is linear.

5.1.2 2

$\mathbb{R} \to \mathbb{R}^3; x \to (x, x, x)$:

Solution:

Take two objects u and v in the space. Then

$$L(u + v) = (u + v, u + v, u + v) = (u, u, u) + (v, v, v) = Lu + Lv \tag{5.3}$$
$$L(ku) = (ku, ku, ku) = k(u, u, u) = kL(u) \tag{5.4}$$

and so it is linear.

5.1.3 3

$\mathbb{R} \to \mathbb{R}; x \to x + 1$:

Solution:

Take two objects u and v in the space. Then

$$L(u+v) = u+v+1 = Lu + Lv - 1 \neq Lu + Lv \tag{5.5}$$

and so it is not linear.

5.1.4 4

Rotation $\mathbb{R}^2 \to \mathbb{R}^2; (x,y) \to (-y,x)$:

Solution:

Take two objects $u = (x_1, y_1)$ and $v = (x_2, y_2)$ in the space. Then

$$L(u+v) = L(x_1+x_2, y_1+y_2) = (-y_1-y_2, x_1+x_2) = (-y_1, x_1) + (-y_2, x_2) = L(u) + L(v) \tag{5.6}$$

$$L(ku) = L(kx_1, ky_1) = (-ky_1, kx_1) = k(-y_1, x_1) = kL(u) \tag{5.7}$$

and so it is linear.

5.1.5 5

$\mathcal{C}[0,1] \to \mathbb{R}; f \to f(0)$:

Solution:

Take two objects $u = f$ and $v = g$ in the space. Then

$$L(u+v) = L(f+g) = (f+g)(0) = f(0) + g(0) = L(u) + L(v) \tag{5.8}$$

$$L(ku) = L(kf) = kf(0) = kL(u) \tag{5.9}$$

and so it is linear.

5.1.6 6

$\mathbb{R}^2 \to \mathbb{R}; (x,y) \to \sqrt{x^2+y^2}$:

Solution:

Take two objects $u = (x_1, y_1)$ and $v = (x_2, y_2)$ in the space. Then

$$L(u+v) = L(x_1+x_2, y_1+y_2) = \sqrt{(x_1+x_2)^2 + (y_1+y_2)^2} \neq \sqrt{x_1^2+y_1^2} + \sqrt{x_2^2+y_2^2} = L(u) + L(v) \tag{5.10}$$

and so it is not linear.

5.1.7 7

$\mathcal{C}[0,1] \to \mathbb{R}; f \to \|f\|$:

Solution:

Take $u = f$ in the space. Then

$$L(ku) = \|kf\| = |k|\,\|f\| = |k|L(u) \neq kL(u) \tag{5.11}$$

and so it is not linear (k can be negative).

5.1.8 8

$\mathbb{R}^3 \to \mathbb{R}; (x, y, z) \to \|(x, y, z)\|_\infty$:

Solution:

For the same reason that 7 doesn't work, this doesn't work. (namely, that norms cannot be negative, whereas k can be).

5.1.9 9

$\mathbb{R}^3 \to \mathbb{R}; (x, y, z) \to \|(x, y, z)\|_1$:

Solution:

For the same reason that 7 doesn't work, this doesn't work.

5.1.10 10

$\mathbb{R}^3 \to \mathbb{R}; (x, y, z) \to x + y + z$:

Solution:

Take two objects $u = (x_1, y_1, z_1)$ and $v = (x_2, y_2, z_2)$ in the space. Then

$$\begin{aligned} L(u+v) &= L(x_1+x_2, y_1+y_2, z_1+z_2) = x_1+x_2+y_1+y_2+z_1+z_2 \\ &= x_1+y_1+z_1+x_2+y_2+z_2 = L(u)+L(v) \end{aligned} \tag{5.12}$$

$$L(ku) = L(kx_1, ky_1, kz_1) = kx_1 + kx_2 + ky_1 = k(x_1+y_1+z_1) = kL(u) \tag{5.13}$$

and so it is linear.

5.1.11 11

$\mathcal{C}[0,1] \to \mathcal{C}[0,1]; f \to g,\, g(x) = [f(x)]^2$:

Solution:

Take two objects $u = f$ and $v = g$ in the space. Then

$$L(u + v) = L(f + g) = (f + g)(0) = f(0) + g(0) = L(u) + L(v) \tag{5.14}$$
$$L(ku) = L(kf) = kf(0) == kL(u) \tag{5.15}$$

and so it is linear.

5.1.12 12

$\mathcal{C}[0, 1] \to \mathcal{C}[0, 1]; f \to g$, $g(x) = f(x^2)$:

Solution:

Take two objects $u = f$ and $v = g$ in the space. Then

$$L(u + v) = L(f + g) = (f + g)(x^2) = f(x^2) + g(x^2) = L(u) + L(v) \tag{5.16}$$
$$L(ku) = L(kf) = kf(x^2) = kL(u) \tag{5.17}$$

and so it is linear.

5.1.13 13

$\mathcal{C}[0, 1] \to \mathbb{R}^6; f \to v$, where v is the vector $[f(0), f(0.2), f(0.4), f(0.6), f(0.8), f(1.0)]$. (This correspondence is involved when we deal with a function specified by a table.:

Solution:

Take two objects f and g in the space (with corresponding vectors u and v). Then

$$L(f + g) = u + v = L(f) + L(g) \tag{5.18}$$
$$L(kf) = ku = kL(u) \tag{5.19}$$

and so it is linear.

5.1.14 14

$\mathcal{C}[0, 1] \to \mathbb{R}; f \to \int_0^1 f(x)\,\mathrm{d}x - 0.5[f(0) + f(1)]$ (This has to with the error when the area under a curve is estimated by the trapezium rule):

Solution:

Take two objects f and g in the space. Then

$$\begin{aligned}L(f+g) &= \int_0^1 [f(x)+g(x)]\,\mathrm{d}x - 0.5[f(0)+g(0)+f(1)+g(1)] \\ &= \int_0^1 f(x)\,\mathrm{d}x - 0.5[f(0)+f(1)] + \int_0^1 g(x)\,\mathrm{d}x - 0.5[g(0)+g(1)] \\ &= L(f)+L(g)\end{aligned} \tag{5.20}$$

$$L(kf) = \int_0^1 kf(x)\,\mathrm{d}x - 0.5[kf(0)+kf(1)] = k\left\{\int_0^1 f(x)\,\mathrm{d}x - 0.5[f(0)+f(1)]\right\} = kL(f) \tag{5.21}$$

and so it is linear.

5.1.15 15

Difference operator. $f \to g$ where $g(x) = f(x+h) - f(x)$:

Solution:

Take two objects f and g in the space. Then

$$L(f+g) = f(x+h)+g(x+h)-f(x)-g(x) = f(x+h)-f(x)+g(x+h)-g(x) = L(f)+L(g) \tag{5.22}$$

$$L(ku) = kf(x+h)-kf(x) = k\left[f(x+h)-f(x)\right] = kL(u) \tag{5.23}$$

and so is linear.

5.2 Exercise in Matrix Norm

In the plane $S\,(O,1)$ is a circle when the ℓ_2 metric is used, a square with horizontal and vertical sides when the ℓ_∞ metric is used, and a tilted square when the ℓ_1 metric is used. Find and draw the figures to which these 'spheres' are mapped when the matrix $A = \begin{bmatrix}1 & 1\\ 0 & 2\end{bmatrix}$ acts on them. For each space, draw the appropriate $S(O,r)$ which is just large enough to contain all the output points. Deduce the value of $\|A\|$ for each of the three cases.

Solution:

We are being asked how x and y are changed for each of these metrics. Note

$$\begin{bmatrix}1 & 1\\ 0 & 2\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}x+y\\ 2y\end{bmatrix} \tag{5.24}$$

For ℓ_2 we had

$$x^2+y^2 = 1 \tag{5.25}$$

$$\tag{5.26}$$

as our defining equation. Thus, for our list of points, given an (x, y) have it go to $(x + y, 2y)$ so that we get an elongated ellipse instead of a circle.

To find the maximum distance this gets from the origin, we use

$$d = \sqrt{(x+y)^2 + (2y)^2} = \sqrt{(x + \sqrt{1-x^2})^2 + 4(1-x^2)} = \sqrt{x^2 + 2x\sqrt{1-x^2} + 5(1-x^2)} \tag{5.27}$$

$$\frac{\mathrm{d}d}{\mathrm{d}x} = \frac{2x + 2\sqrt{1-x^2} + \frac{(-2x)x}{\sqrt{1-x^2}} - 10x}{2d} = 0 \tag{5.28}$$

Solving for x we find:

$$-8x + 2\sqrt{1-x^2} - \frac{2x^2}{\sqrt{1-x^2}} = 0 \tag{5.29}$$

$$\sqrt{1-x^2} - \frac{x^2}{\sqrt{1-x^2}} = 4x \tag{5.30}$$

$$(1-x^2) - 2x^2 + \frac{x^4}{1-x^2} = 16x^2 \tag{5.31}$$

$$1 + \frac{x^4}{1-x^2} = 19x^2 \tag{5.32}$$

$$(1-x^2) + x^4 = 19x^2 - 19x^4 \tag{5.33}$$

$$20x^4 - 20x^2 + 1 = 0 \tag{5.34}$$

$$x^4 - x^2 + \frac{1}{20} = 0 \tag{5.35}$$

$$x^2 = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{1}{20}} = \frac{1}{2} \pm \sqrt{\frac{1}{5}} \tag{5.36}$$

$$x = \pm\sqrt{\frac{1}{2} \pm \sqrt{\frac{1}{5}}} \tag{5.37}$$

If we look back at our original equation, we know that only two of them can actually satisfy the equations, and so we find the x's that are good are

$$x = -\sqrt{\frac{1}{2} + \sqrt{\frac{1}{5}}}, \sqrt{\frac{1}{2} - \frac{1}{\sqrt{5}}} \tag{5.38}$$

Let's use the latter value for a test, as both should give equivalent answers, then

$$x \equiv \xi = \sqrt{\frac{1}{2} - \frac{1}{\sqrt{5}}} \approx 0.2297 \tag{5.39}$$

$$y \equiv \psi = \sqrt{1 - \frac{1}{2} + \frac{1}{\sqrt{5}}} = \sqrt{\frac{1}{2} + \frac{1}{\sqrt{5}}} \approx 0.973 \tag{5.40}$$

Thus, the maximum distance will be

$$\begin{aligned} d = \sqrt{(\xi+\psi)^2+4\psi^2} &= \sqrt{\frac{1}{2} - \cancel{\frac{1}{\sqrt{5}}} + \frac{1}{2} + \cancel{\frac{1}{\sqrt{5}}} + 2\sqrt{\frac{1}{4}-\frac{1}{5}} + 2 + \frac{4}{\sqrt{5}}} \\ &= \sqrt{3 + 2\sqrt{\frac{1}{20}} + \frac{4}{\sqrt{5}}} = \sqrt{3 + \frac{1+4}{\sqrt{5}}} \\ &= \sqrt{3+\sqrt{5}} \approx 2.28825 \end{aligned} \tag{5.41}$$

Thus the limiting sphere is $S(O, 3+\sqrt{5}) \approx S(O, 2.288)$.

For ℓ_∞ the square becomes a parallelogram with $x \to x+y$ and $y \to 2y$. In this case, we can use that the maximum distance we ever get to is $y = 2$, so $S(O, 2)$ is the limiting sphere.

For ℓ_1 we get a stranger looking parallelogram (at an angle). Here, the maximum distance will clearly be for $x = 0$, $y = 1$ as we then get a distance $|0+1|+|2| = 3$. It is possible to prove this algebraically, because the maximum for $|x+y|+|2y|$ over the range $-1 \leq x, y \leq 1$ by looking at the boundaries and extrema, but it rather clear visually. Thus the limiting sphere is $S(O, 3)$, simply enough.

We can see all of these in Figure 5.1.

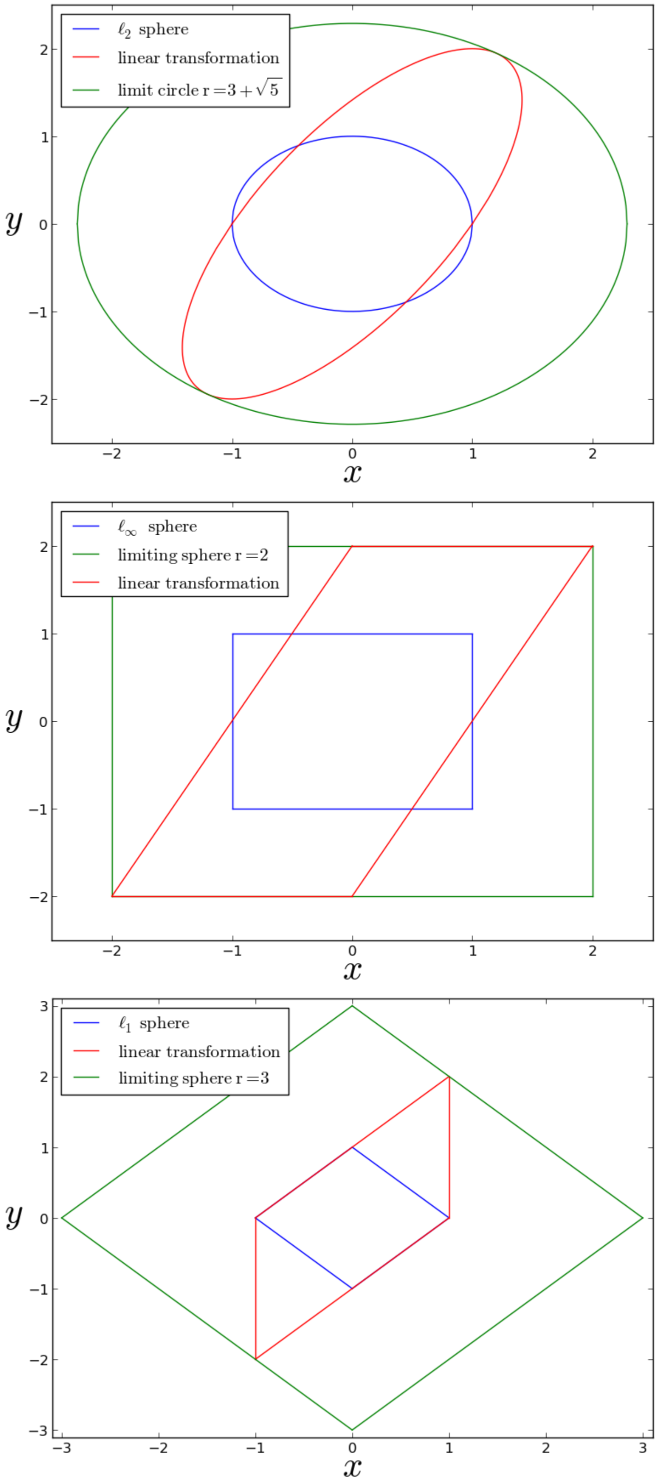

Figure 5.1: This shows the original sphere, the transformation of the sphere, and the new limiting sphere.

with generating code

chapter4/drawshapes.py

```
#!/usr/bin/env python2

```

```
import numpy as np
import matplotlib.pyplot as plt

x1=np.linspace(-1,1,201)
y1=np.sqrt(1-x1**2)
y2=-np.sqrt(1-x1**2)

xr1=x1+y1
yr1=2*y1

xr2=x1+y2
yr2=2*y2

fnum=3+np.sqrt(5)
limx = np.linspace(-np.sqrt(fnum),np.sqrt(fnum),501)
limy1= np.sqrt(fnum-limx**2)
limy2= -np.sqrt(fnum-limx**2)


fig=plt.figure()
ax=fig.add_subplot(111)

ax.plot(x1,y1,'b',label=r'$\ell_2$'+'␣'+r'$\rm{sphere}$')
ax.plot(x1,y2,'b')

ax.plot(xr1,yr1,'r',label=r'$\rm{linear\␣transformation}$')
ax.plot(xr2,yr2,'r')

ax.plot(limx,limy1,'g',label=r'$\rm{limit\␣circle\␣}r=3+\sqrt{5}$')
ax.plot(limx,limy2,'g')

ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y$',fontsize=30)
ax.set_ylabel('$y$',fontsize=30,rotation='horizontal')
#plt.title(r'Real$(n)$')
ax.set_xlim([-2.5,2.5])
ax.set_ylim([-2.5,2.5])

ax.legend(loc='best',prop={'size':15})

plt.tight_layout()
plt.savefig('ell2A.png',bbox_inches='tight')

plt.clf()

xa=np.linspace(-1,1,201)
ya=np.linspace(-1,1,201)
ones=np.ones(201)

fig=plt.figure()
ax=fig.add_subplot(111)

ax.plot(xa,ones,'b',label=r'$\ell_\infty$'+'␣'+r'$\rm{sphere}$')
ax.plot(xa,-ones,'b')
ax.plot(ones,ya,'b')
ax.plot(-ones,ya,'b')

ax.plot(2*xa,2*ones,'g',label=r'$\rm{limiting\␣sphere\␣}r=2$')
ax.plot(2*xa,-2*ones,'g')
ax.plot(2*ones,2*ya,'g')
ax.plot(-2*ones,2*ya,'g')

ax.plot(xa+ones,2*ones,'r',label=r'$\rm{linear\␣transformation}$')
ax.plot(xa-ones,-2*ones,'r')
ax.plot(ones+ya,2*ya,'r')
ax.plot(-ones+ya,2*ya,'r')


ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y$',fontsize=30)
ax.set_ylabel('$y$',fontsize=30,rotation='horizontal')
```

```
#plt.title(r'Real$(n)$')
ax.set_xlim([-2.5,2.5])
ax.set_ylim([-2.5,2.5])

ax.legend(loc='best',prop={'size':15})

plt.tight_layout()
plt.savefig('ellinfA.png',bbox_inches='tight')

plt.clf()

xr=np.linspace(0,1,5)
xl=np.linspace(-1,0,5)
yr=1-xr
yl=-xr

fig=plt.figure()
ax=fig.add_subplot(111)

ax.plot(xr,yr,'b',label=r'$\ell_1$'+'␣'+r'$\rm{sphere}$')
ax.plot(xr,xl,'b')
ax.plot(xl,xr,'b')
ax.plot(xl,yl,'b')

ax.plot(xr+yr,2*yr,'r',label=r'$\rm{linear\␣transformation}$')
ax.plot(xr+xl,2*xl,'r')
ax.plot(xl+xr,2*xr,'r')
ax.plot(xl+yl,2*yl,'r')

ax.plot(3*xr,3*yr,'g',label=r'$\rm{limiting\␣sphere\␣}r=3$')
ax.plot(3*xr,3*xl,'g')
ax.plot(3*xl,3*xr,'g')
ax.plot(3*xl,3*yl,'g')

ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y$',fontsize=30)
ax.set_ylabel('$y$',fontsize=30,rotation='horizontal')
#plt.title(r'Real$(n)$')
ax.set_xlim([-3.1,3.1])
ax.set_ylim([-3.1,3.1])

ax.legend(loc='best',prop={'size':15})

plt.tight_layout()
plt.savefig('ell1A.png',bbox_inches='tight')
```

5.3 Exercises on operator norms

5.3.1 1

Let $w = 2x - 3y + 4z$. Which point of the form $(\pm 1, \pm 1, \pm 1)$ make w largest? If f is the function $\mathbb{R}^3 \to \mathbb{R}$, $(x, y, z) \to w$, what is $\|f\|_\infty$?

Solution:

Clearly to maximize w we need all of (x, y, z) to be such that x is as large and z as large as they can be while y be as negative as possible.

In this case $(1, -1, 1)$ will give the largest possible value of $w = 9$.

Let's choose a unit vector in the direction of largest increase, which will be $(0, 0, 1)$ and so $\|f\|_\infty = 4$.

5.3.2 2

Generalize from your answer to question 1. What is $\|f\|_\infty$ for $f : (x, y, z) \to ax + by + cz$?

Solution:

We simply take the sum of $|a|, |b|, |c|$, since a unit vector in that direction will be increased the most, and in the ∞-norm, we simply take the max of the components which will be the sum of the absolute values. Thus, $\|f\|_\infty = |a| + |b| + |c|$.

5.3.3 3

What is $\|f\|_\infty$ for $f : (x_1, x_2, \ldots, x_n) \to \sum_{r=1}^{n} a_r x_r$?

Solution:

Via the same reasoning as above, for a max norm such as the ∞-norm, we have $\|f\|_\infty = \sum_{r=1}^{n} |a_r|$.

5.3.4 4

Let

$$\begin{aligned} w_1 &= 2x_1 - 3x_2 + 4x_3 \\ w_2 &= x_1 + x_2 + x_3 \end{aligned}$$

Consider $f : x \to w$, where $x = (x_1, x_2, x_3)$ and $w = (w_1, w_2)$. What x with $\|x\|_\infty = 1$ makes $\|w\|_\infty$ a maximum? What is $\|f\|$, it being understood the ℓ_∞ norm applies both to input and output?

Solution:

Just looking at the possibilities, we want to maximize w_1 since the maximum of w_1 is larger than the possible maximum of w_2 with $(x_1, x_2, x_3) = (\pm 1, \pm 1, \pm 1)$ being the unit vectors in the ∞-norm that are most useful. Then

$$x = (1, -1, 1) \tag{5.42}$$

$$w = (9, 1) \tag{5.43}$$

is the maximum since 9 is the largest that either w_1 or w_2 could ever be. Thus $\|f\| = 9$.

5.3.5 5

Would the answer to question 4 be different if, the rest of the equation being unchanged, w_2 were altered (a) to $x_1 + 2x_2 - 4x_3$ or (b) to $x_1 + 3x_2 - 6x_3$?

Solution:

Yes, then for (a) $x = (1, 1, -1)$ leading to $\|f\| = 7$ and (b) $x = (1, 1, -1)$ leading to $\|f\| = 10$.

5.3.6 6

Investigate the generalization of the problems posed in questions 4 and 5, with the aim of finding a formula for $\|f\|$ where $f : x \to w$ is specified by $w_r = \sum_s a_{rs}x_s$ with $1 \le r \le m$ and $1 \le s \le n$.

Solution:

It will clearly just be the maximum over the w_r. Thus

$$\|f\| = \max_r \sum_{s=1}^{n} |a_{rs}| \tag{5.44}$$

which is the max of the absolute value row sum.

5.3.7 7

Let $w = 2x - 3y + 4z$. What is the maximum value $|w|$ can have have subject to $|x| + |y| + |z| = 1$? What is $\|f\|$ for $f : (x, y, z) \to w$ if the ℓ_1 norm is used for the input?

