Solutions for Exercises in: A First Look at Numerical Functional Analysis W. W. Sawyer

John L. Weatherwax   Delbert D. Franz

May 3, 2007

1 Introduction

This small book, published in 1978, is one of the best introductions to functional analysis with a clear focus on numerical analysis. The book contains occasional exercises to hone the skills of the reader. Working exercises in a mathematics book can sometimes be an exercise in frustration when one has no clue of how to start. Perhaps that condition arises from lack of mathematical maturity, mental block, or perhaps not understanding quite what the author of the question had in mind. In a normal classroom setting, there would be opportunity to ask a question of the instructor, or perhaps even a fellow student. When reading a mathematics text, outside of a classroom setting, neither option exists. Consequently this solutions manual is aimed at motivated readers of the above book who desire to get a better understanding of functional analysis.

A brief note on arrangement. The major heading for each set of exercises will be the page number of the book, perhaps followed by the section title from the book that relates to the exercises. Also references to the book will be given as: Sawyer.

2 Page 11: 2.3 Continuity and distance

The following is a short python program that interactively requests a starting value, and then prints all values until the convergence tolerance is met.

```
eps = 1.e-6
err = 1e30
x1 = input('Starting value=')
knt = 0
print "knt= %3d" % knt," x= %12.7f" % x1

while err > eps:
    x2 = 2*x1**3/(3*(x1**2 - 1))
    knt += 1
    print "knt= %3d" % knt," x= %12.7f" % x2
    err = abs(x1 - x2)
    x1 = x2
```

This program was run for each suggested starting values. The following table shows the resulting iterates and the final value when the convergence tolerance was set at 0.000001.

i	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
0	0.810000	0.780000	0.779000	0.775000	0.774700	0.774600	0.774600	0.770000
1	-1.030224	-0.807886	-0.801591	-0.777021	-0.775217	-0.774617	-0.774617	-0.747618
2	-11.879800	1.012109	0.960617	0.789318	0.778330	0.774717	0.774717	0.631602
3	-7.976385	28.367170	-7.653365	-0.869658	-0.797406	-0.775317	-0.775317	-0.279452
4	-5.402505	18.934978	-5.190864	1.799316	0.928271	0.778931	0.778931	0.015781
5	-3.729447	12.658625	-3.593957	1.735644	-3.855432	-0.801162	-0.801162	-0.000003
6	-2.678903	8.492079	-2.597034	1.732062	-2.755676	0.957231	0.957231	0.000000
7	-2.075085	5.740994	-2.032745	1.732051	-2.115732	-6.985362	-6.985362	-0.000000
8	-1.801841	3.947087	-1.787839	1.732051	-1.816230	-4.754343	-4.754343	
9	-1.735907	2.811877	-1.734558		-1.737563	-3.316275	-3.316275	
10	-1.732064	2.146002	-1.732056		-1.732077	-2.431986	-2.431986	
11	-1.732051	1.827489	-1.732051		-1.732051	-1.951227	-1.951227	
12	-1.732051	1.739041	-1.732051		-1.732051	-1.764190	-1.764190	
13		1.732093				-1.732908	-1.732908	
14		1.732051				-1.732051	-1.732051	
15		1.732051				-1.732051	-1.732051	

The iteration solves for the roots of the polynomial: $x^3 - 3x$. Starting values of -1.5, 1.5, and 0.0 find the roots much more quickly and with less erratic behavior. However, finding the approximate root values for a more complex equation often proves difficult. Consequently, the example is indicative of problems encountered when solving nonlinear equations. Good first estimates are always valuable!

3 Page 16-17 2.3 Continuity and distance

Problem 1 - the ℓ_1 distance

D.1: Arguments to absolute value are real numbers and the result is a real number.
D.2: Absolute value ≥ 0 by definition
D.3: Zero distance implies point identity
D.4: Order of values in absolute value operation has no effect on result.
D.5: Show that

$$|x_0 - x_1| + |y_0 - y_1| \leq [|x_0 - x_2| + |y_0 - y_2|] + [|x_2 - x_1| + |y_2 - y_1|]\,. \tag{1}$$

Insert $-x_2 + x_2$ and $-y_2 + y_2$ into the corresponding terms on the LHS (left-hand side) of Eq. 1. This insertion leaves the value unchanged. The LHS then becomes

$$|x_0 - x_1| + |y_0 - y_1| = |x_0 - x_2 + x_2 - x_1| + |y_0 - y_2 + y_2 - y_1|. \tag{2}$$

Each term on the RHS (right-hand side) of Eq. 2 is bounded by

$$|x_0 - x_2 + x_2 - x_1| \leq |x_0 - x_2| + |x_2 - x_1| \tag{3}$$

and

$$|y_0 - y_2 + y_2 - y_1| \leq |y_0 - y_2| + |y_2 - y_1| \tag{4}$$

because the LHS's of Ineq. 3 and 4 are subject to cancellation between differences in coordinates, that is, for any two real numbers, a, b, we have

$$|a + b| \leq |a| + |b|. \tag{5}$$

If both are zero we get equality; if both are of the same sign, we get equality again, and if they differ in sign, we get inequality.