Solution:

These two questions are equivalent, we may note. The simplest way to figure this out would be to graph it, but three dimensions is challenging, and we'd have to deal with a cube. Thus, we can instead use that one of the edge points must be a maximum because there are no extrema for w except at boundaries. Thus we would need to check $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$ and $(0, 0, \pm 1)$. Therefore we see that $(0, 0, 1)$ produces the largest value and so $\|f\|_1 = 4$.

5.3.8 8

Generalize from your answer to question 7. What is $\|f\|$ for $f(x, y, z) \to ax + by + cz$, the ℓ_1 norm applying to the input?

Solution:

We will have the largest value be $\|f\| = \max(|a|, |b|, |c|)$ via the same reasoning as in the previous question.

5.3.9 9

What is $\|f\|$ for $(x_1, x_2, \ldots, x_n) \to \sum_{r=1}^{n} a_r x_r$ the ℓ_1 norm applying to the input?

Solution:

Via the same reasoning we maximize over the arguments, thus $\|f\| = \max_r |a_r|$.

5.3.10 10

Let

$$\begin{aligned} w_1 &= 7x_1 + 2x_2 \\ w_2 &= -3x_1 + 6x_2 \end{aligned}$$

What is the maximum value of $|w_1| + |w_2|$ subject to $|x_1| + |x_2| = 1$? For $f : x \to w$, where $x = (x_1, x_2)$ and $w = (w_1, w_2)$ what x with $\|x\|_1 = 1$ makes $\|w\|_1$ a maximum? With these norms for x and w, what is $\|f\|$?

Solution:

This is a bit more complicated because we now need to think about maximizing w_1 and w_2 simultaneously. However, we are again aided by the fact that we need only check boundaries as there are no extrema in w_1 or w_2. Thus, we see that we need only check w_1 and w_2 for $(\pm 1, 0)$ and $(0, \pm 1)$. This will clearly just be summing the "columns" so we see $7 + 3 = 10$ is greater than $2 + 6 = 8$. Thus, the maximum $\|x\|_1$ is $x = (\pm 1, 0)$ with $\|w\|_1 = 10$, so that $\|f\| = 10$.

5.3.11 11

How would the answer to question 10 have to be modified if, th rest of the equation remaining unaltered, the equation for w_2 was changed to (a) $w_2 = -3x_1 + 9x_2$, (b) $w_2 = -3x_1 + 8x_2$, (c) $w_2 = 3x_1 + 8x_2$

Solution:

Now we have (a) $x = (0, \pm 1)$, $\|f\| = 11$ (b) here there is no "unique" vector as $x = (0, \pm 1)$ or $x = (\pm 1, 0)$ yields $\|f\| = 10$ and (c) is the exact same as (b) as the negative sign makes no difference because of absolute values.

5.3.12 12

Investigate the general question of a formula for $\|f\|$ with $f : \ell_1 \to \ell_1$, $x \to w$ where $w_r = \sum_s a_{rs} x_s$ with $1 \le r \le m$ and $1 \le s \le n$.

Solution:

Clearly, our reasoning implies that

$$\|f\| = \max_s \sum_{r=1}^{m} |a_{rs}| \tag{5.45}$$

which is the max of the absolute value column sum.

5.3.13 13

If

$$c = \int_0^1 (3x^2 + 2x + 1) f(x)\,\mathrm{d}x\,,$$

what is the maximum value c can take for a continuous function f subject to $|f(x)| \leq 1$ for $0 \leq x \leq 1$?

Solution:

Clearly, choose $f(x) = 1$ throughout the interval so that the rest of the integral (which is completely positive) can be maximized. Thus

$$c = \int_0^1 \mathrm{d}x\ (3x^2 + 2x + 1) = 1 + 1 + 1 = 3 \tag{5.46}$$

is the maximum.

5.3.14 14

What function, *not necessarily continuous*, makes c maximum where

$$c = \int_0^{2\pi} f(x)\sin(x)\,\mathrm{d}x \qquad |f(x)| \leq 1$$

for $x \in [0, 2\pi]$? If we also require f to be continuous what is the supremum for c? Is there any continuous f that makes c actually equal to the supremum value?

Solution:

The function is simply $\mathrm{sgn}(\sin(x))$ which is discontinuous. The supremum for c will be 4 because

$$\int_0^{\pi} \sin(x)\,\mathrm{d}x = -\cos(\pi) + \cos(0) = 2 \tag{5.47}$$

and sin is antisymmetric around π on this interval. There is no continuous f that will actually hit the supremum value, because then it would be $\mathrm{sgn}(\sin(x))$, but we can get arbitrarily close.

5.3.15 15

Discuss the general question of $\|T\|$, where T is in $\mathcal{C}[a, b]$ and T maps $f \to c$ where

$$c = \int_a^b \phi(x) f(x)\,\mathrm{d}x$$

for a given continuous function ϕ. Question 13 throws some light on the situation when $\phi(x)$ is positive throughout, and question 14 on the case when $\phi(x)$ changes sing within the interval.

Solution:

We see that $\|T\|$ is $\int_a^b |\phi(x)|$ if $|f(x)| \leq 1$ is a restriction.

5.3.16 16

Let $T : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$ $f \to g$ be defined by

$$g(x) = \int_0^1 (x^2 + 2xy + 3y^2) f(y)\,\mathrm{d}y\,.$$

What f with $\|f\| = 1$ makes $\|g\|$ maximum? What is $\|T\|$?

Solution:

Given the domain of $g(x)$ as $[0,1]$, then we see we wish to make the integral as large as possible since it is completely positive. Thus if $f(y) = 1$, then

$$g(x) = x^2 + x + 1 \tag{5.48}$$

and so $\|g\| = 3$. WE also then see that $\|T\| = 3$, since this is the maximum possible for $\|g\| = Tf$.

5.3.17 17

(i) Let k be a prescribed number, for which $0 \leq k \leq 1$. Find the supremum of $|c|$ where

$$c = \int_0^1 (k - y) f(y)\,\mathrm{d}y$$

and $f \in \mathcal{C}[0,1]$ with $\|f\| = 1$. What values of k in the interval $[0,1]$ make $\sup |c|$ a maximum, and what is the value of this maximum?

(ii) Let $g(x) = \int_0^1 (x - y) f(y)\,\mathrm{d}y$. What is $\|T\|$ for $T : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$, $f \to g$?

Solution:

(i) The supremum will be where we take $f(y) = \operatorname{sgn}(k - y)$. This will then give

$$|c| = |\int_0^k (k-y)\,\mathrm{d}y + \int_k^1 (y-k)\,\mathrm{d}y| = |k^2 - \frac{k^2}{2} + \frac{1-k^2}{2} - k(1-k)| = |\frac{1}{2} - k + k^2| \tag{5.49}$$

we need to check $k = 0, k = 1, k = 1/2$ for extrema.

$$\sup |c| \overset{k=0}{=} \frac{1}{2} \tag{5.50}$$

$$\sup |c| \overset{k=1}{=} \frac{1}{2} - 1 + 1 = \frac{1}{2} \tag{5.51}$$

$$\sup |c| \overset{k=1/2}{=} \frac{1}{2} - 1/2 + 1/4 = 0 \tag{5.52}$$

And so the $\max \sup |c| = \frac{1}{2}$ on this interval.

(ii) This is given by our answer above since the max possible is given by either $x = 0$ or $x = 1$ we get $\|T\| = \frac{1}{2}$.

5.3.18 18

Investigate $\|T\|$ for $T : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$, $f \to g$ where

$$g(x) = \int_0^1 K(x,y) f(y) \, \mathrm{d}y$$

the kernel $K(x,y)$ being continuous. Note: If $K(x,y)$ were given by a table, for example by its values when x and y are both multiples of 0.1, an approximate treatment would lead us to consider equations of the type $g_r = \sum_s k_{rs} f_s$. Thus it is likely that there will be some resemblance between the answer to this question and the answer to question 6 above.

Solution:

Given the information from the above, we expect the answer to be of a form similar to

$$\|T\| \to \max_x |K(x,y)| \tag{5.53}$$

and indeed, if $|f(y)| \leq 1$ we find the supremum by assuming that we integrate $\int_0^1 |K(x,y)| \, \mathrm{d}y$. We will get some function from this $g(x)$ and we now need to choose the x to get the supremum.

Thus,

$$\|T\| = \sup_x \int_0^1 |K(x,y)| \, \mathrm{d}y \tag{5.54}$$

5.4 Bounded, Linear Operators

5.4.1 Exercise 1

Let $T : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$ be defined by $Tf = g$ where $g(x) = \int_0^x t f(t) \, \mathrm{d}t$. Find T and discuss the convergence of the series $\mathbb{1} + T + T^2 + \cdots + T^n + \cdots$ in the space of bounded operators. Examine the series produced by the iteration

$$f_{n+1}(x) = 1 + \int_0^x t f_n(t) \, \mathrm{d}t \quad \text{with } f_0(x) \equiv 0$$

Identify the analytic function given by this series. Is it a solution of the integral equation $f(x) = 1 + \int_0^x t f(t) \, \mathrm{d}t$?

Solution:

T is defined by the above equation, so I'm not sure what is meant by find T. Note

$$T^2 f = \int_0^x \mathrm{d}t \; t \int_0^t \mathrm{d}t' \; t' f(t') \tag{5.55}$$

And generally speaking,

$$T^n = \int_0^x \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2 \; t_2 \cdots \int_0^{t_{n-1}} \mathrm{d}t_n t_n \tag{5.56}$$

We also can see that if $h = \|f\|$, then

$$\int_0^x tf(t)\,\mathrm{d}t < \int_0^x \mathrm{d}t\ th = \frac{x^2}{2}h < h = \|f\| \tag{5.57}$$

Thus, $\|Tf\| < \|f\|$ and so $\|T\| < 1$. Note that this strict inequality is true of $\left\|T^k f\right\| < \left\|T^{k-1} f\right\| < \cdots < \|f\|$. Hence (with $T^0 \equiv \mathbb{1}$) and $\|T\| = a$ we have (using $\left\|T^k\right\| \leq \|T\|^k$)

$$\left\|\sum_{k=0}^{\infty} T^k\right\| < \sum_{k=0}^{\infty} \left\|T^k\right\| < \sum_{k=0}^{\infty} a^k = \frac{1}{1-a} \tag{5.58}$$

Thus this must converge.

For $f_0(x) = 0$ we see

$$f_1 = 1 \tag{5.59}$$

$$f_2 = 1 + \int_0^x \mathrm{d}t\ t = 1 + \frac{x^2}{2} \tag{5.60}$$

$$f_3 = 1 + \int_0^x \mathrm{d}t\ t(1 + \frac{t^2}{2}) = 1 + \frac{x^2}{2} + \frac{x^4}{4(2)} \tag{5.61}$$

$$f_4 = 1 + \int_0^x \mathrm{d}t\ t(1 + t^2/2 + t^4/8) = 1 + \frac{x^2}{2} + \frac{x^4}{8} + \frac{x^6}{48} \tag{5.62}$$

$$f_n = 1 + \int_0^x \mathrm{d}t\ tf_{n-1}(t) = \sum_{j=1}^{n} \frac{t^{2j-2}}{(2j-2)!!} = \sum_{k=0}^{n-1} \frac{t^{2k}}{(2k)!!} \tag{5.63}$$

where $(2k)!! = (2k)(2k-2)(2k-4)\cdots(2)$. Note that

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \tag{5.64}$$

$$e^{t^2} = \sum_{i=0}^{\infty} \frac{(t^2)^i}{i!} = \sum_{i=0}^{\infty} \frac{t^{2i}}{i!} \tag{5.65}$$

$$e^{t^2/2} = \sum_{i=0}^{\infty} \frac{(t^2/2)^i}{i!} = \sum_{i=0}^{\infty} \frac{t^{2i}}{2^i(i!)} = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)[2[i+1])\cdots 2} = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!!} \tag{5.66}$$

Thus, the series yields $f(t) = e^{t^2/2}$.

Then (using $\frac{\mathrm{d}}{\mathrm{d}t}e^{t^2/2} = te^{-t^2/2}$)

$$1 + \int_0^x \mathrm{d}t\ te^{t^2/2} = 1 + \int_0^x \mathrm{d}t\ \frac{\mathrm{d}}{\mathrm{d}t}e^{t^2/2} = 1 + e^{-x^2/2} - 1 = e^{-x^2/2} = f(x) \tag{5.67}$$

and so it does satisfy the integral equation.

5.4.2 2

Let matrix

$$M = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}$$

Find the matrix N given by the infinite series $N = \mathbb{1} + M + M^2 + \cdots + M^n + \cdots$, when this series converges. What condition must a and b satisfy if the series that appear in the four entries of N are all to converge? Find algebraic expressions for the entries of N when this condition is satisfied. On the analogy of results in elementary algebra, it appears plausible that $(\mathbb{1} - M)N = \mathbb{1}$. Is this equation in fact verified? What are the eigenvalues of M?

Solution:

First let's find M^n for a couple of M.

$$M^2 = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} = \begin{bmatrix} ab & 0 \\ 0 & ab \end{bmatrix} = ab\mathbb{1} \tag{5.68}$$

$$M^3 = \begin{bmatrix} ab & 0 \\ 0 & ab \end{bmatrix} \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} = ab \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \tag{5.69}$$

$$M^{2n} = (ab)^n \mathbb{1} \tag{5.70}$$

$$M^{2n+1} = (ab)^n \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \tag{5.71}$$

Thus, we find

$$N = \sum_{k=0}^{\infty} M^k = \left(\sum_{k=0}^{\infty} (ab)^k \right) \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} \tag{5.72}$$

If this is to converge we require $ab < 1$ When this expression is satisfied, we find

$$N = \sum_{k=0}^{\infty} M^k = \frac{1}{1-ab} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{1-ab} & \frac{a}{1-ab} \\ \frac{b}{1-ab} & \frac{1}{1-ab} \end{bmatrix} \tag{5.73}$$

Note

$$\det M = -ab \tag{5.74}$$

$$\det N = \frac{1}{(1-ab)^2} - \frac{ab}{(1-ab)^2} = \frac{1-ab}{1-ab^2} = \frac{1}{1-ab} \tag{5.75}$$

We find

$$MN = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{1-ab} & \frac{a}{1-ab} \\ \frac{b}{1-ab} & \frac{1}{1-ab} \end{bmatrix} = \begin{bmatrix} \frac{ab}{1-ab} & \frac{a}{1-ab} \\ \frac{b}{1-ab} & \frac{ab}{1-ab} \end{bmatrix} \tag{5.76}$$

Thus,

$$(\mathbb{1} - M)N = \begin{bmatrix} \frac{1}{1-ab} & \frac{a}{1-ab} \\ \frac{b}{1-ab} & \frac{1}{1-ab} \end{bmatrix} - \begin{bmatrix} \frac{ab}{1-ab} & \frac{a}{1-ab} \\ \frac{b}{1-ab} & \frac{ab}{1-ab} \end{bmatrix} = \begin{bmatrix} \frac{1-ab}{1-ab} & 0 \\ 0 & \frac{1-ab}{1-ab} \end{bmatrix} = \mathbb{1} \tag{5.77}$$

is confirmed. Note that the eigenvalues of M are given by $\lambda^2 - ab = 0$ or $\lambda = \pm\sqrt{ab}$.

5.4.3 3

Let $M : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$ be defined by $Mf = g$ where

$$g(x) = \int_0^x yf(y)\,\mathrm{d}y + \int_x^1 xf(y)\,\mathrm{d}y$$

Find $\|M\|$. Will the iteration defined by $f_{n+1} = \mathbb{1} + Mf_n$ converge? Here $\mathbb{1}$ denotes the function with the constant value 1, and we suppose $f_0(x) \equiv 0$.

Solution:

Given $\|f\| = 1$ choosing $f = 1$ will maximize this because both integrals have positive arguments. Thus

$$g(x) = \int_0^x y\,\mathrm{d}y + \int_x^1 x\,\mathrm{d}y = \frac{x^2}{2} + x(1-x) = x - \frac{x^2}{2} = x\left(1 - \frac{x}{2}\right) \tag{5.78}$$

To maximize this, we see we need to check the endpoints and

$$\frac{\mathrm{d}g}{\mathrm{d}x} = 0 = 1 - x \tag{5.79}$$

or $x = 1$, which is an endpoint. Clearly this is maximum at $x = 1$ yielding

$$\|g(x)\| \overset{\|f\|=1}{=} \frac{1}{2} \tag{5.80}$$

and thus $\|M\| = \frac{1}{2}$. Given this, we do expect $f_{n+1} = \mathbb{1} + Mf_n$ to converge. In fact we can show this because $\|M\| < 1$.

As an example, we find

$$f_0 = 0 \tag{5.81}$$

$$f_1 = 1 \tag{5.82}$$

$$f_2 = 1 + x(1 - \frac{x}{2}) \tag{5.83}$$

$$f_3 = \frac{x^4}{24} - \frac{x^3}{6} - \frac{x^2}{2} + \frac{4x}{3} + 1 \tag{5.84}$$

$$f_4 = -\frac{x^6}{720} + \frac{x^5}{120} + \frac{x^4}{24} - \frac{2x^3}{9} - \frac{x^2}{2} + \frac{22x}{15} + 1 \tag{5.85}$$

$$f_5 = \frac{x^8}{40320} - \frac{x^7}{5040} - \frac{x^6}{720} + \frac{x^5}{90} + \frac{x^4}{24} - \frac{11x^3}{45} - \frac{x^2}{2} + \frac{479x}{315} + 1 \tag{5.86}$$

Note that $\frac{22}{15} \approx 1.4667$ and $\frac{479}{315} \approx 1.52063$.

It is not at all obvious what this is becoming. But we can use Leibnitz's rule to find

$$f(x) = 1 + \int_0^x yf(y)\,\mathrm{d}y + \int_x^1 xf(y)\,\mathrm{d}y \tag{5.87}$$

$$f'(x) = \cancel{xf(x)} + \int_x^1 f(y)\,\mathrm{d}y - \cancel{xf(x)} = \int_x^1 f(y)\,\mathrm{d}y \tag{5.88}$$

$$f''(x) = -f(x) \tag{5.89}$$

Thus, our most general solution is

$$f(x) = A\cos(x) + B\sin(x) \tag{5.90}$$

Using $f(0) = 1$ we see that $A = 1$. Using $f'(1) = 0$ then

$$\begin{align} f'(1) = -\sin(1) + B\cos(1) = 0 \tag{5.91} \\ B = \tan(1) \tag{5.92} \end{align}$$

and so the solution is

$$f(x) = \cos(x) + \tan(1)\sin(x) \tag{5.93}$$

We note $\tan(1) \approx 1.55741$, and that our power series is getting close to this value with the fractions employed.

Chapter 6

Differentiation and Integration

6.1 Iteration Exercises

6.1.1 1

A well-known algorithm for finding the square root of a number a uses the iteration $x \to f(x)$ with $f(x) = \frac{1}{2}[x + a/x]$. In what region is the convergence of the iteration guaranteed by the condition $f'(x) \leq k < 1$? Does the iteration converge for any initial value, x_0, outside this region? Are there any circumstances in which the iteration could converge to the other square root, $-\sqrt{a}$? What would happen if someone tried to use this algorithm to calculate the square root of a negative number?

Solution:

We have

$$f'(x) = \frac{1}{2} - \frac{a}{2x^2} = \frac{1}{2}\left[1 - \frac{a}{x^2}\right] \tag{6.1}$$

This leads to

$$
\begin{align}
|f'(x)| &< 1 \tag{6.2}\\
-2 &< \frac{a}{x^2} - 1 < 2 \tag{6.3}\\
-1 &< \frac{a}{x^2} < 3 \tag{6.4}\\
-1 &> \frac{x^2}{a} > \frac{1}{3} \tag{6.5}\\
-a &> x^2 > \frac{a}{3} \tag{6.6}\\
& \tag{6.7}
\end{align}
$$

Where I have assumed $a > 0$. Because $x^2 > 0$ for all real numbers, the first inequality is always true, that is, $-1 < a/x^2$ identically and so we can ignore the $-a > x^2$ part. Thus we only have $x^2 > a/3$ as our condition.

If $a < 0$ we instead find

$$|f'(x)| < 1 \tag{6.8}$$

$$-2 < 1 - \frac{a}{x^2} < 2 \tag{6.9}$$

$$-3 < \frac{-a}{x^2} < 1 \tag{6.10}$$

$$\frac{1}{-3} > \frac{x^2}{-a} > 1 \tag{6.11}$$

$$\frac{a}{3} > x^2 > -a \tag{6.12}$$

Here we note that because $-a/x^2 > 0$ we can ignore that inequality and so $x^2 > -a$ is our condition for convergence.

Note that with $x_1(x) = \frac{1}{2}\left[x_0 + \frac{a}{x_0}\right]$ and

$$x_2 = \frac{1}{2}\left[x_1 + \frac{a}{x_1}\right] = \frac{1}{2}\left[\frac{1}{2}\left[x_0 + \frac{a}{x_0}\right] + \frac{a}{\frac{1}{2}\left[x_0 + \frac{a}{x_0}\right]}\right] \tag{6.13}$$

$$x_{n+1} = \frac{1}{2}\left[x_n + \frac{a}{x_n}\right] \tag{6.14}$$

$$x_n^2 - 2x_{n+1}x_n + a = 0 \tag{6.15}$$

If we assume $a > 0$, $x_n > 0$ then the discriminant above must be greater than or equal to zero, so

$$4x_{n+1}^2 - 4a \geq 0 \tag{6.16}$$

$$x_{n+1}^2 \geq a \tag{6.17}$$

Thus,

$$x_n - x_{n+1} = x_n - \frac{x_n + \frac{a}{x_n}}{2} = \frac{x_n}{2} + \frac{a}{2x_n} = \frac{1}{2x_n}\left(x_n^2 - a\right) \tag{6.18}$$

Taking n large enough then $x_n^2 \geq a$ since $x_{n+1}^2 \geq a$ and so $x_n - x_{n+1} \geq 0$.

So we see that $x_n \geq \sqrt{a}$ and $x - x_{n+1} \geq 0$ as n increases. Because x_n is a decreasing sequence, it is rather clear it must approach the $\sqrt{a}$ value.

So for $a > 0$ and $x_0 > 0$ we have convergence no matter what happens.

We can get the negative square root by choosing $x_0 < 0$. Then we must check that (using that $x_n < 0$ so $-x_n > 0$)

$$|x_n| - |x_{n+1}| = x_n - \frac{x_n + \frac{a}{x_n}}{2} = \frac{x_n}{2} + \frac{a}{2x_n} = \frac{1}{2x_n}\left(x_n^2 - a\right) \tag{6.19}$$

and so the sequence $x_n - x_{n+1}$ is getting closer to zero, and will converge.

If $a < 0$ our above argument would seem to indicate that so long as $x^2 > -a$, we should get convergence, but this is not true. This is because given $x_0^2 > -a$, x_1 is reduced and so on, until $x_n^2 < -a$ and then we no longer converge to the correct result.

chapter6/sqrootiteration.py

```python
#!/usr/bin/env python2

import numpy as np

# Find square root of number a with initial guess x_0
def iterate(a,xin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  xg1=xin+2*tol
  while ((np.abs(xg2-xg1)>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=0.5*(xg1+a/xg1)
    itercount+=1
    xg4=0.5*(1 - a/xg2**2)
  return a,xg2,xin,xg4,itercount

print iterate(-1,-1.9)
print iterate(-1,1.1)
print iterate(-1,3.1)
```

6.1.2 2

If $f(x) = \sin(x) + 0.5x$, then $|f'(x)| \leq 0.5$ in the interval $[\pi/2, \pi]$. From this information, obtain an interval that contains all points of the iteration, $x_{n+1} = f(x_n)$, with $x_0 = 2$. Estimate how many iterations will be needed to solve $2\sin x - x = 0$ with an error less than 10^{-9}. Carry out the iteration, observe the number of iterations actually required and the magnitude of $|f'(x)|$ in the various intervals $[x_n, x_{n+1}]$.

Solution:

Let's find $x_1 = \sin(2) + 0.5(2) = 1.909$ so $|x_1 - x_0| \approx 0.0907$. Thus, our estimate of the distance is

$$|x_0 - x_1|/(1-k) \approx \frac{0.0907}{1-0.5} \approx 0.1814 \tag{6.20}$$

is the size of the region we expect to get farthest away from x_0 which is still within our region, so we're fine. We expect

$$10^{-9} = (0.5)^n \frac{|x_0 - x_1|}{(1-0.5)} \tag{6.21}$$

$$10^{-9} = e^{n \ln 0.5} \frac{|x_0 - x_1|}{1-0.5} \tag{6.22}$$

$$e^{\ln\left(10^{-9}(1-0.5)/|x_0-x_1|\right)} = e^{n \ln 0.5} \tag{6.23}$$

$$n = \frac{\ln 10^{-9} \frac{1-0.5}{|x_0-x_1|}}{\ln 0.5} \approx 27.4 \tag{6.24}$$

so about 28 iterations.

Running the code we find the answer is $x \approx 1.89549$ and it takes only 12 iterations. This is consistent with what was seen in Sawyer for his guess. Our guess is conservative, and the algorithm does quite a bit better.

Note the iteration comes from using if a is a fixed point of $f(x)$, so $f(a) = a$ then we can use $x_{n+1} = f(x_n)$ as an iteration.

Thus, in our case $2\sin(x) - x = 0$ we write it as $x = 2\sin(x)$ and then $x = \sin(x) + 0.5x$.

chapter6/siniteration.py

```
#!/usr/bin/env python2

import numpy as np

# Find solution to 2 sin(x) -x = 0
def iterate(xin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  xg1=xin+2*tol
  while ((np.abs(xg2-xg1)>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=np.sin(xg1)+0.5*xg1
    itercount+=1
  return xg2,xin,itercount

print iterate(2,1e-9)
```

6.1.3 3

Find $f'(4.5)$ for (a) $f(x) = \tan x$ (b) $f(x) = \pi + \tan^{-1}(x)$. The equation $x = \tan x$ has a solution in the neighborhood of $x = 4.5$. Which of the two functions mentioned above should be chosen to find this solution by means of the iteration $x_{n+1} = f(x_n)$ with $x_0 = 4.5$?

Solution:

(a) We have $\frac{\mathrm{d}\tan x}{\mathrm{d}x} = \sec^2(x)$ and $\sec^2(4.5) \approx 22.5$.

(b) To find the derivative of the arctangent, we use $\tan y = x$ (making triangles with this we find $\cos(y) = 1/\sqrt{1+x^2}$) so $y = \tan^{-1}(x)$ and

$$\sec^2 y\,\mathrm{d}y = \mathrm{d}x\frac{\mathrm{d}y}{\mathrm{d}x} \qquad = \frac{1}{\sec^2 y} = \cos^2(y) = \frac{1}{1+x^2} \tag{6.25}$$

so $f'(4.5) = 0.047$ is the value for this method.

We clearly should use method (b) since the derivative is much smaller, and so we expect much faster convergence.

Indeed, we find it takes 4 iterations for (b), but (a) actually converges to a different solution and takes nearly 5000 iterations to get there.

chapter6/taniteration.py

```
#!/usr/bin/env python2

import numpy as np

# Find solution to 2 sin(x) -x = 0
def iterate1(xin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  xg1=xin+2*tol
  while ((np.abs(xg2-xg1)>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=np.tan(xg1)
    itercount+=1
  return xg2,xin,itercount
```

```

def iterate2(xin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  xg1=xin+2*tol
  while ((np.abs(xg2-xg1)>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=np.pi + np.arctan(xg1)
    itercount+=1
  return xg2,xin,itercount

print iterate1(4.5)
print iterate2(4.5)
```

6.2 Differentiation Exercises

6.2.1 1

Find the 2×2 matrix that represents $f'(x,y)$ when $f(x,y)$ is specified by

6.2.1.1 a

$(-y,x)$

Solution:

Then the matrix is

$$\begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{6.26}$$

6.2.1.2 b

$(x+y,x-y)$

Solution:

Then the matrix is

$$\begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{6.27}$$

6.2.1.3 c

$(x+y,x+y)$

Solution:

Then the matrix is

$$\begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{6.28}$$

6.2.1.4 d

$(x^2 - y^2, 2xy)$

Solution:

Then the matrix is

$$\begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & -2y \\ 2y & 2x \end{bmatrix} \tag{6.29}$$

6.2.1.5 e

$(e^x \cos y, e^x \sin y)$

Solution:

Then the matrix is

$$\begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{bmatrix} \tag{6.30}$$

6.2.2 2

In section 6.2, generalized differentiation was discussed in the context of functions $\mathbb{R}^2 \to \mathbb{R}^2$. Adapt this discussion (a) to functions, $\mathbb{R}^2 \to \mathbb{R}$, such as $(x, y) \to x^2 + y^2$, (b) to functions $\mathbb{R} \to \mathbb{R}^2$ such as $t \to (a \cos t, b \sin t)$.

Solution:

(a) We use the Frechét definition so

$$f((x, y) + (h_x, h_y)) - f((x, y)) = M(h_x, h_y) + e((h_x, h_y)) \tag{6.31}$$

where $\|e((h_x, h_y))\| \to 0$ as $\|(h_x, h_y)\| \to 0$

In our example $(x, y) \to x^2 + y^2$, we see

$$\begin{aligned} f((x, y) + (h_x, h_y)) - f((x, y)) &= (x + h_x)^2 + (y + h_y)^2 - x^2 - y^2 \\ &= 2xh_x + h_x^2 + 2yh_y + h_y^2 = M(h_x, h_y) + e((h_x, h_y)) \\ M(a, b) &= 2a + 2b \end{aligned} \tag{6.32}$$

We can note that the Frechét derivative is the divergence of a vector from this definition. If we do a Taylor approximation in multiple dimensions it will be clear $M = \sum_i \frac{\partial f_i}{\partial x_i}$.

(b) We use the Frechét definition so

$$f((x, y) + (h_x, h_y)) - f((x, y)) = M(h_x, h_y) + e((h_x, h_y)) \tag{6.33}$$

where $\|e((h_x, h_y))\| \to 0$ as $\|(h_x, h_y)\| \to 0$

In our example $t \to (a\cos t, b\sin t)$, we see

$$\begin{aligned} f(t+h) - f(t) &= (a\cos(t+h), b\sin(t+h)) - (a\cos t, b\sin t) \\ &= (a\left[-\cos(t) + \cos(t+h)\right], b\left[-\sin(t) + \sin(t+h)\right]) \end{aligned} \tag{6.34}$$

Now as $h \to 0$ we find

$$\begin{aligned} &f(t+h) - f(t) \\ &= \left(a\left[-\cos(t) + \cos(t) - h\sin(t) - \frac{h^2}{2}\cos(t) + \mathcal{O}(h^3)\right],\right. \\ &\qquad\qquad \left. b\left[-\sin(t) + \sin(t) + h\cos(t) - \frac{h^2}{2}\sin(t) + \mathcal{O}(h^3)\right]\right) \end{aligned} \tag{6.35}$$

$$= (-ah\sin(t), bh\cos(t)) + e(h) = (-a\sin(t), b\cos(t))\,h + e(h) \tag{6.36}$$

Thus $Mt = (-a\sin(t), b\cos(t))$ is the linear transformation. In general, we'd find for $t \to (f_1(t), f_2(t), \ldots, f_N(t))$ that

$$Mt = \left(\frac{\partial f_1}{\partial t}, \frac{\partial f_2}{\partial t}, \ldots \frac{\partial f_N}{\partial t}\right) \tag{6.37}$$

so that M is the gradient operator.

6.3 Exercises On Functional Calculus Derivatives

6.3.1 Exercise 1

Let F be defined by $f \to g$. For a number of cases the value of $g(x)$ is shown below. For each of these, write the expression $F'(f)h \cdot (x)$ which defines the derivative $F'(f)$.

6.3.1.1 a

$[f(x)]^2$

Solution:

We simply get $F'(f) = 2f$.

6.3.1.2 b

$[f(x)]^n$

Solution:
Here we get $F'(f) = n[f(x)]^{n-1}$.

6.3.1.3 c

$x^3[f(x)]^2$

Solution:

We find $F'(f) = 2x^3 f$ because the x is unchanging for this operation.

6.3.1.4 d

$\sin(f(x))$

Solution:

We find $F'(f) = \cos(f(x))$.

6.3.1.5 e

$\int_0^x t[f(t)]^2 \, \mathrm{d}t$

Solution:

Here we find

$$F'(f) = \int_0^x 2tf(t) \, \mathrm{d}t \tag{6.38}$$

6.3.1.6 f

$f'(x)$

Solution:

Let's use the definition of $f'(x)$ so

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{6.39}$$

so

$$F'(f) = \lim_{h \to 0} \left[\frac{(f+h_1)(x+h) - (f+h_1)(x)}{h} - \frac{f(x+h) - f(x)}{h} \right] = \lim_{h \to 0} \frac{h_1(x+h) - h_1(x)}{h} = h_1'(x) \tag{6.40}$$

As there is nothing multiplying $h_1(x)$ we therefore say that the derivative is zero I'd think.

Note the simpler way to find this is

$$F'(f)h \cdot (x) = (f+h)'(x) - f'(x) = h'(x) \tag{6.41}$$

We then must say that there is no well defined $F'(f)$ unless we are willing to say $F'(f) = h'(x)/h(x)$ which has no functional dependence on f itself.

6.3.1.7 g

$f(x)f'(x)$

Solution:

We find

$$\begin{aligned} F'(f)h\cdot(x) &= (f+h)(x)(f+h)'(x) - f(x)f'(x) = \cancel{f(x)f'(x)} + f(x)h'(x) + h(x)f'(x) - \cancel{f(x)f'(x)} \\ &= f'(x)h(x) + f(x)h'(x) \end{aligned} \tag{6.42}$$

Once again we are faced with saying there is no pure $F'(f)$ since we would need to say

$$F'(f) = f'(x) + \frac{f(x)h'(x)}{h(x)} \tag{6.43}$$

6.3.1.8 h

$x\int_0^1 [f'(t)]^2\,\mathrm{d}t$

Solution:

Here we find

$$g = x\int_0^1 \mathrm{d}t\ \left[\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}\right]^2 = x\int_0^1 \mathrm{d}t\ \lim_{h\to 0}\frac{f(x+h)f(x+h)-2f(x)f(x+h)+f(x)f(x)}{h^2} \tag{6.44}$$

so

$$\begin{aligned} F'(f) = x\int_0^1 \mathrm{d}t\ \lim_{h\to 0}\Bigg[&\frac{(f+h_1)(x+h)(f+h_1)(x+h) - 2(f+h_1)(x)(f+h_1)(x+h)}{h^2} \\ &+ \frac{(f+h_1)(x)(f+h_1)(x)}{h^2} - \frac{f(x+h)f(x+h)-2f(x)f(x+h)+f(x)f(x)}{h^2}\Bigg] \end{aligned} \tag{6.45}$$

$$\begin{aligned} &= x\int_0^1 \mathrm{d}t\ \Bigg[\lim_{h\to 0}\frac{\cancel{f(t+h)f(t+h)} + 2f(t+h)h_1(t+h)}{h^2} \\ &\frac{-2f(t+h)h_1(t) - \cancel{2f(t)f(t+h)} - 2f(t)h_1(t+h) + \cancel{f(t)f(t)} + 2f(t)h_1(t)}{h^2} \\ &- \frac{\cancel{f(t+h)f(t+h)} - \cancel{2f(t)f(t+h)} + \cancel{f(t)f(t)}}{h^2} + \mathcal{O}(h_1^2)\Bigg] \end{aligned} \tag{6.46}$$

$$= x\int_0^1 \mathrm{d}t\ \lim_{h\to 0} 2\left[\frac{f(t+h)-f(t)}{h^2}h_1(t+h) - \frac{f(t+h)-f(t)}{h^2}h_1(t)\right] \tag{6.47}$$

Then we use

$$\frac{f(t+h)-f(t)}{h^2}h_1(t+h) = \frac{f(t+h)-f(t)}{h^2}(h_1(t) + hh_1'(t) + \mathcal{O}(h^2)) \tag{6.48}$$

so that we get

$$= x\int_0^1 \mathrm{d}t\ \lim_{h\to 0} 2\left[\frac{f(t+h)-f(t)}{h}h_1'(t)\right] \tag{6.49}$$

$$= x\int_0^1 \mathrm{d}t\ 2f'(t)h_1'(t) = x\int_0^1 \mathrm{d}t\ \left[2[f'(t)h_1(t)]' - 2f''(t)h_1(t)\right] \tag{6.50}$$

$$= x[2f'(1)h_1(1) - 2f'(0)h_1(0)] - x\int_0^1 \mathrm{d}t\ 2f''(t)h_1(t) \tag{6.51}$$

So we find

$$F'(f) = 2xf'(1) - 2xf'(0) + x\int_0^1 -2f''(t)\,\mathrm{d}t \tag{6.52}$$

More simply we could use

$$\begin{aligned} F'(f)h\cdot(x) &= x\int_0^1 \mathrm{d}t\;\left[[(f+h)'(t)]^2 - [f'(t)]^2\right]^2 = x\int_0^1 \mathrm{d}t\;[\cancel{f'(t)f'(t)} + 2f'(t)h'(t) - \cancel{f'(t)f'(t)}] \\ &= x\int_0^1 \mathrm{d}t\; 2f'(t)h'(t) = 2xf'(1)h(1) - 2xf'(0)h(0) - x\int_0^1 \mathrm{d}t\; 2f''(t)h(t) \end{aligned} \tag{6.53}$$

which is identical. We are again confronted with $F'(f)$ is not really well defined, although $F'(f)h\cdot(x)$ certainly is.

6.3.1.9 i

$\int_0^x \mathrm{d}t\;\;\{[f(t)]^2 + [f'(t)]^2\}$

Solution:

Here we find from our previous results that

$$F'(f) = \int_0^x 2f(t)\,\mathrm{d}t + 2f'(1) - 2f'(0) + x\int_0^1 -2f''(t)\,\mathrm{d}t \tag{6.54}$$

Or more accurately

$$\begin{aligned} F'(f)h\cdot(x) &= \int_0^x \mathrm{d}t\;\;[2f(t)h(t) + 2f'(t)h'(t)] \\ &= \int_0^x \mathrm{d}t\;\;[2f(t)h(t) - 2f''(t)h(t)] + 2xf'(1)h(1) - 2xf'(0)h(0) \end{aligned} \tag{6.55}$$

6.3.1.10 j

$\left[\int_0^x \mathrm{d}t\;\; f(t)\right]^2$

Solution:

Now we find

$$\begin{aligned} F'(f)h\cdot(x) &= \left[\int_0^x (f+h)(t)\,\mathrm{d}t\right]^2 - \left[\int_0^x f(t)\,\mathrm{d}t\right]^2 = \left[\int_0^x f(t)\,\mathrm{d}t + \int_0^x h(t)\,\mathrm{d}t\right]^2 - \left[\int_0^x f(t)\,\mathrm{d}t\right]^2 \\ &= \cancel{\left[\int_0^x f(t)\,\mathrm{d}t\right]^2} + 2\left[\int_0^x f(t)\,\mathrm{d}t\right]\left[\int_0^x h(t)\,\mathrm{d}t\right] - \cancel{\left[\int_0^x f(t)\,\mathrm{d}t\right]^2} \end{aligned} \tag{6.56}$$

6.4 Exercise 2

Let S be the function $\mathcal{C}[0,1] \to \mathcal{C}[0,1]$, $f \to g$, where

$$g(x) = \frac{x}{8} + \int_0^x [f(t)]^2 \, \mathrm{d}t$$

Find the derivative $S'(f)$ and show that $\|f\| \leq 0.25 \Rightarrow \|S'(f)\| \leq 0.5$. Hence show that the iteration $f_{n+1} = Sf_n$ with $f_0 = 0$ converges to a function belonging in $\overline{B}(0, 0.25)$.

Solution:

From the above we have

$$\begin{aligned} S'(f)h \cdot (x) &= \cancel{\frac{x}{8}} + \int_0^x [(f+h)(t)]^2 \, \mathrm{d}t - \cancel{\frac{x}{8}} - \int_0^x [f(t)]^2 \, \mathrm{d}t \\ &= \int_0^x \left[\cancel{[f(t)]^2} + 2f(t)h(t) - \cancel{[f(t)]^2} \right] = \int_0^x \mathrm{d}t \; 2f(t)h(t) \end{aligned} \tag{6.57}$$

$$S'(f) = 2 \int_0^x \mathrm{d}t \; f(t) \tag{6.58}$$

So assuming $\|f\| \leq 0.25$, then clearly

$$\|S'(f)\| = \left\| 2 \int_0^x \mathrm{d}t \; f(t) \right\| \leq 2 \int_0^1 \mathrm{d}t \;\; \|f(t)\| \leq 2(0.25) \leq 0.5 \tag{6.59}$$

Then $f_0 = 0$ implies $f_1 = \frac{x}{8}$ so $\|f_1\| = \frac{1}{8} = 0.125$ and so we see that we will get

$$f_n \leq \sum_{i=0}^{\infty} \frac{x^{2n+1}}{2^n(8)} = \frac{x}{4(2-x^2)} \leq \frac{1}{4} \tag{6.60}$$

This series is arrived at because

$$f_1 = \frac{x}{8} \tag{6.61}$$

$$f_2 = \frac{x}{8} + \int_0^x \frac{x^2}{16} = \frac{x}{8} + \frac{x^3}{192} < \frac{x}{8} + \frac{x^3}{16} \tag{6.62}$$

where because the integration will yield $n!$ in the denominator and $n! > 2^n$, our inequality will always hold.

Thus, $f_{n+1} = Sf_n$ will converge with $f_0 = 0$.

Note that the $f(x)$ that satisfies

$$f(x) = \frac{x}{8} + \int_0^x [f(t)]^2 \,\mathrm{d}t \tag{6.63}$$

$$f'(x) = \frac{1}{8} + [f(x)]^2 \tag{6.64}$$

$$f(x) = \frac{\tan(\frac{x}{2\sqrt{2}} + C)}{2\sqrt{2}} \tag{6.65}$$

$$f'(x) = \frac{\sec^2(\frac{x}{2\sqrt{2}} + C)}{8} \tag{6.66}$$

$$\frac{1}{8} + [f(x)]^2 = \frac{1 + \tan(\frac{x}{2\sqrt{2}} + C)^2}{8} = \frac{\sec^2(\frac{x}{2\sqrt{2}} + C)}{8} \tag{6.67}$$

We then can note that the series for this $f(x)$ matches that given by f_n as n increases.

6.5 Exercises on Iteration

In these questions the symbols f, S, and ϕ have the meanings attached to them in section 6.6

6.5.1 1

Find the operator S corresponding to the function $f : \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \to (xy + 0.07, x^2 + y^2 - 0.41)$ with the initial $(x_0, y_0) = (0.1, -0.6)$. Show that $S'(0.1 + X, -0.6 + Y)$ is given by a 2×2 matrix, M with $m_{11} = m_{22} = (2X + 12Y)/7$ and $m_{12} = m_{21} = (12X + 2Y)/7$. Hence show that $|X| \leq t$, $|Y| \leq t$ implies $\|S'(0.1 + X, -0.6 + Y)\| \leq 4t$. Deduce that the iteration $t \to \phi(t) = 0.0315 + 2t^2$, may be used for comparison with the modified Newton-Raphson iteration. Verify this by carrying out both iterations. Compare this method with that used at the beginning of Section 6.4

Solution:

We have (assuming $(x, y) \to (f_x, f_y)$)

$$f' = \begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} y & x \\ 2x & 2y \end{bmatrix} \tag{6.68}$$

So that

$$f'(x, y)^{-1} = \begin{bmatrix} \frac{y}{y^2 - x^2} & \frac{x}{2x^2 - 2y^2} \\ \frac{x}{x^2 - y^2} & \frac{y}{2y^2 - 2x^2} \end{bmatrix} \tag{6.69}$$

Thus,

$$(x_{n+1}, y_{n+1}) = (x_n, y_n) - \begin{bmatrix} \frac{y_0}{y_0^2 - x_0^2} & \frac{x_0}{2x_0^2 - 2y_0^2} \\ \frac{x_0}{x_0^2 - y_0^2} & \frac{y_0}{2y_0^2 - 2x_0^2} \end{bmatrix} (x_n y_n + 0.07, x_n^2 + y_n^2 - 0.41) \tag{6.70}$$

noting

$$f'(x_0,y_0)^{-1} = \begin{bmatrix} -\frac{12}{7} & -\frac{1}{7} \\ -\frac{2}{7} & -\frac{6}{7} \end{bmatrix} \tag{6.71}$$

so that

$$(x_{n+1},y_{n+1}) = \left(\frac{1}{7}\left(12x_ny_n + x_n(x_n+7)+y_n^2\right) + \frac{43}{700}, \frac{2}{7}\left(3x_n^2 + x_ny_n + 3y_n^2\right) + y_n - \frac{58}{175}\right) \tag{6.72}$$

Thus,

$$S' = \begin{bmatrix} \frac{\partial S_x}{\partial x} & \frac{\partial S_x}{\partial y} \\ \frac{\partial S_y}{\partial x} & \frac{\partial S_y}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{1}{7}\left(7+2x+12y\right) & \frac{2}{7}(6x+y) \\ \frac{2}{7}(6x+y) & \frac{1}{7}\left(7+2x+12y\right) \end{bmatrix} \tag{6.73}$$

We see clearly that $m_{11} = m_{22}$ and $m_{12} = m_{21}$ from this and so $S'(0.1+X,-0.6+Y)$ yields

$$S'(0.1+X,-0.6+Y) = \begin{bmatrix} \frac{2}{7}(X+6Y) & \frac{2}{7}(6X+Y) \\ \frac{2}{7}(6X+Y) & \frac{2}{7}(X+6Y) \end{bmatrix} \tag{6.74}$$

So that for $|X| \le t$ and $|Y| \le t$ we have

$$S'(0.1+X,-0.6+Y) = \begin{bmatrix} \frac{2}{7}(X+6Y) & \frac{2}{7}(6X+Y) \\ \frac{2}{7}(6X+Y) & \frac{2}{7}(X+6Y) \end{bmatrix} \le \begin{bmatrix} 2t & 2t \\ 2t & 2t \end{bmatrix} \tag{6.75}$$

and so

$$\|S'(0.1+X,-0.6+Y)\|_\infty \le 4t \tag{6.76}$$

We know that $\phi'(t) = 4t$ will therefore yield a good answer. Thus $\phi(t) = 2t^2 + c$ is required. We need (x_1,y_1) which yields

$$(x_1,y_1) = (\frac{39}{350}, \frac{221}{350}) \tag{6.77}$$

$$(x_1,y_1) - (x_0,y_0) = (\frac{2}{175}, \frac{-11}{350}) \approx (0.0114286, -0.0314286) \tag{6.78}$$

Thus choosing $\phi(0) > 0.0315$ will do and so indeed

$$\phi(t) = 2t^2 + 0.0315 \tag{6.79}$$

is a good iteration.