Consequently, adding Ineq. 3 and 4, and using the result to replace the RHS of Eq. 2 yields the triangle inequality for the least first-power norm in Ineq. 1.

Problem 1 - the ℓ_∞ distance

D.1 though D.4: Essentially the same as for ℓ_1

D.5: By definition

$$d_\infty = \max\big(|x_0 - x_1|, |y_0 - y_1|\big). \tag{6}$$

Now insert $-x_2 + x_2$ and $-y_2 + y_2$ in the corresponding term in Eq. 6. This leaves the distance unchanged since we have added zero. This yields

$$d_\infty = \max\big(|x_0 - x_2 + x_2 - x_1|, |y_0 - y_2 + y_2 - y_1|\big). \tag{7}$$

Then use Ineq. 5 to rewrite this equation as an inequality

$$d_\infty \leq \max\big(|x_0 - x_2| + |x_2 - x_1|, |y_0 - y_2| + |y_2 - y_1|\big). \tag{8}$$

Now comes the move that is not always obvious at first. We assert that

$$d_\infty \leq \max\big(|x_0-x_2|+|x_2-x_1|, |y_0-y_2|+|y_2-y_1|\big) \leq \max\big(|x_0-x_2|, |y_0-y_2|\big)+\max\big(|x_2-x_1|, |y_2-y_1|\big). \tag{9}$$

To see this more clearly, note that each addend or argument in Ineq. 9 is positive. Therefore, let $A = |x_0 - x_2|$, $B = |x_2 - x_1|$, $C = |y_0 - y_2|$, and $D = |y_2 - y_1|$. We then assert that

$$\max\big(A + B, C + D\big) \leq \max\big(A, C\big) + \max\big(B, D\big). \tag{10}$$

Now if $A + B$ in Ineq. 10 is the larger, then $\max(A, C) + \max(B, D)$ cannot be smaller because at least $\max(A, C) \geq A$ and $\max(B, D) \geq B$! On the other hand, if $C + D$ is the larger, then $\max(A, C) \geq C$ and $\max(B, D) \geq D$ and again $\max(A, C)+\max(B, D)$ cannot be smaller than $C+D$. Consequently the triangle inequality follows from the definition of the uniform norm.

Problem 2 - the triangle inequality for the ℓ_2 distance

We give a derivation using algebra which is based on a derivation in L. B. Rall(2) (See references in Sawyer). It is an attempt to decipher the terse and unmotivated nine-line proof given by Rall! We again consider three distinct points in the plane. The triangle inequality for the Euclidean norm then becomes

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \leq \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2} + \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2} \tag{11}$$

The main idea is to apply a sequence of algebraic transformation to this inequality eventually reducing it to one that is "trivially" true. To begin this process, our first step is to simplify notation. Let $p_1 = x_2 - x_3$, $p_2 = y_2 - y_3$, $q_1 = x_3 - x_1$, and $q_2 = y_3 - y_1$. From these equations solve for x_2, y_2, x_1, and y_1, and then find that $x_2 - x_1 = p_1 + q_1$ and $y_2 - y_1 = p_2 + q_2$. Using these results in Ineq. 11 then yields

$$\sqrt{(p_1 + q_1)^2 + (p_2 + q_2)^2} \leq \sqrt{p_1^2 + p_2^2} + \sqrt{q_1^2 + q_2^2}. \tag{12}$$

We start by squaring the LHS and the RHS of Ineq. 12 which does not effect the truth of the inequality. For the square of the LHS we find, after expanding the argument to the square root,

$$p_1^2 + 2p_1q_1 + q_1^2 + p_2^2 + 2p_2q_2 + q_2^2. \tag{13}$$

while the square of the RHS becomes

$$p_1^2 + p_2^2 + 2\sqrt{(p_1q_1)^2 + (p_1q_2)^2 + (p_2q_1)^2 + (p_2q_2)^2} + q_1^2 + q_2^2. \tag{14}$$

Since the above expressions have many terms in common we can cancel them from both sides. In particular p_1, p_2, q_1, and q_2 are common to both sides and can be subtracted giving the following inequality

$$2(p_1q_1 + p_2q_2) \leq 2\sqrt{(p_1q_1)^2 + (p_1q_2)^2 + (p_2q_1)^2 + (p_2q_2)^2}. \tag{15}$$

Dividing both sides by 2 and then squaring again we obtain the following

$$(p_1q_1)^2 + 2p_1q_1p_2q_2 + (p_2q_2)^2 \leq (p_1q_1)^2 + (p_1q_2)^2 + (p_2q_1)^2 + (p_2q_2)^2. \tag{16}$$