We can implement these in python.

chapter6/Stiteration.py

```
#!/usr/bin/env python2

import numpy as np



s1=12/7.
s2=1/7.
s3=43/700.
```

```
s4=6/7.
s5=2/7.
s6=58/175.

# Find solution to crossing point of xy=-0.07 and x^2+y^2=0.41
# Note this is equivalent to solving x**4-0.41x**2+0.0049=0
#  with solutions x = +-0.630618,+-0.111202
def xyiterate(xin,yin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=np.array([xin,yin])
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[0]=xg2[0]
    xg1[1]=xg2[1]
    #use iteration
    xg2[0]=s1*xg1[0]*xg1[1]+s2*xg1[0]**2 + xg1[0] + s2*xg1[1]**2 + s3
    xg2[1]=s4*xg1[0]**2+s5*xg1[0]*xg1[1]+s4*xg1[1]**2+ xg1[1]-s6
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err

# Find solution t_{n+1}=phi(t_n) for phi(x) = 0.0315+2t**2
def titerate(tin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=tin
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.abs(np.max(xg2-xg1))>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=0.0315+ 2.*xg1**2
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err

# Slower way of solving xyiterate above
def xyiterate2(xin,yin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=np.array([xin,yin])
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[0]=xg2[0]
    xg1[1]=xg2[1]
    #use iteration
    xg2[0]=xg1[0]*xg1[1]+xg1[0]+0.07
    xg2[1]=xg1[0]**2+xg1[1]**2+xg1[1]-0.41
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err


print xyiterate(0.1,-0.6)
print titerate(0)
print xyiterate2(0.1,-0.6)
```

The results are

iterate	N-R $\lVert (x_n, y_n) - (x_{n-1}, y_{n-1}) \rVert_\infty$	$\lvert \phi(t_n) - \phi(t_{n-1}) \rvert$	6.4 $\lVert (x_n, y_n) - (x_{n-1}, y_{n-1}) \rVert_\infty$
1	0.031428571428571472	0.0315	2.0000000000020002e-06
2	0.00085597667638492858	0.0019845000000000002	0.039999999999999925
3	4.7467115329880016e-05	0.00025792348050000108	0.011699999999999822
4	2.6599557030326793e-06	3.4678804174792521e-05	0.0032269499999999507
5	1.5187785007420018e-07	4.6829928239747187e-06	0.0010892640927249175
6	—	6.3275557129344184e-07	0.00027077859784385705
7	—	—	0.00010285104490803665
8	—	—	2.1657419990717131e-05
9	—	—	1.0118774463219182e-05
10	—	—	2.9775669416892692e-06
11	—	—	1.0604413587245176e-06
12	—	—	4.6128826033942083e-07

Table 6.1: Error versus iterates for two methods and the test method. N-R is the Newton-Raphson method and 6.4 is the method from section 6.4. We see that $t \to \phi(t)$ is a conservative guess for N-R and that N-R is much better than the method of 6.4, as we expected. This program went until the error was less than 10^{-6}.

6.5.2 2

For $f : \mathbb{R} \to \mathbb{R}$, $(x, y) \to (x^2 + y^2 - 200, y^3 + xy - x^3)$ and the initial $(x_0, y_0) = (10, 10)$ show that $[f'(x_0, y_0)]^{-1} = N$, where

$$N = \frac{1}{12000} \begin{bmatrix} 310 & -20 \\ 290 & 20 \end{bmatrix}$$

Show that

$$S'(x, y) = \mathbb{1} - N \begin{bmatrix} 2x & 2y \\ y - 3x^2 & 3y^2 + x \end{bmatrix}$$

What is $S'(x_0, y_0)$? Is this result an accident or will it always happen with the Newton-Raphson method? Find $S'(10 + X, 10 + Y)$ and show that $\|S'(10 + X, 10 + Y)\|_\infty$ does not exceed $(t/3) + t^2/100$ when $|X| \leq t$ and $|Y| \leq t$. (This can be shown by a crude approximation in which every term is replaced by its absolute value.) Hence show that the comparison function ϕ can be defined by $\phi(t) = \frac{1}{6} + \frac{t^2}{6} + \frac{t^3}{300}$. Check this by carrying out the iterations both of S and of ϕ.

Solution:

First

$$f'(x, y) = \begin{bmatrix} \frac{\partial f_x}{\partial x} & \frac{\partial f_x}{\partial y} \\ \frac{\partial f_y}{\partial x} & \frac{\partial f_y}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ y - 3x^2 & 3y^2 + x \end{bmatrix} \tag{6.80}$$

Thus,

$$[f'(x_0, y_0)]^{-1} = \begin{bmatrix} 2x & 2y \\ y - 3x^2 & 3y^2 + x \end{bmatrix} = \begin{bmatrix} \frac{3y^2+x}{2(x+y)(3yx+x-y)} & -\frac{y}{(x+y)(3yx+x-y)} \\ -\frac{y-3x^2}{2(x+y)(3yx+x-y)} & \frac{x}{(x+y)(3yx+x-y)} \end{bmatrix} = \begin{bmatrix} \frac{31}{1200} & -\frac{1}{600} \\ \frac{29}{1200} & \frac{1}{600} \end{bmatrix} = N \tag{6.81}$$

as desired. Thus,

$$\begin{aligned} (x_{n+1}, y_{n+1}) = S(x_n, y_n) &= (x_n, y_n) - [f'(x_0, y_0)]^{-1} f(x_n, y_n) & (6.82) \\ &= (x_n, y_n) \\ &\quad - \left(\frac{31(-200 + x_n^2 + y_n^2) + 2(x_n^3 - x_n y_n - y_n^3)}{1200}, \frac{29(-200 + x_n^2 + y_n^2) + 2(y_n^3 + x_n y_n - x_n^3)}{1200} \right) & (6.83) \\ = \left(x_n - \frac{31(-200 + x_n^2 + y_n^2) - 2(y_n^3 + x_n y_n - y_n^3)}{1200}, y_n - \frac{29(-200 + x_n^2 + y_n^2) + 2(y_n^3 + x_n y_n - x_n^3)}{1200} \right) & & (6.84) \end{aligned}$$

thus

$$\begin{aligned} S'(x, y) = \begin{bmatrix} \frac{\partial S_x}{\partial x} & \frac{\partial S_x}{\partial y} \\ \frac{\partial S_y}{\partial x} & \frac{\partial S_y}{\partial y} \end{bmatrix} &= \begin{bmatrix} 1 - \frac{62x+2(3x^2-y)}{1200} & -\frac{62y-2(x+3y^2)}{1200} \\ -\frac{58x-2(3x^2-y)}{1200} & 1 - \frac{58y+2(x+3y^2)}{1200} \end{bmatrix} = \mathbb{1} - \begin{bmatrix} \frac{62x+2(3x^2-y)}{1200} & \frac{62y-2(x+3y^2)}{1200} \\ \frac{58x-2(3x^2-y)}{1200} & \frac{58y+2(x+3y^2)}{1200} \end{bmatrix} & (6.85) \\ &= \mathbb{1} - N \begin{bmatrix} 2x & 2y \\ y - 3x^2 & 3y^2 + x \end{bmatrix} & (6.86) \end{aligned}$$

This will always be true since N is independent of x, y

$$S(x,y) = \mathbb{1}(x,y) - \overbrace{[f'(x_0,y_0)]^{-1}}^{N} f(x,y) \tag{6.87}$$
$$S'(x,y) = \mathbb{1} - N f'(x,y) \tag{6.88}$$

which is exactly what we found.

Thus,

$$S'(10+X, 10+Y) = \begin{bmatrix} \frac{1}{600}(Y - X(3X+91)) & \frac{1}{600}(X + Y(3Y+29)) \\ \frac{1}{600}(X(3X+31) - Y) & \frac{1}{600}(-X - Y(3Y+89)) \end{bmatrix} \tag{6.89}$$

Using

$$\|S'(10+X,10+Y)\|_\infty \le \left\| \begin{bmatrix} \frac{1}{600}(|Y| + |X|(3|X|+91)) & \frac{1}{600}(|X| + |Y|(3|Y|+29)) \\ \frac{1}{600}(|X|(3|X|+31) + |Y|) & \frac{1}{600}(|X| + |Y|(3|Y|+89)) \end{bmatrix} \right\|_\infty \tag{6.90}$$
$$\le \max\left\{ \frac{3t^2 + 92t + 3t^2 + 30t}{600}, \frac{3t^2 + 32t + 3t^2 + 90t}{600} \right\} \tag{6.91}$$
$$\le \max\left\{ \frac{t^2}{100} + \frac{122}{600}t, \frac{t^2}{100} + \frac{122t}{600} \right\} \tag{6.92}$$
$$\le \frac{t^2}{100} + \frac{t}{3} \tag{6.93}$$

Now let's find

$$\|(x_1,y_1) - (x_0,y_0)\|_\infty = \|(59/6, 61/6) - (10,10)\|_\infty = \|(1/6, -1/6)\|_\infty = \frac{1}{6} \tag{6.94}$$

Thus, we have

$$\phi(t) = \frac{t^3}{300} + \frac{t^2}{6} + a = \frac{t^3}{300} + \frac{t^2}{6} + \frac{1}{6} \tag{6.95}$$

as our test function so that $|\phi(0)| > \|(x_1 - x_0, y_1 - y_0)\|_\infty$.

Carrying out the iterations with the following program will yield the table below. We find $(x,y) \approx (10.16521699, 9.83200709)$.

chapter6/s10iteration.py

```python
#!/usr/bin/env python2

import numpy as np



s1=31*2/12.
s2=31/1200.
s3=1/600.
s4=29*2/12.
s5=29/1200.

# Find solution crossing of x^2+y^2-200=0 and y^3+xy-x^3=0
def xyiterate(xin,yin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=np.array([xin,yin])
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[0]=xg2[0]
    xg1[1]=xg2[1]
    #use iteration
    xg2[0]=xg1[0]+s1-s2*(xg1[0]**2+xg1[1]**2)+s3*(xg1[1]**3+xg1[0]*xg1[1]-xg1[0]**3)
    xg2[1]=xg1[1]+s4-s5*(xg1[0]**2+xg1[1]**2)-s3*(xg1[1]**3+xg1[0]*xg1[1]-xg1[0]**3)
    print xg2
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err

# Find solution t_{n+1}=phi(t_n) for phi(t) = 1/6+t^2/6+t^3/300
s6=1/6.
s7=1/300.
def titerate(tin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=tin
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.abs(np.max(xg2-xg1))>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=s6*(1+xg1**2)+xg1**3
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err


print xyiterate(10.,10.)
print titerate(0)
```

iterate	N-R $\|(x_n, y_n) - (x_{n-1}, y_{n-1})\|_\infty$	$\lvert\phi(t_n) - \phi(t_{n-1})\rvert$
1	0.16666666666666607	0.16666666666666666
2	0.0014969135802456179	0.0092592592592592449
3	4.8116429892886003e-05	0.0013439579840471283
4	8.8120274632785822e-07	0.00020485532225994474
5	—	3.1446710607341277e-05
6	—	4.8324814503619695e-06
7	—	7.4274005490426731e-07

Table 6.2: Error versus iterates for the method and the test method. N-R is the Newton-Raphson method. We see that $t \to \phi(t)$ is a conservative guess for N-R . This program went until the error was less than 10^{-6}.

6.5.3 3

The function $f : \mathbb{R}^n \to \mathbb{R}^n$ is defined by $f(v) = Mv + g(v)$ with constant matrix M. The equation, $f(v) = 0$ is to be solved by the Newton-Raphson method with the initial vector v_0. It is given that $g'(v_0) = 0$. Prove that S is the function $v \to -M^{-1}g(v)$. Find S for the function $f : \mathbb{R}^2 \to \mathbb{R}^2$, $(x, y) \to (-37x + 9y + x^5 + y^5 + 25, 4x - 28y + x^3y^3 + 18)$ with $(x_0, y_0) = (0, 0)$. Show that, when $|x| \leq t$ and $|y| \leq t$, then

$$\|S'(x,y)\|_\infty \leq 0.28t^4 + 0.222t^5$$

Deduce that a possible comparison function ϕ is given by

$$\phi(t) = 0.862 + 0.056t^5 + 0.037t^6$$

Carry out the iterations both of S and ϕ. (Note: for the relevant values of t an improved choice of ϕ can be found)

Solution:

We begin with

$$f'(v) = M + g'(v) \tag{6.96}$$
$$f'(v_0) = M + g'(v_0) = M \tag{6.97}$$

So

$$[f'(v_0)]^{-1} = M^{-1} \tag{6.98}$$

so that

$$f(v_n) = v_{n+1} = v_n - M^{-1}f(v_n) = v_n - M^{-1}(Mv_n + g(v_n)) = v_n - v_n - M^{-1}g(v_n) \tag{6.99}$$
$$S(v_n) = -M^{-1}g(v_n) \tag{6.100}$$
$$f(v) = -M^{-1}g(v) \tag{6.101}$$

as desired. For $f : \mathbb{R}^2 \to \mathbb{R}^2$ we see that

$$\begin{aligned} f(x,y) &= (-37x + 9y + x^5 + y^5 + 25, 4x - 28y + x^3y^3 + 18) \\ &= \begin{bmatrix} -37 & 9 \\ 4 & -28 \end{bmatrix} (x,y) + (x^5 + y^5 + 25, x^3y^3 + 18) \end{aligned} \tag{6.102}$$
$$= Mv + g(v) \tag{6.103}$$

with $v = (x, y)$. Note

$$g'(x,y) = \begin{bmatrix} 5x^4 & 5y^4 \\ 3x^2y^3 & 3x^3y^2 \end{bmatrix} \tag{6.104}$$

so $g'(v_0) = 0$ as needed for our theorem above. Thus,

$$S(v) = M^{-1}g(v) \tag{6.105}$$
$$S'(v) = M^{-1}g'(v) = \begin{bmatrix} -37 & 9 \\ 4 & -28 \end{bmatrix}^{-1} \begin{bmatrix} 5x^4 & 5y^4 \\ 3x^2y^3 & 3x^3y^2 \end{bmatrix} = \frac{-1}{1000} \begin{bmatrix} 28 & 9 \\ 4 & 37 \end{bmatrix} \begin{bmatrix} 5x^4 & 5y^4 \\ 3x^2y^3 & 3x^3y^2 \end{bmatrix} \tag{6.106}$$
$$= \frac{-1}{1000} \begin{bmatrix} 28(5x^4) + 9(3x^2y^3) & 28(5y^4) + 9(3x^3y^2) \\ 4(5x^4) + 37(3x^2y^3) & 4(5y^4) + 37(3x^3y^2) \end{bmatrix} \tag{6.107}$$
$$= \frac{-1}{1000} \begin{bmatrix} 140x^4 + 27x^2y^3 & 140y^4 + 27x^3y^2 \\ 20x^4 + 111x^2y^3 & 20y^4 + 111x^3y^2 \end{bmatrix} \tag{6.108}$$

Thus, we find for $|x| < t$ and $|y| < t$ that

$$\left\| \frac{-1}{1000} \begin{bmatrix} 140x^4 + 27x^2y^3 & 140y^4 + 27x^3y^2 \\ 20x^4 + 111x^2y^3 & 20y^4 + 111x^3y^2 \end{bmatrix} \right\|_\infty \leq \left\| \frac{1}{1000} \begin{bmatrix} 140t^4 + 27t^5 & 140t^4 + 27t^5 \\ 20t^4 + 111t^5 & 20t^4 + 111t^5 \end{bmatrix} \right\|_\infty \tag{6.109}$$

$$\leq \max \left\{ \frac{280}{1000}t^4 + \frac{54}{1000}t^5, \frac{40}{1000}t^4 + \frac{222}{1000}t^5 \right\} \leq \frac{280}{1000}t^4 + \frac{222}{1000}t^5 \tag{6.110}$$

Thus,

$$\phi'(t) = 0.28t^4 + 0.222t^5 \tag{6.111}$$

$$\phi(t) = a + 0.056t^5 + 0.037t^6 \tag{6.112}$$

The first iteration yields

$$(x_1, y_1) = \frac{1}{1000} \begin{bmatrix} 28 & 9 \\ 4 & 37 \end{bmatrix} (25, 18) = (0.862, 0.766) \tag{6.113}$$

So making $|\phi(0)| > \|(x_1 - x_0, y_1 - y_0)\|_\infty = 0.862$. Therefore a good test function is

$$\phi(t) = 0.862 + 0.056t^5 + 0.037t^6 \tag{6.114}$$

Note

$$S(v_n) = (\frac{28x_n^5 + 28y_n^5 + 700}{1000} + \frac{9x_n^3y_n^3 + 162}{1000}, \frac{4x_n^5 + 4y_n^5 + 100}{1000} + \frac{37x_n^3y_n^3 + 666}{1000}) \tag{6.115}$$

$$= \left(\frac{28x_n^5 + 28y_n^5 + 9x_n^3y_n^3 + 862}{1000}, \frac{4x_n^5 + 4y_n^5 + 37x_n^3y_n^3 + 766}{1000} \right) \tag{6.116}$$

We can test as we have before and form a table. Note the approximate solution is $(x, y) \approx (0.88871945, 0.78179593)$.

chapter6/sqiteration.py

```
#!/usr/bin/env python2

import numpy as np



s1=0.028
s2=0.009
s3=0.862
s4=0.004
s5=0.037
s6=0.766

# Find solution crossing of -37x+9y+x^5+y^5+25=0 and 4x-28y+x^3y^3+18=0
def xyiterate(xin,yin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=np.array([xin,yin])
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[0]=xg2[0]
    xg1[1]=xg2[1]
    #use iteration
    xg2[0]=s1*(xg1[0]**5+xg1[1]**5)+s2*xg1[0]**3*xg1[1]**3+s3
    xg2[1]=s4*(xg1[0]**5+xg1[1]**5)+s5*xg1[0]**3*xg1[1]**3+s6
    print xg2
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err

# Find solution t_{n+1}=phi(t_n) for phi(t) = 0.862+0.056t^5+0.037*t^6
s7=0.862
s8=0.056
s9=0.037
def titerate(tin,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=tin
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.abs(np.max(xg2-xg1))>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=s7+s8*xg1**5+s9*xg1**6
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err


print xyiterate(0.,0.)
print titerate(0.)
```

iterate	N-R $\|(x_n, y_n) - (x_{n-1}, y_{n-1})\|_\infty$	$\lvert\phi(t_n) - \phi(t_{n-1})\rvert$
1	0.8619999999999999	0.8619999999999999
2	0.023300905889619972	0.041830741063889398
3	0.0029489501040541599	0.012117311147522125
4	0.00040447316229497154	0.0040036202722723013
5	5.6095039640324806e-05	0.001375264700676726
6	7.7920815866328041e-06	0.00047855614292680038
7	1.0826349360337773e-06	0.0001672674888629766
8	1.5042662826481035e-07	5.8554859815318494e-05
9	—	2.0509239661681278e-05
10	—	7.1848634624060992e-06
11	—	2.5171918762723067e-06
12	—	8.819099203138947e-07

Table 6.3: Error versus iterates for the method and the test method. N-R is the Newton-Raphson method. We see that $t \to \phi(t)$ is a conservative guess for N-R . This program went until the error was less than 10^{-6}.

6.5.4 4

The function $S : \mathcal{C}[0,a] \to \mathcal{C}[0,1]$, where $a > 0$, is defined by $y \to z$ with

$$z(x) = \int_0^x [y(t) + t]^2 \, \mathrm{d}t$$

Show that with $y_0(x) = 0$, the behavior of the iteration of S can be estimated by means of the comparison function ϕ, with $\phi(t) = a^3/3 + a^2 t + at^2$. Show that the iteration of S arises naturally from the Newton-Raphson procedure for solving the differential equation $\frac{\mathrm{d}u}{\mathrm{d}x} - (x+u)^2 = 0$, with the condition $u(0) = 0$ and the initial $u_0(x) = -x$. Verify that $u_1 = y_0$.

Solution:

We begin with

$$\begin{aligned} S'(y)h\cdot(x) &= \int_0^x [y(t)+h(t)+t]^2 - [y(t)+t]^2\,\mathrm{d}t \\ &= \int_0^x \left[\cancel{[y(t)]^2} + 2y(t)h(t) + \cancel{[h(t)]^2} + \cancel{t^2} + 2t[\cancel{y(t)} + h(t)] - \cancel{[y(t)]^2} - \cancel{2ty(t)} - \cancel{t^2}\right]\,\mathrm{d}t \\ &= \int_0^x [2y(t)h(t) + 2th(t)]\,\mathrm{d}t \end{aligned} \tag{6.117}$$

$$S'(y) = 2\int_0^x [y(t)+t]\,\mathrm{d}t \tag{6.118}$$

Thus, we need for $\|y(t)-y_0\| < s$ (where $y_0 \equiv 0$) we have

$$\|S'(y)\| \le 2\sup\left\{\int_0^x |y(t)+t|\,\mathrm{d}t\right\} \le 2\sup\left\{\int_0^x s\,\mathrm{d}t + \int_0^x t\,\mathrm{d}t\right\} \tag{6.119}$$

$$2\sup\left\{\int_0^a s\,\mathrm{d}t + \int_0^a t\,\mathrm{d}t\right\} \le 2as + a^2 \tag{6.120}$$

or if we rewrite in terms of t,

$$\|S'(y)\| \le \phi'(t) = a^2 + 2at \tag{6.121}$$

Now, if we plug in $y_0 = 0$ we find

$$y_1 = \int_0^x t^2\,\mathrm{d}t = \frac{x^3}{3} \tag{6.122}$$

which is maximized by $x = a$ for this interval. This yields that $\phi(0) = \frac{a^3}{3}$ so

$$\phi(t) = \frac{a^3}{3} + a^2t + at^2 \tag{6.123}$$

just as desired.

For the differential equation $u' = (x+u)^2$, We note the error function can simply be

$$f(u) = u' - x^2 - 2xu - u^2 \tag{6.124}$$

Now

$$f'(u)h\cdot(x) = \cancel{u'} + h' - \cancel{x^2 - 2xu} - 2xh - \cancel{u^2} - 2uh - \cancel{h^2} - \cancel{u'} + \cancel{x^2 + 2xu + u^2} \tag{6.125}$$

$$= h' - 2xh - 2uh \tag{6.126}$$

Now, we can use that $u_0 = -x$ which gives

$$f'(u_0)h\cdot(x) \equiv k(x) = h' - 2xh + 2xh = h' \tag{6.127}$$

Thus, using $h(0) = 0$ since $u(0) = 0$ and we need agreement at this point for $k(x)$ as well, so that

$$\int_0^x \mathrm{d}s\ k(s) = h(x) - h(0) = h(x) \tag{6.128}$$

This of course would yield

$$[f'(u_0)]^{-1}[g(x)] = \int_0^x \mathrm{d}t\ \ [g(t)] \tag{6.129}$$

so

$$u_{n+1} = u_n - \int_0^x \mathrm{d}t\ f(u_n) = u_n - \int_0^x \mathrm{d}t\ \left[u_n' - (t + u_n)^2\right] = u_n - u_n + \cancel{u_n(0)} + \int_0^x \mathrm{d}t\ (u_n + t)^2 \tag{6.130}$$

$$= \int_0^x \mathrm{d}t\ (u_n + t)^2 \tag{6.131}$$

which is the S we desired since $u_n(0) = 0$. We find

$$u_1 = \int_0^x [-x + x]^2 = 0 = y_0 \tag{6.132}$$

as desired.

6.5.5 5

The function $f : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$, $y \to z$ has

$$z(x) = b - y(x) + ax[y(x)]^2 + a\int_1^x [y(s)]^2 \,\mathrm{d}s \qquad \text{with } a > 0\,, b > 0\,.$$

With initial $y_0(x) \equiv 0$, find the function S used to solve the equation $f(y) = 0$. (The inverse of $f'(y_0)$, often hard to find is here immediately evident.) Show that

$$S'(y_0)h \cdot (x) = 2axy(x)h(x) + a\int_1^x 2y(s)h(s)\,\mathrm{d}s \tag{6.133}$$

Deduce that $\phi(t) = b + at^2$ can be used in the Kantorovich comparison iteration. For the case $a = 1$, $b = 0.09$, find y_1, y_2, y_3, and compare $\|y_{n+1} - y_n\|$ with $t_{n+1} - t_n$ for $n = 0, 1, 2$. What unusual thing happens in this example?

Solution:

We immediately have

$$\begin{aligned} f'(y)h \cdot (x) = {} & \cancel{b} - \cancel{y} - h + ax\left(\cancel{y^2} + 2yh + \cancel{h^2}\right) + a\int_1^x \mathrm{d}s\; \cancel{y^2} + 2yh + \cancel{h^2} \\ & - \left\{\cancel{b - y(x) + ax[y(x)]^2 + a\int_1^x [y(s)]^2 \,\mathrm{d}s}\right\} \end{aligned} \tag{6.134}$$

$$= -h + 2axyh + 2a\int_1^x \mathrm{d}s\; yh \tag{6.135}$$

$$= -h(1 + 2axy) + 2a\int_1^x \mathrm{d}s\; yh \tag{6.136}$$

With $y_0 = 0$ we find

$$f'(y_0)h \cdot (x) = -h \tag{6.137}$$

and so $f'(y_0) = -1$. Thus,

$$S(y_n) = y_n + f(y_n) = y_n + b - y_n + axy_n^2 + a\int_1^x y_n^2 \,\mathrm{d}s \tag{6.138}$$

$$S(y) = y - [f'(y_0)]^{-1}f(y) = b + axy^2 + a\int_1^x y^2 \,\mathrm{d}s \tag{6.139}$$

Thus,

$$\begin{aligned} S'(y)h \cdot x = {} & \cancel{b} + ax(\cancel{y^2} + 2yh + \cancel{h^2}) + a\int_1^x (\cancel{y^2} + 2yh + \cancel{h^2})\,\mathrm{d}s \\ & - \left\{\cancel{b + axy^2 + a\int_1^x y^2 \,\mathrm{d}s}\right\} \end{aligned} \tag{6.140}$$

$$= 2axyh + 2a\int_1^x \mathrm{d}s\; yh \tag{6.141}$$

$$S'(y) = 2axy + 2a\int_1^x \mathrm{d}s\; y \tag{6.142}$$

matching what we desired. Thus, with $\|y-y_0\| = \|y\| \leq t$

$$\|S'(y)\| \leq 2a\,\|xy\| + 2a\sup\left\{\int_1^x \mathrm{d}s\ \ \|y\|\right\} = 2axt + 2a(x-1)t = 4axt - 2at \leq 2at \tag{6.143}$$

Thus $\phi'(t) = 2at$ is a good test. We now need y_1, so

$$y_1 = b + axy_0^2 + a\int_1^x \mathrm{d}s\ y_0^2 = b \tag{6.144}$$

and so $\phi(0) = b$ yielding

$$\phi(t) = at^2 + b \tag{6.145}$$

as desired.

We have

$$y_0 = 0 \tag{6.146}$$
$$y_1 = b \tag{6.147}$$
$$y_2 = b + axb^2 + a\int_1^x \mathrm{d}sb^2 = b + axb^2 + ab^2(x-1) = b + 2axb^2 - ab^2 \tag{6.148}$$
$$\begin{aligned} y_3 &= b + ax[b + 2ab^2x - ab^2]^2 + a\int_1^x \mathrm{d}s\ [b - ab^2 + 2ab^2s]^2 \\ &= \frac{16}{3}a^3b^4x^3 - 6a^3b^4x^2 + 2a^3b^4x - \frac{a^3b^4}{3} + 6a^2b^3x^2 - 4a^2b^3x + 2ab^2x - ab^2 + b \end{aligned} \tag{6.149}$$

Thus, for $a = 1$ and $b = 0.09$ we find

$$y_0 = 0 \tag{6.150}$$
$$y_1 = 0.09 \tag{6.151}$$
$$y_2 = b + 2axb^2 - ab^2 = 0.09 + 0.0081(-1+x) + 0.0081x = 0.0819 + 0.0162x \tag{6.152}$$
$$\begin{aligned} y_3 &= \frac{16}{3}a^3b^4x^3 - 6a^3b^4x^2 + 2a^3b^4x - \frac{a^3b^4}{3} + 6a^2b^3x^2 - 4a^2b^3x + 2ab^2x - ab^2 + b \\ &= 0.0818781 + 0.0134152x + 0.00398034x^2 + 0.00034992x^3 \end{aligned} \tag{6.153}$$

Thus,

$$\|y_1 - y_0\| = 0.09 \tag{6.154}$$
$$\|y_2 - y_1\| = 0.0081 \tag{6.155}$$
$$\|y_3 - y_2\| = 0.00152361 \tag{6.156}$$

where we see the norm is achieved at $x = 1$ for all of the function tested. For ϕ we find

$$t_0 = 0 \tag{6.157}$$
$$t_1 = at_0^2 + b = b \tag{6.158}$$
$$t_2 = at_1^2 + b = ab^2 + b \tag{6.159}$$
$$t_3 = at_2^2 + b = a(ab^2+b)^2 + b = a(a^2b^4 + 2ab^3 + b^2) + b = a^3b^4 + 2a^2b^3 + ab^2 + b \tag{6.160}$$

and for $a = 1$ and $b = 0.09$

$$|t_1 - t_0| = 0.09 \tag{6.161}$$
$$|t_2 - t_1| = |ab^2| = 0.0081 \tag{6.162}$$
$$|t_3 - t_2| = |a^3b^4 + 2a^2b^3| = 0.00152361 \tag{6.163}$$

And so we see we get exactly the same as our test function.

We can note that numerically we find a similar result (see Figure 6.1).

chapter6/intiteration.py

```python
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

# Find solution for a function y(x), such that f(y)=0
# Here f(y) = b - y +ax[y(x)]^2 + a int_1^x ds [y(s)]^2
# So S(y) = b+ axy^2 + a int_1^x ds y^2
#
# xarray should be the values used for xin
# xin should be an array that is a numerical approximation
#  to the initial function guess over x
def funiterate(xarray,xin,a,b,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  a=1.*a
  b=1.*b
  alen=np.shape(xin)[0]
  ints=np.zeros(alen)
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[:]=np.copy(xg2[:])
    #use iteration
    for j in range(alen):
      ints[j]=-np.trapz(xg1[j:]**2,xarray[j:])
    xg2=b+a*xarray*xg1**2 + a*ints
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err[1:]

# Find solution t_{n+1}=phi(t_n) for phi(x) = 0.0315+2t**2
def titerate(tin,a,b,tol=1e-6,maxitercount=1e5):
  itercount=0
  a=1.*a
  b=1.*b
  xg2=tin
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.abs(np.max(xg2-xg1))>tol)and(itercount<maxitercount)):
    xg1=xg2
    xg2=a*xg1**2+b
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err[1:]

# Find solution for a function y(x), such that f(y)=0
# Here f(y) = y(x) + int_0^x y(s) ds - b - a[y(x)]^2
# So S(y) = a[y(x)]^2+b e^(-x) - e^(-x) int_0^x ds a e^(s)[(y(s)]^2
#
# xarray should be the values used for xin
# xin should be an array that is a numerical approximation
#  to the initial function guess over x
def funiterate2(xarray,xin,a,b,tol=1e-6,maxitercount=1e5):
  itercount=0
  xg2=xin
  a=1.*a
  b=1.*b
  alen=np.shape(xin)[0]
  ints=np.zeros(alen)
  #ensure first iteration occurs
  xg1=xg2+2*tol
  err = [np.max(np.abs(xg2-xg1))]
  while ((np.max(np.abs(xg2-xg1))>tol)and(itercount<maxitercount)):
    #copy xg1 from xg2
    xg1[:]=np.copy(xg2[:])
    #use iteration
    for j in range(alen):
      ints[j]=np.trapz(np.exp(xarray[:j])*xg1[:j]**2,xarray[:j])
    xg2=a*xg1**2 + b*np.exp(-xarray) - np.exp(-xarray)*a*ints
    err.append(np.max(np.abs(xg2-xg1)))
    itercount+=1
  return xg2,itercount,err[1:]

# print out solutions
x=np.linspace(0,1,101)
y=0*x
print funiterate(x,y,1,0.09)
print titerate(0,1,0.09)

print funiterate2(x,y,1,0.09)
print titerate(0,1,0.09)

# actual solutions, ya1 for problem 5 and ya2 for problem 6
xa=np.linspace(0,1,1001)
ya1=0.0818737 + 0.0134081*xa + 0.00329524*xa**2 + 0.00110903*xa**3 + 0.00020512*xa**4 + 0.0000302779*xa**5 + 3.24987e-6*xa**6 + 1.39936e-7*xa**7
ya2= 2.58494e-25 + 1.39936e-7*np.exp(-8*xa) + 3.24987e-6*np.exp(-7*xa) + 0.0000302779*np.exp(-6*xa) + 0.00020512*np.exp(-5*xa) + 0.00110903*np.exp(-4*xa) + 0.00329524*np.exp(-3*xa) + 0.0134081*np.exp(-2*xa) + 0.0818737*np.exp(-xa)


# make plot for problem 5
x=np.linspace(0,1,3)
y=x*0.
g=funiterate(x,y,1,0.09)[0]

x1=np.linspace(0,1,11)
y1=x1*0.
g1=funiterate(x1,y1,1,0.09)[0]

x2=np.linspace(0,1,101)
y2=x2*0.
g2=funiterate(x2,y2,1,0.09)[0]

x3=np.linspace(0,1,501)
y3=x3*0.
g3=funiterate(x3,y3,1,0.09)[0]

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,g,label=r'$'+str(np.shape(x)[0])+r'\rm{\ points}$')
ax.plot(x1,g1,label=r'$'+str(np.shape(x1)[0])+r'\rm{\ points}$')
ax.plot(x2,g2,label=r'$'+str(np.shape(x2)[0])+r'\rm{\ points}$')
ax.plot(x3,g3,label=r'$'+str(np.shape(x3)[0])+r'\rm{\ points}$')
ax.plot(xa,ya1,label=r'$y_4(x)$')

plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')
```

```

plt.tight_layout()
plt.savefig('approxy1.png',bbox_inches='tight')

# make plot for problem 6
x=np.linspace(0,1,3)
y=x*0.
g=funiterate2(x,y,1,0.09)[0]

x1=np.linspace(0,1,11)
y1=x1*0.
g1=funiterate2(x1,y1,1,0.09)[0]

x2=np.linspace(0,1,101)
y2=x2*0.
g2=funiterate2(x2,y2,1,0.09)[0]

x3=np.linspace(0,1,501)
y3=x3*0.
g3=funiterate2(x3,y3,1,0.09)[0]

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,g,label=r'$'+str(np.shape(x)[0])+r'\rm{\␣points}$')
ax.plot(x1,g1,label=r'$'+str(np.shape(x1)[0])+r'\rm{\␣points}$')
ax.plot(x2,g2,label=r'$'+str(np.shape(x2)[0])+r'\rm{\␣points}$')
ax.plot(x3,g3,label=r'$'+str(np.shape(x3)[0])+r'\rm{\␣points}$')
ax.plot(xa,ya2,label=r'$y_4(x)$')

plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('approxy2.png',bbox_inches='tight')
```

iterate	N-R $\|(x_n, y_n) - (x_{n-1}, y_{n-1})\|_\infty$	$\lvert\phi(t_n) - \phi(t_{n-1})\rvert$
1	0.089999999999999997	0.089999999999999997
2	0.00810000000000001	0.0080999999999999961
3	0.0015236100000000086	0.0015236100000000086
4	0.00030125366943209442	0.00030125366943209442
5	6.011470992249579e-05	6.011470992249579e-05
6	1.2017522165411187e-05	1.2017522165411187e-05
7	2.4032878085944454e-06	2.4032878085944454e-06
8	4.8064889809906752e-07	4.8064889809906752e-07

Table 6.4: Error versus iterates for the method and the test method with the function $y_n(x)$ for y being represented at 101 points from $x = [0, 1]$ evenly spaced. N-R is the Newton-Raphson method. We see that $t \to \phi(t)$ is still basically perfect as an estimate. This program went until the error was less than 10^{-6}.

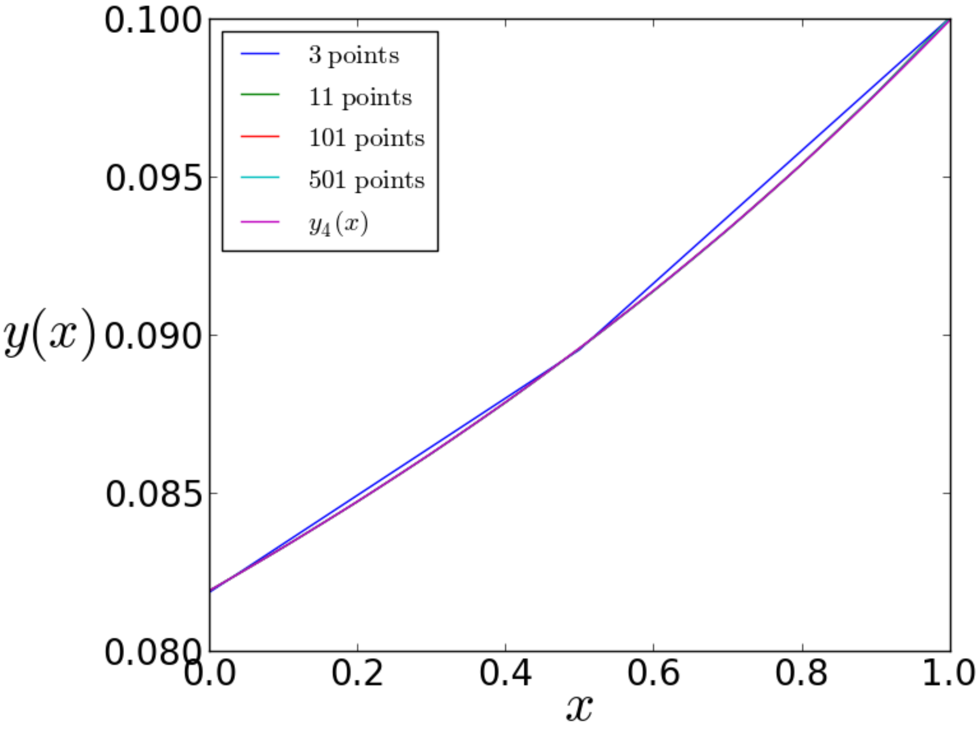

Figure 6.1: We see that as we make the vector $y(x)$ have more and more points so that it approaches the function $y(x)$, we get a better and better result. Note that the analytical $y_4(x)$ is drawn for comparison.

6.5.6 6

Show that, if $f(y) = Ly - g(y)$ where L is a linear operator and $g'(y_0) = 0$ then $S(y) = L^{-1}g(y)$. Show that, if

$$h(x) + \int_0^x h(s)\,\mathrm{d}s = k(x)$$

then

$$h(x) = k(x) - e^{-x}\int_0^x e^s k(s)\,\mathrm{d}s\,.$$

Let $f : \mathcal{C}[0,c] \to \mathcal{C}[0,c]$, $y \to z$ have

$$z(x) = y(x) + \int_0^x y(s)\,\mathrm{d}s - a[y(x)]^2 - b\,.$$

The constants a, b, c are all positive. With the help of the results above, find S, the function iterated to solve the equation $f(y) = 0$, the initial y_0 having $y_0(x) \equiv 0$. Show that

$$S'(y)h \cdot (x) = 2ay(x)h(x) - 2ae^{-x}\int_0^x e^s y(s)h(s)\,\mathrm{d}s\,.$$

Find a suitable function ϕ for the Kantorovich comparison iteration. Compute y_1, y_2, y_3 for $a = 1, b = 0.09$. Do the results suggest any connection between this question and question 5 above?

Solution:

We have

$$f'(y) = L - g'(y)f'(y_0) = L - g'(y_0) = L \tag{6.164}$$

$$[f'(y_0)]^{-1} = L^{-1} \tag{6.165}$$

so $S(y) = L^{-1}g(y)$ as desired.

Then

$$h(x) = k(x) - e^{-x}\int_0^x e^s k(s)\,\mathrm{d}s \tag{6.166}$$

$$h'(x) - k'(x) = e^{-x}\int_0^x e^s k(s)\,\mathrm{d}s - e^{-x}e^x k(x) = -k(x) + e^{-x}\int_0^x e^s k(s)\,\mathrm{d}s = -h(x) \tag{6.167}$$

$$h'(x) + h(x) = k'(x) \tag{6.168}$$

Just as we desired. We need only integrate with $h(0) = k(0) = 0$, which should hold if we keep our boundaries fixed. Alternatively, we can integrate by parts

$$h(x) + \int_0^x h(s)\,\mathrm{d}s = k(x) - e^{-x}\int_0^x \mathrm{d}s\ e^s k(s) + \int_0^x \mathrm{d}s\ \left[k(s) - e^{-s}\int_0^s e^t k(t)\right] \tag{6.169}$$

$$= k(x) + \int_0^x \mathrm{d}s\ k(s) - \int_0^x \mathrm{d}s\ \left[e^{-x}e^s k(s) + e^{-s}\int_0^s \mathrm{d}t\ e^t k(t)\right] \tag{6.170}$$

where

$$\int_0^x \mathrm{d}s\ \left[e^{-x}e^s k(s) + e^{-s}\int_0^s \mathrm{d}t\ e^t k(t)\right] = \int_0^x \mathrm{d}s\ \left[e^{-x}e^s k(s) + \frac{\mathrm{d}}{\mathrm{d}s}\left[-e^{-s}\int_0^s \mathrm{d}t\ e^t k(t)\right] - (-e^s)e^s k(s)\right] \tag{6.171}$$

$$= \int_0^x \mathrm{d}s\ \left[e^{-x}e^s k(s) + k(s)\right] - e^{-x}\int_0^x \mathrm{d}t\ e^t k(t) = \int_0^x \mathrm{d}s\ k(s) + e^{-x}\int_0^x \mathrm{d}s\ e^s k(s) - e^{-x}\int_0^x \mathrm{d}s\ e^s k(s) \tag{6.172}$$

$$= \int_0^x \mathrm{d}s\ k(s) \tag{6.173}$$

Thus,

$$h(x) + \int_0^x h(s)\,\mathrm{d}s = k(x) + \int_0^x \mathrm{d}s\ k(s) - \int_0^x \mathrm{d}s\ \left[e^{-x}e^s k(s) + e^{-s}\int_0^s \mathrm{d}t\ e^t k(t)\right] \tag{6.174}$$

$$= k(x) + \int_0^x \mathrm{d}s\ k(s) - \int_0^x \mathrm{d}s\ k(s) = k(x) \tag{6.175}$$

as desired without us having to imply anything about boundaries.

Now for our f we can write it as

$$f(y) = y(x) + \int_0^x y(s)\,\mathrm{d}s - b - a[y(x)]^2 = L(y) - a[y(x)]^2 - b \tag{6.176}$$

where L is the linear operator

$$Ly = y + \int_0^x y(s)\,\mathrm{d}s \tag{6.177}$$

$$g(y) = a[y(x)]^2 + b \tag{6.178}$$

$$g'(y) = 2a[y(x)] \tag{6.179}$$

We find L^{-1} via the same mechanism as we have previously.

$$k(x) = Lh(x) \tag{6.180}$$

$$L^{-1}k(x) = h(x) \tag{6.181}$$

so we say

$$k(x) = h + \int_0^x h(s)\,\mathrm{d}s \tag{6.182}$$

and so we have

$$L^{-1}k(x) = k(x) - e^{-x}\int_0^x e^s k(s)\,\mathrm{d}s \tag{6.183}$$

Thus, with $y_0 \equiv 0$ we have $g'(y_0) = 0$ so we apply our result and find

$$S(y) = L^{-1}g(y) = a[y(x)]^2 + b - e^{-x}\int_0^x \mathrm{d}s\ e^s\left[a[y(s)]^2 + b\right] \tag{6.184}$$

$$= a[y(x)]^2 + \cancel{b - e^{-x}e^{x}}b + e^{-x}b - e^{-x}\int_0^x \mathrm{d}s\ ae^s[y(s)]^2 \tag{6.185}$$

$$= a[y(x)]^2 + e^{-x}b - e^{-x}\int_0^x \mathrm{d}s\ ae^s[y(s)]^2 \tag{6.186}$$

$$S'(y)h\cdot(x) = 2ay(x)h(x) - 2ae^{-x}\int_0^x \mathrm{d}s\ e^s y(s)h(s) \tag{6.187}$$

as desired.

If we assume $\|y\| < t$ then

$$\|S'(y)\| \leq 2at - 2ae^{-x}\int_0^x e^s t\,\mathrm{d}s \leq 2at - 2ae^{-x}t(e^x - 1) \leq 2ate^{-x} \leq 2at \tag{6.188}$$

Thus we can choose $\phi'(t) = 2at$. We find

$$\|y_1\| = \left\|e^{-x}b\right\| \leq b \tag{6.189}$$

and so if we choose $\phi(0) = b$ then $\|y_1 - y_0\| > \phi(0)$. So

$$\phi(t) = b + at^2 \tag{6.190}$$

We find

$$y_2 = be^{-2x}\left(e^x - ab\left(e^x - 2\right)\right) \tag{6.191}$$

$$y_3 = \frac{1}{3}be^{-4x}\left(-e^{3x}\left(a^3b^3 + 3ab - 3\right) + 16a^3b^3 - 18a^2b^2e^x(ab-1) + 6abe^{2x}(ab-1)^2\right) \tag{6.192}$$

for the given a and b, we find

$$y_1 = 0.09e^{-x} \tag{6.193}$$

$$y_2 = e^{-2x}\left(0.0162\ + 0.0819e^x\right) \tag{6.194}$$

$$y_3 = e^{-4x}\left(e^x\left(e^x\left(0.0134152\ + 0.0818781e^x\right) + 0.00398034\right) + 0.00034992\right) \tag{6.195}$$

$$\|y_1 - y_0\| = 0.09 \tag{6.196}$$

$$\|y_2 - y_1\| = 0.081 \tag{6.197}$$

$$\|y_3 - y_2\| = 0.00152361 \tag{6.198}$$

From our previous problem, we find from $\phi(t)$ that

$$t_0 = 0 \tag{6.199}$$

$$t_1 = at_0^2 + b = b \tag{6.200}$$

$$t_2 = at_1^2 + b = ab^2 + b \tag{6.201}$$

$$t_3 = at_2^2 + b = a(ab^2 + b)^2 + b = a(a^2b^4 + 2ab^3 + b^2) + b = a^3b^4 + 2a^2b^3 + ab^2 + b \tag{6.202}$$

and for $a = 1$ and $b = 0.09$

$$\begin{aligned}
|t_1 - t_0| &= 0.09 && (6.203)\\
|t_2 - t_1| &= |ab^2| = 0.0081 && (6.204)\\
|t_3 - t_2| &= |a^3b^4 + 2a^2b^3| = 0.00152361 && (6.205)
\end{aligned}$$

So we see there is a strong similarity between these two problems. Note that we are choosing $f(y) = 0$ for both problems with for this problem and then the previous problem, respectively,

$$\begin{aligned}
f(y) &= y(x) + \int_0^x y(s)\,\mathrm{d}s - a[y(x)]^2 - b && (6.206)\\
f(y) &= b - y(x) + a\int_1^x [y(s)]^2\,\mathrm{d}s + ax[y(x)]^2 && (6.207)
\end{aligned}$$

So we see there is some similarity, although it is still a bit surprising that the iterations behave so similarly.

We can note that numerically we find a similar result (see Figure 6.2). See previous problem for the code.

iterate	N-R $\|(x_n, y_n) - (x_{n-1}, y_{n-1})\|_\infty$	$\|\phi(t_n) - \phi(t_{n-1})\|$
1	0.089999999999999997	0.089999999999999997
2	0.0080999999999999961	0.0080999999999999961
3	0.0015236100000000086	0.0015236100000000086
4	0.00030125366943209442	0.00030125366943209442
5	6.011470992249579e-05	6.011470992249579e-05
6	1.2017522165411187e-05	1.2017522165411187e-05
7	2.4032878085944454e-06	2.4032878085944454e-06
8	4.8064889809906752e-07	4.8064889809906752e-07

Table 6.5: Error versus iterates for the method and the test method with the function $y_n(x)$ for y being represented at 101 points from $x = [0, 1]$ evenly spaced. N-R is the Newton-Raphson method. We see that $t \to \phi(t)$ is still basically perfect as an estimate. This program went until the error was less than 10^{-6}.

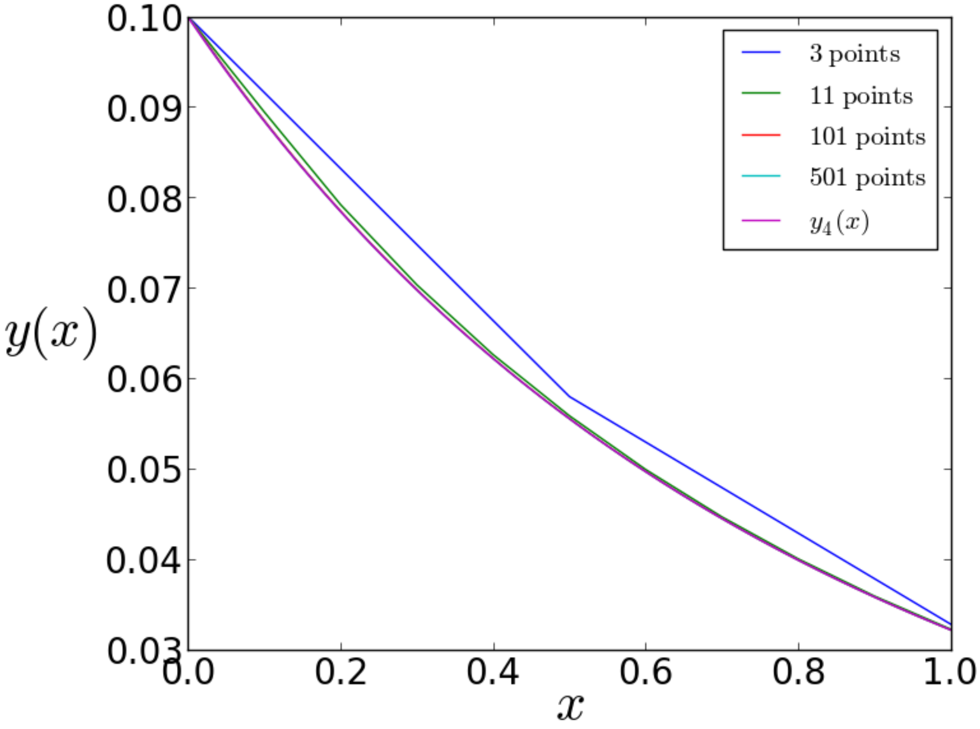

Figure 6.2: We see that as we make the vector $y(x)$ have more and more points so that it approaches the function $y(x)$, we get a better and better result. Note that the analytical $y_4(x)$ is drawn for comparison.

If we use a solver, like Mathematica, we find that the solutions to $f(y) = 0$ for problem 5 and problem 6 scale (for $a = 1$) respectively as (where W is the Lambert W function)

$$y_{\text{P5}}(x) \sim W(-x)/(-x) \tag{6.208}$$

$$y_{\text{P6}}(x) \sim W(-e^{-x-C}) \tag{6.209}$$

for $C > 1.7$ to ensure proper behavior for our domain. Note that for $W = W(z)$ that

$$z(1+W)\frac{\mathrm{d}W}{\mathrm{d}z} = W \tag{6.210}$$

From which by plugging in solutions of the correct form, we can get the solutions for $f(y) = 0$. It is usually easiest to see by taking a deriviative with respect to x of $f(y) = 0$ and matching the differential equation to that above with $y \to W/x, W(e^{-x})$.

Chapter 7

Euclidean Space

7.1 Exercise on minimum distances

The functions $u+mv$, where u and v are fixed functions and the number m varies, form a straight line in $\mathcal{C}[0,1]$. If $u(x) \equiv 1$, $v(x) \equiv x$, find for what values of m the distance, in the metric of $\mathcal{C}[0,1]$, between $u+mv$ and the function 0 takes its minimum value.

Solution:

If we take $u \equiv 1$ and $v \equiv x$ then

$$u + mv = 1 + mx \tag{7.1}$$

Then

$$\|u + mv - 0\| = \|1 + mx\| = \begin{cases} 1 & m \leq 0 \\ 1+m & m \geq 0 \end{cases} \tag{7.2}$$

because $x \in [0,1]$ so that the largest that the function can be is either 1 or $1+m$ away depending on the sign of m.

Thus, $m \leq 0$ are the minimum values.

7.2 Exercises On Euclidean Distance

7.2.1 1

Let $O = (0,0,0)$, $A = (0,1,1)$, $B = (1,0,1)$, $C = (1,1,0)$. Show that any vector in the plane OBC is of the form $(s+t,t,s)$. What condition must s and t satisfy if this vector is to be perpendicular to OB? Find the point, M, of the plane OBC that is nearest to A. Calculate the distances $\|OM\|$ and $\|AM\|$, and show that these agree with the values that could be deduced without the use of coordinates, from the geometry of the regular tetrahedron, $OABC$, with side $\sqrt{2}$.

Solution:

In the plane OBC Let $OB = P$ and $OC = Q$. Then the associated unit vectors are respective $p = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$ and $q = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0)$. Any thing in this plane can be written as $sP + tQ$ with s and t being real numbers. Thus

$$s(1,0,1) + t(1,1,0) = (s,0,s) + (t,t,0) = (s+t,t,s) \tag{7.3}$$

as desired. If this vector (call it R) is to be perpendicular to $OB = P = (1,0,1)$ then

$$R \cdot P = 0 = (s+t) + s = 2s + t = 0 \tag{7.4}$$

That is $2s = -t$. So $R = (-s, -2s, s)$.

Since we have a formula for any point on the plane, we can minimize the length of the line from any point on the plane to A.

$$\begin{align}
MA = T &= (s+t, t-1, s-1) \tag{7.5}\\
d^2 = \|MA\| &= (s+t)^2 + (t-1)^2 + (s-1)^2 \tag{7.6}\\
\frac{\partial d^2}{\partial t} &= 2(s+t) + 2(t-1) = 0 \Rightarrow 2s + 4t - 2 = 0 \Rightarrow s + 2t = 1 \tag{7.7}\\
\frac{\partial d^2}{\partial s} &= 2(s+t) + 2(s-1) = 0 \Rightarrow 2t + 4s - 2 = 0 \Rightarrow 2s + t = 1 \tag{7.8}\\
& \tag{7.9}
\end{align}$$

Thus, we require

$$\begin{align}
2s + 4t - (2s + t) &= 2 - 1 \tag{7.10}\\
3t &= 1 \tag{7.11}\\
t &= \frac{1}{3} \tag{7.12}\\
s = 1 - 2t = 1 - \frac{2}{3} &= \frac{1}{3} \tag{7.13}
\end{align}$$

Thus, $s = t = \frac{1}{3}$ and closest point on the plane is $M = (\frac{2}{3}, \frac{1}{3}, \frac{1}{3})$.

To double check, let's use the two perpendiculars in the plane we have. Thus, first OB and A. IF $OB = tv$ and $A = u$ then the minimum t_0 is given by

$$t_0 = \frac{(0,1,1) \cdot (1,0,1)}{(1,0,1) \cdot (1,0,1)} = \frac{1}{2} \tag{7.14}$$

so the vector is $(1/2, 0, 1/2)$ and for R (choose this vector as $(1,2,-1)$ for simplicity) and A again we find

$$t_0 = \frac{(0,1,1) \cdot (1,2,-1)}{(1,2,-1) \cdot (1,2,-1)} = \frac{2-1}{1+4+1} = \frac{1}{6} \tag{7.15}$$

so the vector is $(1/6, 1/3, -1/6)$ Adding these together yields

$$M = (1/2, 0, 1/2) + (1/6, 1/3, -1/6) = (2/3, 1/3, 1/3) \tag{7.16}$$

as we expected. Then

$$\|OM\| = \sqrt{16+1+1}/3 = \sqrt{18}/3 = \sqrt{2} \tag{7.17}$$

$$\|AM\| = \sqrt{|(-2/3, 2/3, 2/3)|^2} = \sqrt{3(4)/9} = \frac{2}{\sqrt{3}} \tag{7.18}$$

We note that $\|AM\|$ should be the height of a tetrahedron which is given by $\frac{\sqrt{6}}{3}\sqrt{2} = \sqrt{\frac{12}{9}} = \frac{2}{\sqrt{3}}$ as it should. $\|OM\|$ is the length of a side, and so it is reassuring that it is $\sqrt{2}$.

7.3 2

Let $A = (3,4,5)$ and $B = (5,4,3)$. Show that C, the point on the line $x = y = z$ nearest to A is also the point of that line nearest to B. Describe in Euclid's language the figure formed by the points O, A, B, C and their joins.

Solution:
let the point on the line specified by $x = y = z$ be parameterized as (t,t,t), or $t(1,1,1)$. Thus,

$$t_0 = \frac{(1,1,1)\cdot(3,4,5)}{3} = \frac{(1,1,1)\cdot(5,4,3)}{3} = 4 \tag{7.19}$$

so the point $C = (4,4,4)$.

We see that this forms an isosceles triangle.

7.4 3

Let p_1, p_2, and p_3 be three mutually perpendicular vectors of unit length. Let $v = c_1p_1 + c_2p_2 + c_3p_3$. (The numbers c_1, c_2, c_3 are the coordinates of v for axes in the directions of p_1, p_2, and p_3.) Find expressions giving c_1, c_2 and c_3 in terms of scalar products of the four vectors. Verify that $p_1 = (\frac{6}{7}, \frac{3}{7}, \frac{2}{7})$, $p_2 = (\frac{2}{7}, \frac{-6}{7}, \frac{3}{7})$ and $p_3 = (\frac{3}{7}, \frac{-2}{7}, \frac{-6}{7})$ are perpendicular and of unit length. Find the coordinates, c_r, for $v = (5,1,1)$ and show that these are the same as the coordinates of v in the original system. About what line would the original axes of the coordinates have to be rotated to bring them into coincidence with p_1, p_2, and p_3?

Solution:

We have

$$c_1 = \frac{v\cdot p_1}{\sqrt{p_1\cdot p_1}} = v\cdot p_1 \tag{7.20}$$

$$c_2 = \frac{v\cdot p_2}{\sqrt{p_2\cdot p_2}} = v\cdot p_2 \tag{7.21}$$

$$c_3 = \frac{v\cdot p_3}{\sqrt{p_3\cdot p_3}} = v\cdot p_3 \tag{7.22}$$

we have

$$p_1 \cdot p_2 = (6/7, 3/7, 2/7) \cdot (2/7, -6/7, 3/7) = \frac{12 - 18 + 6}{49} = 0 \tag{7.23}$$
$$p_1 \cdot p_3 = (6/7, 3/7, 2/7) \cdot (3/7, -2/7, -6/7) = \frac{18 - 6 - 12}{49} = 0 \tag{7.24}$$
$$p_2 \cdot p_3 = (2/7, -6/7, 3/7) \cdot (3/7, -2/7, -6/7) = \frac{6 + 12 - 18}{49} = 0 \tag{7.25}$$
$$p_i \cdot p_i = \frac{36 + 9 + 4}{49} = 1 \tag{7.26}$$

For $v = (5, 1, 1)$ we find

$$c_1 = p_1 \cdot v = \frac{30 + 3 + 2}{7} = 5 \tag{7.27}$$
$$c_2 = p_2 \cdot v = \frac{10 - 6 + 3}{7} = 1 \tag{7.28}$$
$$c_3 = p_3 \cdot v = \frac{15 - 2 - 6}{7} = 1 \tag{7.29}$$

They would have to be rotated about the line formed by extending the vector $v = (5, 1, 1)$ because this is an invariant in both systems, and so must be conserved through the rotation.

7.5 Exercises on Crude Least Squares

7.5.1 1

Prove that the values of f and g taken for the x-values $-2, -1, 0, 1, 2$ give perpendicular vectors if $f(x) = 1$ and $g(x) = x^2 - 2$. Find the values of a and b that make $a + b(x^2 - 2)$ the best approximation (in the sense of least squares) to $e^{-x^2/8}$ for the x-values listed above. Tabulate these, and compare them with the errors for the approximation $1 - x^2/8$ given by the beginning of the Taylor series for $e^{-x^2/8}$.

Solution:

We have

$$(f, g) = 1(2) + 1(-1) + 1(-2) + 1(-1) + 1(2) = 2 - 1 - 2 - 1 + 2 = 0 \tag{7.30}$$

let $p = f$, $q = g$, and $u = e^{-x^2/8}$. Then we want $ap + bq = u$ to be least squares minimized. Note

$$p = [1, 1, 1, 1, 1] \tag{7.31}$$
$$q = [2, -1, -2, -1, 2] \tag{7.32}$$
$$u = [0.607, 0.882, 1, 0.882, 0.607] \tag{7.33}$$

We have

$$(f, f) = 5 \tag{7.34}$$

$$(g, g) = 2^2 + 1 + 2^2 + 1 + 2^2 = 14 \tag{7.35}$$

$$(f, u) = 3.978 \tag{7.36}$$

$$(g, u) = -1.339 \tag{7.37}$$

$$\tag{7.38}$$

Thus,

$$a = \frac{(p, u)}{(p, p)} \approx \frac{3.978}{5} \approx 0.796 \tag{7.39}$$

$$b = \frac{(q, u)}{(q, q)} \approx \frac{-1.339}{14} \approx -0.0956 \tag{7.40}$$

Thus, collecting terms, (note I used full machine precision, rather than 3 significant figures)

x	-2	-1	0	1	2
$ap + bq$	0.60434372	0.89124468	0.98687833	0.89124468	0.60434372
$1 - \frac{x^2}{8}$	0.5	0.875	1.	0.875	0.5
$e^{-x^2/8}$	0.60653066	0.8824969	1.	0.8824969	0.60653066
Error $ap + bq$	0.00218694	-0.00874778	0.01312167	-0.00874778	0.00218694
Error Taylor Series	0.10653066	0.0074969	0.	0.0074969	0.10653066

Table 7.1: Comparison table of approximations.

7.5.2 2

Verify that 1 and $x^2 - 3.5$ give perpendicular vectors when the x-values $0, 1, 2, 3$ are successively substituted. Find the least squares approximation of $a + b(x^2 - 3.5)$ to $\cos 30x^\circ$ for these values. If you differentiated the least squares approximation to the sine function, would you expect to get the approximation to the cosine function?

Solution:

Then $p = [1, 1, 1, 1]$ and $q = [-3.5, -2.5, 0.5, 5.5]$, with $u = [1, \sqrt{3}/2, 0.5, 0]$ so that

$$(p, q) = -3.5 + -2.5 + 0.5 + 5.5 = -6 + 6 = 0 \tag{7.41}$$

$$(p, p) = 4 \tag{7.42}$$

$$(q, q) = 49 \tag{7.43}$$

$$(u, p) = 2.366 \tag{7.44}$$

$$(u, q) = -5.415 \tag{7.45}$$

so

$$a = \frac{(u, p)}{(p, p)} \approx \frac{2.366}{4} \approx 0.5915 \tag{7.46}$$

$$b = \frac{(u, q)}{(q, q)} \approx \frac{-5.415}{49} \approx -0.1105 \tag{7.47}$$

Note, that I would expect the differentiation of the approximation of the cosine to do a good job of approximating the sine, so long as we are using continuous, differentiable functions, which we are. I would also expect the approximation to be similar. We see that the differentiation of the approximation does not give the same result as this cosine approximation. The coefficients are not wildly off from each other, but neither are they very close.

We see for the differentiation, we'd get

$$a_{\text{diff}} = 0.516061 \leftrightarrow a = 0.5915 \tag{7.48}$$

$$b_{\text{diff}} = -0.061005 \leftrightarrow b = -0.1105 \tag{7.49}$$

Thus, collecting terms, (note I used full machine precision, rather than 3 significant figures)

x	0	1	2	3
$ap + bq$	0.9782966	0.8677851	0.5362506	-0.0163069
$\cos(30x°)$	1.	0.866025404	0.5	0
Error $ap + bq$	0.0217034	-0.0017597	-0.0362506	0.0163069

Table 7.2: Comparison table of approximations.

7.5.3 3

State problems equivalent to 1 and 2 above as questions about points and planes in Euclidean spaces of suitable numbers of dimensions.

Solution:

For problem 1, what point on the planes in the subspace of $\mathcal{E}^5$ spanned by vectors p and q given below, is vector u closest to? Here, $p = [1, 1, 1, 1, 1]$, $q = [2, -1, -2, -1, 2]$ and $u = [0.607, 0.882, 1, 0.882, 0.607]$.

For problem 2, what point on the planes in the subspace of $\mathcal{E}^4$ spanned by vectors p and q given below, is vector u closest to? Here, $p = [1, 1, 1, 1]$, $q = [-3.5, -2.5, 0.5, 5.5]$ and $u = [1, \sqrt{3}/2, 0.5, 0]$.

Chapter 8

Euclidean Space

8.1 Exercises for Least Squares Generalized

8.1.1 1

What in terms of (u, u), (u, v) and (v, v) is the distance of tv from u, when t is chosen to make this a minimum?

Solution:

We use

$$\|tv - u\| = \int \mathrm{d}x\ (tv-u)^2 = \int \mathrm{d}x\ t^2v^2 - \int \mathrm{d}x\ 2tvu + \int \mathrm{d}x\ u^2 = t^2(v,v) - 2t(v,u) + (u,u) \tag{8.1}$$

To minimize we take $\frac{\partial}{\partial t}$ and set to zero to find

$$2t_0(v,v) - 2(v,u) = 0 \tag{8.2}$$

$$t_0 = \frac{(v,u)}{(v,v)} \tag{8.3}$$

$$\tag{8.4}$$

Note that then the minimum may be given by

$$\begin{aligned} \|t_0v - u\|^2 &= \frac{(v,u)^2}{(v,v)} - 2\frac{(v,u)^2}{(v,v)} + (u,u) = (u,u) - \frac{(v,u)^2}{(v,v)} = (u,u) - t_0(v,u) \\ &= \frac{[(u,u)(v,v) - (v,u)^2]}{(v,v)} \end{aligned} \tag{8.5}$$

8.1.2 2

Prove that in $\mathcal{L}_2[0, \pi/2]$, the best approximation to $\sin(x)$ of the form mx is given by $m = 24/\pi^3$. What is the $\mathcal{L}_2$ distance of the approximation from the sin function?

Solution:

We find m via

$$m = \frac{(x, \sin(x))}{(x,x)} = \frac{\int_0^{\pi/2} \mathrm{d}x \; x\sin(x)}{\int_0^{\pi/2} x^2} = \frac{-\int_0^{\pi/2} \mathrm{d}x \; x\frac{\mathrm{d}\cos(x)}{\mathrm{d}x}}{\left[\frac{x^3}{3}\right]_0^{\pi/2}} = \frac{-\left[x\cos(x)|_0^{\pi/2} - \int_0^{\pi/2} \mathrm{d}x \;\; \cos(x)\right]}{\frac{\pi^3}{24}}$$
$$= \frac{0 + \sin(x)|_0^{\pi/2}}{\pi^3/24} = \frac{24}{\pi^3} \tag{8.6}$$

Thus

$$(\sin x, \sin x) = \int_0^{\pi/2} \mathrm{d}x \;\; \sin^2 x = \int_0^{\pi/2} \mathrm{d}x \; \frac{1-\cos(2x)}{2} = \frac{\pi}{4} - \frac{\sin(2x)}{4}|_0^{\pi/2} = \frac{\pi}{4} \tag{8.7}$$

$$\|mx - \sin(x)\|^2 = (\sin(x), \sin(x)) - m(\sin(x), x) = \frac{\pi}{4} - \frac{24}{\pi^3}(1) = \frac{\pi^4 - 96}{4\pi^3} \approx 0.0113613 \tag{8.8}$$

$$\|mx - \sin(x)\| \approx 0.10659 \tag{8.9}$$

8.1.3 3

Show that the algebraic arguments of section 8.3 are justified when (u,v) is defined as $\int_0^{\pi/2} u(x)v(x)\,\mathrm{d}$.

Solution:

It all follows from

$$(u, av + bw) = \int_0^{\pi/2} \mathrm{d}x \;\; u(av + bw) = a\int_0^{\pi/2} \mathrm{d}x \;\; uv + b\int_0^{\pi/2} \mathrm{d}x \;\; uw = a(u,v) + b(u,w) \tag{8.10}$$

8.1.4 4

Let $f(x) \equiv a + bx^2$, $g(x) \equiv cx + kx^3$. Find the angle between the vectors that represent f and g in $\mathcal{L}_2[-1,1]$.

Solution:

We use the angle between two vectors is denoted by

$$\cos\theta = \frac{(f,g)}{\|f\|\,\|g\|} \tag{8.11}$$

Thus,

$$(f,g) = \int_{-1}^{1} \mathrm{d}x \;\; \left[(a + bx^2)x(c + kx^2)\right] = \int_{-1}^{1} \mathrm{d}x \;\; \left[acx + akx^3 + bcx^3 + bkx^5\right]$$
$$= \int_{-1}^{1} \mathrm{d}x \;\; \left[bkx^5 + (bc + ak)x^3 + acx\right] = \left[\frac{bkx^6}{6} + (bc + ak)\frac{x^4}{4} + ac\frac{x^2}{2}\right]_{-1}^{1} = 0 \tag{8.12}$$

Thus, $\cos\theta = 0$ implies $\theta = \frac{\pi}{2}$, thus they are orthogonal.

8.1.5 5

Let $f(x) \equiv 0$, $g(x) \equiv 2$, $h(x) \equiv 1 + 3x$. Find the distances $d(f, g)$, $d(f, h)$, $d(g, h)$ in $\mathcal{L}_2[-1, 1]$. What geometrical figure is formed by the points representing f, g, h in this space? What value does this geometrical representation suggest for the scalar product (g, h)? Test the validity of this suggestion by calculating (g, h) directly as an integral.

Solution:

$$d(f, g) = d(0, 2) = \left[\int_{-1}^{1} \mathrm{d}x \; 2^2\right]^2 = 4x|_{-1}^{1} = 8 \tag{8.13}$$

$$\sqrt{d(f, h)} = \sqrt{d(0, 1 + 3x)} = \int_{-1}^{1} \mathrm{d}x \; (1 + 3x)^2 = \int_{-1}^{1} \mathrm{d}x \; \left[1 + 6x + 9x^2\right] = 2 + 3x^3|_{-1}^{1} = 2 + 6 = 8 \tag{8.14}$$

$$\sqrt{d(g, h)} = \sqrt{d(2, 1 + 3x)} = \int_{-1}^{1} \mathrm{d}x \; (3x - 1)^2 = \int_{-1}^{1} \mathrm{d}x \; \left[1 - 6x + 9x^2\right] 2 + 3x^3|_{-1}^{1} = 2 + 6 = 8 \tag{8.15}$$

This suggests we have an equilateral triangle so $(f, g) = \|f\| \, \|g\| \cos\theta = \sqrt{8}\sqrt{8}\cos(60°) = 4$.

We have

$$(g, h) = \int_{-1}^{1} \mathrm{d}x \; [2 + 6x] = 4 + 0 \tag{8.16}$$

as we thought.

8.1.6 6

Let $f_n(x) \equiv x^n$. Find $\|f_n\|$ (a) if f_n is regarded as an element of $\mathcal{L}_2[0, 1]$, (b) if f_n is regarded as an element of $\mathcal{C}[0, 1]$. In each case, discuss whether f_n tends to a limit as $n \to \infty$ and state the limit if it does.

Solution:

For (a), we have

$$\|f_n\|^2 = \int_{0}^{1} \mathrm{d}x \; x^{2n} = \left.\frac{x^{2n+1}}{2n + 1}\right|_{0}^{1} = \frac{1}{2n + 1} \tag{8.17}$$

As $n \to \infty$, then in $\mathcal{L}_2[0, 1]$ we see that $\lim_{n\to\infty} \|f_n\| = 0$ since the denominator grows without bound.

For (b) we note that x^n will reach its maximum distance from the origin at $x = 1$, which will of course yield $\|f_n\| = 1$ for all n. Thus, there is a limit, and it is $\lim_{n\to\infty} \|f_n\| = 1$.

We note we get very different answer depending on the norm used.

8.1.7 7

Do as in question 6, with $f_n(x) \equiv n^{1/4}e^{-nx}$.

Solution:

For (a) we have

$$\|f_n\|^2 = \int_0^1 \mathrm{d}x \; n^{1/2}e^{-2nx} = n^{1/2}\frac{e^{-2nx}}{-2n}|_0^1 = \frac{1-e^{-2n}}{2\sqrt{n}} \tag{8.18}$$

Clearly, this approaches zero as $n \to \infty$ since $e^{-2n}/\sqrt{n}$ and $\frac{1}{\sqrt{n}}$ both approach zero rapidly as $n \to \infty$.

For (b), we note that the maximum value of e^{-nx} in this interval is 1, so that $\|f_n\| = n^{1/4}$. Note that as $n \to \infty$ that this blows up, so that there is no limit as $n \to \infty$.

8.1.8 8

Let $f(x) \equiv a$ and $g(x) \equiv \cos x$. Find what value of a makes the distance of f from g in $\mathcal{L}_2[-\pi/2, \pi/2]$ a minimum, and determine this minimum distance. For this value of a, what is the distance of f from g in $\mathcal{C}[-\pi/2, \pi/2]$? What value of a would make f nearest to g in $\mathcal{C}[-\pi/2, \pi/2]$, and how far apart would they then be in that space?

Solution:

We use

$$a = \frac{(1, \cos(x))}{(1,1)} = \frac{\int_{-\pi/2}^{\pi/2} \mathrm{d}x \;\; \cos(x)}{\int_{-\pi/2}^{\pi/2} \mathrm{d}x \; 1} = \frac{\sin(x)|_{-\pi/2}^{\pi/2}}{\pi} = \frac{2}{\pi} \tag{8.19}$$

The distance is given by

$$(\cos x, \cos x) = \int_{-\pi/2}^{\pi/2} \mathrm{d}x \;\; \cos^2(x) = \int_{-\pi/2}^{\pi/2} \mathrm{d}x \; \frac{1+\cos(2x)}{2} = \frac{\pi}{2} \tag{8.20}$$

$$\|a - \cos(x)\|^2 = \frac{\pi}{2} - \frac{2}{\pi}2 = \frac{\pi^2 - 8}{2\pi} \approx 0.297557 \tag{8.21}$$

$$\|a - \cos(x)\| \approx 0.545488 \tag{8.22}$$

For $\mathcal{C}[-\pi/2, \pi/2]$ we note that $\left\|\frac{2}{\pi} - \cos(x)\right\| = \sup|\frac{2}{\pi} - \cos(x)| = \frac{2}{\pi} \approx 0.63622$. To make a a minimum in $\mathcal{C}$ we need to choose $a = 0.5$ so that $\|a - \cos(x)\| = 0.5$. This is because $\cos(x)$ varies between 0 and 1 with $x \in [-\pi/2, \pi/2]$ so that choosing the halfway point minimizes the norm.

8.1.9 9

Let $f(x) \equiv ax$ and $g(x) = \sin x$. Find what value of a makes $d(f, g)$ in $\mathcal{L}_2[0, \pi/3]$ a minimum, and find this minimum distance. A well-known crude approximation to $\sin x^\circ$ is $x/60$, which corresponds to $a = 3/\pi$ for radian measure. The value of a that makes f closest to g in $\mathcal{C}[0, \pi/3]$

is 0.869 647 167 . Make tables showing the errors of approximation $ax - \sin x$ for the three values of a referred to in the paragraphs above, and compare them.

Solution:

We use

$$\begin{aligned} a = \frac{(x, \sin(x))}{(x,x)} &= \frac{\int_0^{\pi/3} \mathrm{d}x \ x \sin(x)}{\int_0^{\pi/3} \mathrm{d}x \ x^2} = \frac{-x\cos(x)|_0^{\pi/3} + \int_0^{\pi/3} \mathrm{d}x \ \cos(x)}{\frac{x^3}{3}|_0^{\pi/3}} \\ &= \frac{-\pi/6 + \sin(x)|_0^{\pi/3}}{\pi^3/3^4} = \frac{-\pi/6 + \frac{\sqrt{3}}{2}}{\pi^3/81} \\ &\approx 0.894\,546\,519\,372\,964\,9 \end{aligned} \tag{8.23}$$

Thus, the minimum distance is

$$\int_0^{\pi/3} \mathrm{d}x \ \sin^2(x) = \int_0^{\pi/3} \frac{1-\cos(2x)}{2} = \frac{\pi}{6} - \frac{\sin(2x)|_0^{\pi/3}}{4} = \frac{\pi}{6} - \frac{\sin(2\pi/3)}{4} = \frac{\pi}{6} - \frac{\sqrt{3}}{8} \tag{8.24}$$

$$\begin{aligned} \|a - \sin(x)\|^2 &= \frac{\pi}{6} - \frac{\sqrt{3}}{8} - \frac{(-\pi/6 + \frac{\sqrt{3}}{2})^2}{\pi^3/81} \\ &= -\frac{9\left(3\sqrt{3}-\pi\right)^2}{4\pi^3} - \frac{\sqrt{3}}{8} + \frac{\pi}{6} \approx 0.000\,775\,876 \end{aligned} \tag{8.25}$$

Now for the table

a	$x = 0°$	10°	20°	30°	40°	50°	60°
a_1	0.	-0.01752036	-0.0297645	-0.03161654	-0.01827633	0.01459466	0.07074152
a_2	0.	-0.00698151100	-0.00868680999	0.	0.0238790570	0.0672888902	0.133974596
a_3	0.	-0.02186611	-0.03845602	-0.04465381	-0.03565935	-0.00713412	0.04466698

Table 8.1: Error in the approximation ax for various a for function $\sin(x)$ in $[0, \pi/3]$. $a_1 = 0.894\,546\,519\,372\,964\,9$, $a_2 = 3/\pi$, and $a_3 = 0.869\,647\,167$.

We see that both have errors at different portions of the domain. One might say that a_1 and a_3 do better over the entire range, but $a_2 = 3/\pi$ does an impressive job near the center. We can plot these in Figure 8.1.

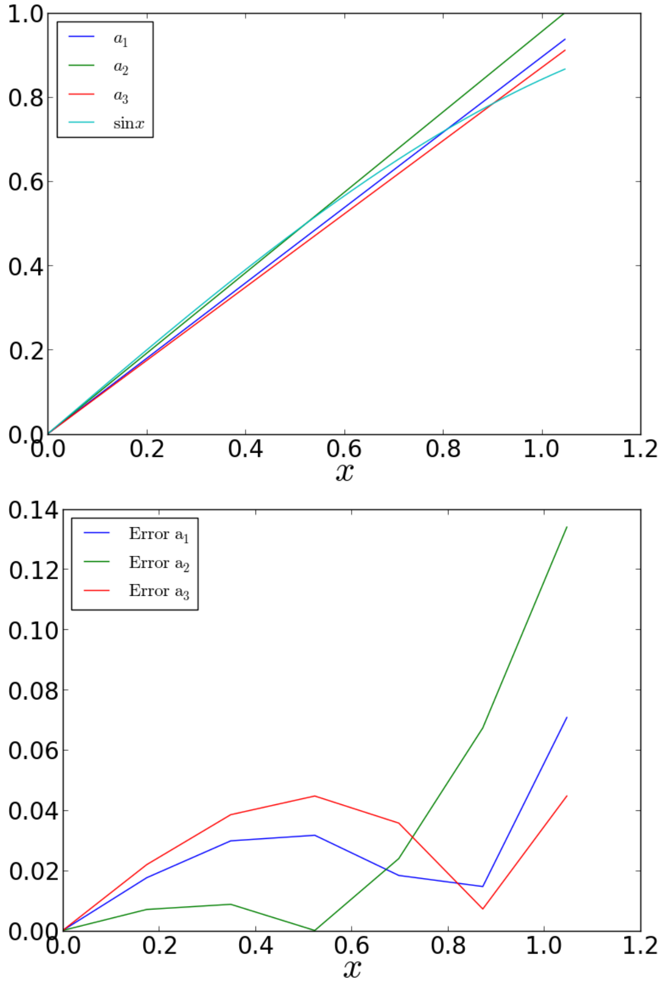

Figure 8.1: A plot of the approximations and the errors.

chapter8/varas.py

```
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

#Test various a's for a x as an approximation to sin(x) in [0,pi/3]
a1=0.8945465193729649
a2=3/np.pi
a3=0.869647167

x=np.linspace(0,np.pi/3,501)

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,a1*x,label=r'$a_1$')
ax.plot(x,a2*x,label=r'$a_2$')
ax.plot(x,a3*x,label=r'$a_3$')
ax.plot(x,np.sin(x),label=r'$\sin x$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
```

```
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('varas.png',bbox_inches='tight')

plt.clf()

x1=np.linspace(0,np.pi/3,7)

y1=a1*x1-np.sin(x1)
y2=a2*x1-np.sin(x1)
y3=a3*x1-np.sin(x1)

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x1,np.abs(y1),label=r'$\rm{Error\ }a_1$')
ax.plot(x1,np.abs(y2),label=r'$\rm{Error\ }a_2$')
ax.plot(x1,np.abs(y3),label=r'$\rm{Error\ }a_3$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('varasError.png',bbox_inches='tight')

print y1
print y2
print y3
```

8.2 Exercises on Chebyshev/Fourier Series

8.2.1 1

Show that for $f(x) \equiv \pi^2 x - x^3$ with the interval $[0, \pi]$, equation

$$c_r = \frac{2}{\pi}(f, p_r) = \frac{2}{\pi}\int_0^\pi \mathrm{d}x \; f(x)\sin(rx)$$

gives the Fourier coefficients, $c_r = 12(-1)^{r+1}/r^3$. Use these values to calculate

$$g(x) = \sum_{r=1}^{10} c_r \sin(rx)$$

for $x = n\pi/4$, with n taking whole number values from $[-4, 12]$ inclusive. Sketch a rough graph of g for $-\pi \leq x \leq 3\pi$, and compare this with the graph of f.

Solution:

We have

$$\begin{aligned}\int_0^\pi \mathrm{d}x\; x\sin(rx) &= -\frac{1}{r}\int_0^\pi \mathrm{d}x\; x\frac{\mathrm{d}\cos(rx)}{\mathrm{d}x} = -\frac{1}{r}\left[\int_0^\pi \mathrm{d}x\;\frac{\mathrm{d}}{\mathrm{d}x}\left[x\cos(rx)\right] - \cos(rx)\frac{\mathrm{d}x}{\mathrm{d}x}\right]\\ &= \frac{-1}{r}\left[x\cos(rx)|_0^\pi - \int_0^\pi \mathrm{d}x\;\;\cos(rx)\right] = \frac{-1}{r}\left[\pi\cos(r\pi) - \cancel{\frac{\sin(rx)}{r}}|_0^\pi\right]\\ &= \frac{-1}{r}\pi(-1)^r = \frac{(-1)^{r+1}\pi}{r}\end{aligned}\tag{8.26}$$

$$\begin{aligned}\int_0^\pi \mathrm{d}x\; x^3\sin(rx) &= -\frac{1}{r}\int_0^\pi \mathrm{d}x\; x^3\frac{\mathrm{d}\cos(rx)}{\mathrm{d}x} = -\frac{1}{r}\left[\int_0^\pi \mathrm{d}x\;\frac{\mathrm{d}}{\mathrm{d}x}\left[x^3\cos(rx)\right] - \cos(rx)\frac{\mathrm{d}x^3}{\mathrm{d}x}\right]\\ &= \frac{-1}{r}\left[x^3\cos(rx)|_0^\pi - \int_0^\pi \mathrm{d}x\;\;\cos(rx)\right]\\ &= \frac{-1}{r}\left[\pi^3\cos(r\pi) - 3\int_0^\pi \mathrm{d}x\; x^2\cos(rx)\right]\\ &= \frac{-1}{r}\left[\pi^3\cos(r\pi) - \frac{3}{r}\int_0^\pi \mathrm{d}x\; x^2\frac{\mathrm{d}\sin(rx)}{\mathrm{d}x}\right]\\ &= \frac{-1}{r}\left[\pi^3(-1)^r - \frac{3}{r}\int_0^\pi \mathrm{d}x\;\;\left\{\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2\sin(rx)\right] - \sin(rx)\frac{\mathrm{d}x^2}{\mathrm{d}x}\right\}\right]\\ &= \frac{-1}{r}\left[\pi^3(-1)^r - \cancel{\frac{3}{r}x^2\sin(rx)|_0^\pi} + \frac{3}{r}2\int_0^\pi \mathrm{d}x\; x\sin(rx)\right]\\ &= \frac{-1}{r}\left[\pi^3(-1)^r + \frac{6}{r}\frac{(-1)^{r+1}\pi}{r}\right]\\ &= \frac{\pi^3(-1)^{r+1}}{r} + \frac{6(-1)^{r+2}\pi}{r^3}\end{aligned}\tag{8.27}$$

Thus,

$$\int_0^\pi \mathrm{d}x\;(\pi^2x - x^3)\sin(rx) = \cancel{\pi^2\frac{(-1)^{r+1}\pi}{r}} - \cancel{\frac{\pi^3(-1)^{r+1}}{r}} - \frac{6(-1)^{r+2}\pi}{r^3} = \frac{6(-1)^{r+3}\pi}{r^3}\tag{8.28}$$

Thus, (using $(-1)^{r+3} = (-1)^2(-1)^{r+1} = (-1)^{r+1}$)

$$\frac{2}{\pi}\int_0^\pi \mathrm{d}x\;\;(\pi^2x - x^3)\sin(rx) = \frac{2}{\pi}\frac{6(-1)^{r+1}\pi}{r^3} = \frac{12(-1)^{r+1}}{r^3}\tag{8.29}$$

Thus, we find Figure 8.2 and table 8.2.

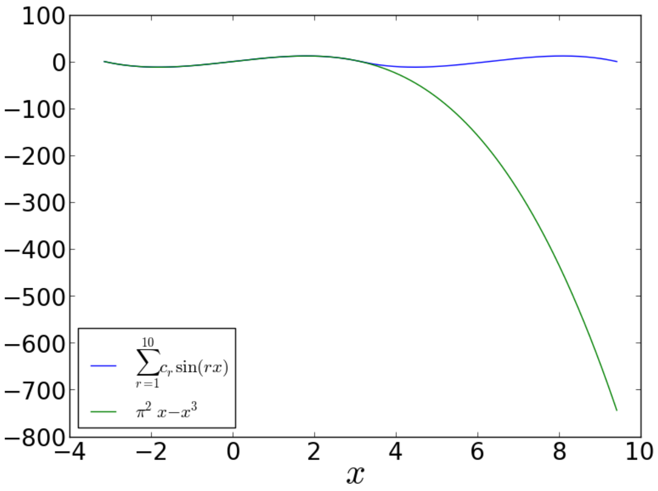

Figure 8.2: A plot of the approximation.

x	$\pi^2x - x^3$	$\sum_{r=1}^{10} c_r \sin(rx)$
-4	10.64092289	24.5215824
-3	-2.57857812	-2.6088132
-2	-11.73275439	-11.7392088
-1	-8.87490147	-8.8696044
0	0.	0.
1	8.87490147	8.8696044
2	11.73275439	11.7392088
3	2.57857812	2.6088132
4	-10.64092289	-24.5215824
5	-10.54632377	-75.65197799
6	-2.77145435	-156.78237359
7	6.70111714	-273.91276919
8	11.88878887	-433.04316479
9	6.77445731	-640.17356039
10	-8.41359761	-901.30395599
11	-11.62179874	-1222.43435159
12	-5.40972637	-1609.56474719

Table 8.2: The values for the approximation and the function for the given values of x.

chapter8/fourierone.py

```
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

#Test various a's for a x as an approximation to sin(x) in [0,pi/3]
x1=np.linspace(-4,12,17)
x=np.linspace(-np.pi,3*np.pi,501)

f1=np.pi**2*x1-x1**3
f=np.pi**2*x-x**3

```

```
def cr(r):
  return 12.*(-1)**(r+1)/(r)**3

maxfour=10
s1=x1*0
s=x*0
for i in range(len(s)):
  for j in range(maxfour):
    s[i]=s[i]+cr(j+1)*np.sin((j+1)*x[i])

for i in range(len(s1)):
  for j in range(maxfour):
    s1[i]=s1[i]+cr(j+1)*np.sin((j+1)*x1[i])

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,s,label=r'$\sum_{r=1}^{10}c_r\sin(rx)$')
ax.plot(x,f,label=r'$\pi^2x-x^3$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('cubicfunction.png',bbox_inches='tight')

print s1
print f1
```

8.2.2 2

(The sawtooth function) Let $f(x) = x$ for $0 \leq x \leq \pi/2$ and $f(x) = \pi - x$ for $\pi/2 < x \leq \pi$. Show that the corresponding Fourier series is

$$g(x) = (4/\pi)\left[\sin(x) - \frac{1}{9}\sin(3x) + \frac{1}{25}\sin(5x) + \cdots\right]$$

with the squares of the odd numbers appearing as denominators. Does the series $g(x)$ converge (a) absolutely, (b) uniformly? Investigate the graph of g, with x taking all real values. Sketch the graph of the derivative f' in $[0, \pi]$. Would you expect a Fourier series for f' to converge uniformly?

Solution:

Let's first construct the series for $0 \leq x \leq \pi/2$. Then

$$\begin{aligned}\int_0^{\pi/2} \mathrm{d}x\ x\sin(rx) &= -\frac{1}{r}\int_0^{\pi/2}\mathrm{d}x\ x\frac{\mathrm{d}\cos(rx)}{\mathrm{d}x} = -\frac{1}{r}\left[\int_0^{\pi/2}\mathrm{d}x\ \frac{\mathrm{d}}{\mathrm{d}x}\left[x\cos(rx)\right] - \cos(rx)\frac{\mathrm{d}x}{\mathrm{d}x}\right] \\ &= \frac{-1}{r}\left[x\cos(rx)|_0^{\pi/2} - \int_0^{\pi}\mathrm{d}x\ \cos(rx)\right] = \frac{-1}{r}\left[\frac{\pi}{2}\cos(r\pi/2) - \frac{\sin(rx)}{r}|_0^{\pi/2}\right] \\ &= \frac{\sin(r\pi/2)}{r^2} - \frac{\pi}{2}\frac{\cos(r\pi/2)}{r}\end{aligned} \tag{8.30}$$

Note that

$$\sin(r\frac{\pi}{2}) = \begin{cases} 0 & r \equiv 0 \mod 4 \\ 1 & r \equiv 1 \mod 4 \\ 0 & r \equiv 2 \mod 4 \\ -1 & r \equiv 3 \mod 4 \end{cases} = \frac{(-1)^r - 1}{2} i^{r+1} \tag{8.31}$$

$$\cos(r\frac{\pi}{2}) = \begin{cases} 1 & r \equiv 0 \mod 4 \\ 0 & r \equiv 1 \mod 4 \\ -1 & r \equiv 2 \mod 4 \\ 0 & r \equiv 3 \mod 4 \end{cases} = \frac{(-1)^r + 1}{2} i^{r} \tag{8.32}$$

For $\pi/2 < x \leq \pi$ we use $u = x - \pi/2$ so that $f(u) = \pi - (u + \pi/2) = \pi/2 - u$. Thus,

$$\int_{\pi/2}^{\pi} \mathrm{d}x \ \pi \sin(rx) = \frac{-\pi\cos(rx)}{r}|_{\pi/2}^{\pi} = \frac{\pi}{r}\left(\cos(r\pi/2) - \cos(r\pi)\right) \tag{8.33}$$

$$\begin{aligned} \int_{\pi/2}^{\pi} \mathrm{d}x \ x\sin(rx) &= -\frac{1}{r}\int_{\pi/2}^{\pi} \mathrm{d}x \ x\frac{\mathrm{d}\cos(rx)}{\mathrm{d}x} = -\frac{1}{r}\left[\int_{\pi/2}^{\pi} \mathrm{d}x \ \frac{\mathrm{d}}{\mathrm{d}x}\left[x\cos(rx)\right] - \cos(rx)\frac{\mathrm{d}x}{\mathrm{d}x}\right] \\ &= \frac{-1}{r}\left[x\cos(rx)|_{\pi/2}^{\pi} - \int_{\pi/2}^{\pi} \mathrm{d}x \ \ \cos(rx)\right] = \frac{-1}{r}\left[\frac{\pi}{2}\left(2\cos(r\pi) - \cos(r\pi/2)\right) - \frac{\sin(rx)}{r}|_{\pi/2}^{\pi}\right] \\ &= \frac{-\sin(r\pi/2)}{r^2} - \frac{\pi}{2r}\left(2\cos(r\pi) - \cos(r\pi/2)\right) \end{aligned} \tag{8.34}$$

So in totality,

$$\int_{\pi/2}^{\pi} \mathrm{d}x \ \ [\pi - x]\sin(rx) = \frac{\pi}{r}\left(\cos(r\pi/2) - \cos(r\pi)\right) + \frac{\sin(r\pi/2)}{r^2} + \frac{\pi}{r}\left(\cos(r\pi) - \frac{1}{2}\cos(r\pi/2)\right) \tag{8.35}$$

$$= \frac{\pi}{2}\frac{\cos(r\pi/2)}{r} + \frac{\sin(r\pi/2)}{r^2} \tag{8.36}$$

Thus, c_r is given by

$$\begin{aligned} c_r &= \frac{2}{\pi}\int_0^{\pi} \mathrm{d}x \ \ f(x)\sin(rx) = \frac{2}{\pi}\frac{\sin(r\pi/2)}{r^2} - \frac{\cos(r\pi/2)}{r} + \frac{\cos(r\pi/2)}{r} + \frac{2}{\pi}\frac{\sin(r\pi/2)}{r^2} \\ &= \frac{4\sin(r\pi/2)}{\pi r^2} \\ &= \begin{cases} \frac{4}{\pi r^2} & r \equiv 1 \mod 4 \\ -\frac{4}{\pi r^2} & r \equiv 3 \mod 4 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8.37}$$

So we recover

$$g(x) = \frac{4}{\pi}\left[\sin(x) - \frac{1}{9}\sin(3x) + \frac{1}{25}\sin(5x) + \cdots\right] \tag{8.38}$$

Note that

$$g(x) \leq |g(x)| \leq \frac{4}{\pi} \sum_{r=1}^{\infty} \frac{1}{(2r-1)^2} \leq \frac{4}{\pi} \sum_{r=1}^{\infty} \frac{1}{r^2} = \frac{4}{\pi}\frac{\pi^2}{6} \tag{8.39}$$

therefore $g(x)$ does converge absolutely, and via Weierstrass, uniformly as well.

The derivative is shown in Figure 8.3. I wouldn't expect uniform convergence. This is because the function is discontinuous, and so we don't expect good convergence properties. We'd find

$$\begin{aligned}
\int_0^{\pi/2} \mathrm{d}x\ \sin(rx) &= \frac{\cos(r0) - \cos(r\pi/2)}{r} = \frac{1 - \cos(r\pi/2)}{r} && (8.40)\\
-\int_{\pi/2}^{\pi} \mathrm{d}x\ \sin(rx) &= \frac{\cos(r\pi) - \cos(r\pi/2)}{r} && (8.41)\\
c_r &= \frac{2}{\pi}\left[\frac{1 - \cos(r\pi/2)}{r} + \frac{\cos(r\pi) - \cos(r\pi/2)}{r}\right] && (8.42)\\
&= \frac{2}{\pi}\left[\frac{1 - 2\cos(r\pi/2) + \cos(r\pi)}{r}\right] && (8.43)\\
&= \begin{cases} 0 & r \equiv 0 \mod 4 \\ 0 & r \equiv 1 \mod 4 \\ \frac{8}{\pi r} & r \equiv 2 \mod 4 \\ 0 & r \equiv 3 \mod 4 \end{cases} && (8.44)
\end{aligned}$$

Or

$$h(x) = \frac{8}{\pi}\left[\frac{1}{2}\sin(2x) + \frac{1}{6}\sin(6x) + \frac{1}{10}\sin(10x) + \cdots\right] \tag{8.45}$$

For the worst case, this is

$$\frac{\pi}{8}|h(x)| \leq \sum_{r=0}^{\infty} \frac{1}{(4r+2)} \leq \frac{1}{8}\sum_{r=1}^{\infty} \frac{1}{r} \tag{8.46}$$

the harmonic series which famously fails to converge.

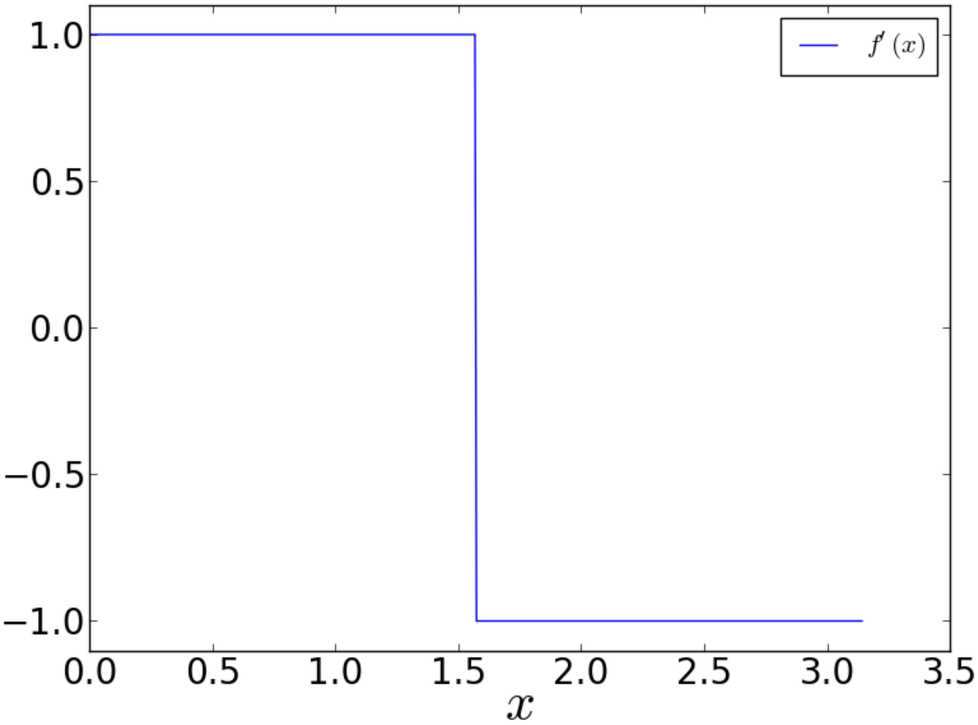

Figure 8.3: A plot of the derivative of $f'(x)$.

chapter8/plotsawp.py

```
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

#Test various a's for a x as an approximation to sin(x) in [0,pi/3]
x=np.linspace(0,np.pi,500)
f=0*x

f[:250]=1
f[250:]=-1

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,f,label=r'$f^\prime(x)$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
ax.set_ylim(-1.1,1.1)
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('sawp.png',bbox_inches='tight')
```

8.2.3 3

Show that the Fourier series

$$\frac{4}{\pi}\sum_{r=0}^{\infty}\left[\frac{\cos[(2r+1)\theta]}{(2r+1)^2}\right]$$

arises from $f(\theta)=\frac{\pi}{2}-\theta$ on $[0,\pi]$. By putting $x=\cos\theta$ obtain the series of Chebyshev polynomials that corresponds to $\sin^{-1}x$. Tabulate the errors produced when $\sin^{-1}x$ is approximated (a) by the terms up to x^5 in its Taylor series, (b) by the Chebyshev series up to the term in $T_5(x)$. Observe the contrast in behavior between these.

Solution:

First, we find

$$\int_0^{\pi}\mathrm{d}\theta\ \cos[(2r+1)\theta]=\frac{\sin[(2r+1)\pi]-\sin[(2r+1)0]}{r}=0 \tag{8.47}$$

$$\int_0^{\pi}\mathrm{d}\theta\ \theta\cos[(2r+1)\theta]=\frac{1}{2r+1}\int_0^{\pi}\mathrm{d}\theta\ \theta\frac{\mathrm{d}\sin(2r+1)\theta}{\mathrm{d}\theta} \tag{8.48}$$

$$=\cancel{\frac{\theta\sin[(2r+1)\theta]}{2r+1}}|_0^{\pi}-\frac{1}{2r+1}\int_0^{\pi}\sin[(2r+1)\theta]=\frac{-1}{(2r+1)^2}\cos[(2r+1)\theta]|_0^{\pi} \tag{8.49}$$

$$=\frac{1-\overbrace{\cos[(2r+1)\theta]}^{-1}}{(2r+1)^2}=\frac{2}{(2r+1)^2} \tag{8.50}$$

Thus, yielding the series

$$g(\theta)=\sum_{r=0}^{\infty}\frac{2}{\pi}\frac{2}{(2r+1)^2}\cos[(2r+1)\theta]=\frac{4}{\pi}\sum_{r=0}^{\infty}\frac{\cos[(2r+1)\theta]}{(2r+1)^2} \tag{8.51}$$

as desired.

We note that for $x=\cos\theta$, $\sqrt{1-x^2}=\sin\theta$, $\sin^{-1}(x)=\pi/2-\theta$ due to the triangle drawn this way.

We remember that

$$T_n(x)=\cos(n\theta) \tag{8.52}$$

So that the series becomes

$$g(x)=\frac{4}{\pi}\sum_{r=0}^{\infty}\frac{T_{2r+1}(x)}{(2r+1)^2} \tag{8.53}$$

For (a), we note that the fastest way to find the coefficients is through Cauchy's theorem. Then,

$$\frac{\mathrm{d}^n f(z)}{\mathrm{d}z^n}=\frac{n!}{2\pi i}\oint_C \mathrm{d}\zeta\ \frac{f(\zeta)}{(\zeta-z)^{n+1}} \tag{8.54}$$

To calculate the series, let's use that we can integrate term by term. So we have

$$\frac{\mathrm{d}\sin^{-1}(x)}{\mathrm{d}x} = \frac{1}{\sqrt{1-x^2}} \tag{8.55}$$

Then

$$\begin{aligned}
\frac{1}{\sqrt{1-x^2}} &\approx 1 + \frac{1}{2}x^2 + \frac{3}{8}x^4 + \cdots && (8.56)\\
\sin^{-1}(x) &\approx \int^x \mathrm{d}x \,\left[1 + \frac{1}{2}x^2 + \frac{3}{8}x^4 + \cdots\right] && (8.57)\\
&\approx x + \frac{x^3}{6} + \frac{3}{40}x^5 + \cdots && (8.58)
\end{aligned}$$

Thus,

$$\begin{aligned}
\sin^{-1}(x) &\approx x + \frac{x^3}{6} + \frac{3x^5}{40} && (8.59)\\
\sin^{-1}(x) &\approx \frac{4}{\pi}\left[T_0 + \frac{T_3}{9} + \frac{T_5}{25}\right] && (8.60)
\end{aligned}$$

Let's divide $[0, \pi]$ into eleven divisions. Thus, table 8.3 and Figure 8.4.

x	Power Series	Chebyshev Series	$\sin^{-1} x$
-1.00000000e+00	9.16666667e-02	-1.41471061e+00	-1.57079633e+00
-9.51056516e-01	7.79989459e-02	-1.29407737e+00	-1.25663706e+00
-8.09016994e-01	4.54043097e-02	-9.86355468e-01	-9.42477796e-01
-5.87785252e-01	1.46320165e-02	-6.13844453e-01	-6.28318531e-01
-3.09016994e-01	1.30843538e-03	-2.79000165e-01	-3.14159265e-01
6.12323400e-17	2.34299938e-66	5.19756244e-17	6.12323400e-17
3.09016994e-01	1.73110699e-03	2.79000165e-01	3.14159265e-01
5.87785252e-01	2.51561097e-02	6.13844453e-01	6.28318531e-01
8.09016994e-01	9.73894813e-02	9.86355468e-01	9.42477796e-01
9.51056516e-01	1.94712928e-01	1.29407737e+00	1.25663706e+00
1.00000000e+00	2.41666667e-01	1.41471061e+00	1.57079633e+00

x	Power Series	Chebyshev Series	$\sin^{-1} x$
-1	0.09166667	-1.41471061	-1.57079633
-0.8	0.04369067	-0.96879382	-0.92729522
-0.6	0.015768	-0.63152681	-0.64350111
-0.4	0.00349867	-0.37574714	-0.41151685
-0.2	0.00024267	-0.17429235	-0.20135792
0.	0.	0.	0.
0.2	0.00029067	0.17429235	0.20135792
0.4	0.00503467	0.37574714	0.41151685
0.6	0.027432	0.63152681	0.64350111
0.8	0.09284267	0.96879382	0.92729522
1.	0.24166667	1.41471061	1.57079633

Table 8.3: The values for the approximation and the function for the given values of x. I used values for x from $x = \cos\theta$ in the first table.

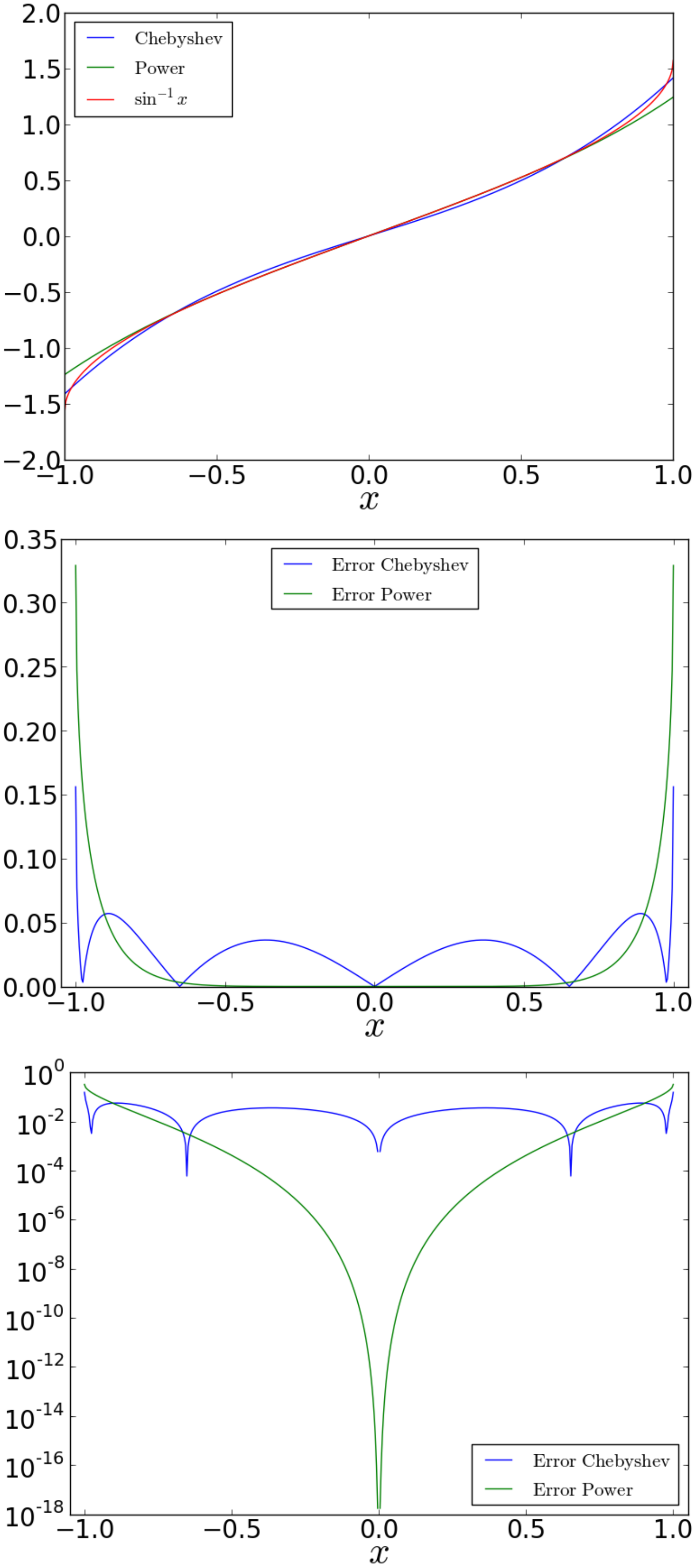

Figure 8.4: The approximations for $\sin^{-1}(x)$ and the error plotted versus x.

We note that the Chebyshev series error, nicely, has an oscillatory behavior throughout the region, and so remains lower overall through the region. The power series is more accurate near $x = 0$, but at the edge regions has larger error.

chapter8/chebpower.py

```
#!/usr/bin/env python2

import numpy as np
import matplotlib.pyplot as plt

#Test various a's for a x as an approximation to sin(x) in [0,pi/3]
def coeffarr(m):
  mm=np.zeros(m)
  for j in range(m)[1::2]:
    mm[j]=1./(j)**2
  return 4/np.pi*mm


theta=np.linspace(0,np.pi,501)
x=np.cos(theta)
x=np.linspace(-1,1,501)

y=np.polynomial.chebyshev.chebval(x,coeffarr(5))
y1=x+x**3/6.+3/40.*x**5
y2=np.arcsin(x)

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,y,label=r'$\rm{Chebyshev}$')
ax.plot(x,y1,label=r'$\rm{Power}$')
ax.plot(x,y2,label=r'$\sin^{-1} x$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('arcsinapprox.png',bbox_inches='tight')

plt.clf()

fig = plt.figure()
ax=fig.add_subplot(111)

ax.plot(x,np.abs(y -y2),label=r'$ \rm{Error\ Chebyshev}$')
ax.plot(x,np.abs(y1-y2),label=r'$\rm{Error\ Power}$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
ax.set_xlim(-1.05,1.05)
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('Errorarcsinapprox.png',bbox_inches='tight')

plt.clf()

fig = plt.figure()
```

```
ax=fig.add_subplot(111)

ax.semilogy(x,np.abs(y -y2),label=r'$␣\rm{Error\␣Chebyshev}$')
ax.semilogy(x,np.abs(y1-y2),label=r'$\rm{Error\␣Power}$')


plt.setp(ax.get_yticklabels(), fontsize=20)
plt.setp(ax.get_xticklabels(), fontsize=20)
ax.set_xlabel('$x$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30)
#ax.set_ylabel('$y(x)$',fontsize=30,rotation='horizontal')
ax.legend(loc='best',prop={'size':15})
ax.set_xlim(-1.05,1.05)
#plt.title(r'Real$(n)$')

plt.tight_layout()
plt.savefig('LogErrorarcsinapprox.png',bbox_inches='tight')

plt.clf()

theta=np.linspace(0,np.pi,11)
x=np.cos(theta)
x=np.linspace(-1,1,11)
print x

y1=np.polynomial.chebyshev.chebval(x,coeffarr(5))
y2=x+x**3/6.+3/40.*x**5
y3=np.arcsin(x)

print y1
print y2
print y3
```