Again a great number of terms can be canceled from both sides. In this case $(p_1q_1)^2$ and $(p_2q_2)^2$ are common to both sides and when canceled we have

$$2p_1q_1p_2q_2 \leq (p_1q_2)^2 + (p_2q_1)^2 \tag{17}$$

Or moving the term on the left hand side to the right we obtain

$$0 \leq (p_1q_2)^2 + (p_2q_1)^2 - 2p_1q_1p_2q_2 \tag{18}$$

In this inequality, note that these terms look like they are all part of a perfect square! In fact it is true that the above becomes

$$0 \leq \left[p1q2 - p2q1\right]^2 \tag{19}$$

which is seen to be trivially true since anything squared is necessarily positive. Since every transformation we have made up to this point is reversible and the final inequality is true the initial inequality must be true also and we have proved the triangle inequality for the ℓ_2 distance.

Expand Ineq. 19 and transpose the cross product term to the LHS of the relation to yield the end of our quest

$$2|p_1q_2||p_2q_2| \leq (p_1q_2)^2 + (p_2q_1)^2. \tag{20}$$

Ineq. 20 shows that Ineq. 18 is true. Ineq. 18 being true, establishes the truth of the assertion at Ineq. 16. Consequently the triangle inequality has been derived from the definition of the Euclidean distance.

This is a long series of steps but we suppose it is somewhat similar to what was done some time ago. However, it is not neat or concise, but we assert it is easier to comprehend at the first encounter. After some familiarity with the subject matter, it is, however, too long and wordy. Thus, in the rich tradition of mathematics, the proofs are shortened by those who understand them, but that action obscures them from the newcomer!

Here is a slightly expanded version of Rall's brief proof using notation that differs. I have added comments in italics that would have helped my understanding of the proof. I also add questions that occur on first encounter with this approach to proof!

We must prove that

$$\sqrt{(p_1 + q_1)^2 + (p_2 + q_2)^2} \leq \sqrt{p_1^2 + p_2^2} + \sqrt{q_1^2 + q_2^2} \tag{21}$$

Where did this come from? It looks like a distance of some kind but what happened to the coordinate points? To do this, observe for real p, q, that

$$0 \leq (|p| - |q|)^2 \tag{22}$$

or *expanding and transposing the cross product term*

$$2|pq|| \leq p^2 + q^2. \tag{23}$$

What does this equation have to do with what we are proving? There is a hint of what it might be for but one has to dig for it. Some words on why this, would have helped.

Setting $p = p_1q_2$, $q = p_2q_1$, we have *after adding* $(p_1q_2)^2 + (p_2q_1)^2$ *to both sides of the expression*

$$(p_1q_2)^2 + 2|p_1q_2||q_2a_1| + (p_2q_2)^2 \leq (p_1q_1)^2 + (q_1q_2)^2 + (p_2q_1)^2 + (p_2q_2)^2 \tag{24}$$

or

$$|q_1q_1| + |p_2q_2| \leq \sqrt{|p_1q_1|^2 + |p_1q_2|^2 + |p_2q_1|^2 + |p_2q_2|^2}. \tag{25}$$

Consequently

$$(p_1 + q_1)^2 + (p_2 + q_2)^2 \leq p_1^2 + 2|p1q1| + q_1^2 + p_2^2 + 2|p_2q_2| + q_2^2 \leq \left[\sqrt{p_1^2 + p_2^2} + \sqrt{q_1^2 + q_2^2}\right]^2. \tag{26}$$

Square RHS of Ineq. 21, then force cross product terms to be positive always, thus changing equality into potential inequality, then substitute the RHS of Ineq. 24 for the sum of the absolute values of the cross product terms in Ineq. 26. Then note that we have the square of the LHS of Ineq. 21. Just a few extra words and equations would have reduced the comprehension time from about two hours to perhaps 10 minutes!

Problem 3 : Hamming distance in five-bit signals

Part a:

The Hamming distance is a real number by definition, an integer, and is never negative. Furthermore it is only zero when the bit patterns are identical and the distance is independent of which signal is considered to be first in order. Consequently the first four axioms of distance are satisfied by this definition of a distance between points, where a point is a five-bit signal.

To prove that the triangle inequality follows from this definition, imagine that two distinct points out of the possible 32 points are selected and treat these as defining the base line of the triangle. Then in the next step, select a third point, distinct from the previous baseline points, and show that the sum of the distances from each baseline point to this third point is always ≥ the distance between the two baseline points. The algebra, if it can be called that, of this metric is not easily expressible because it is a count of the bits that differ between two signals. Consequently we list the possible distances between the base line points and then show that it is never the case that the distance can be shortened by taking the path that passes through the third point.

We first note that the possible distances between the baseline points are: $1, 2, 3, 4, 5$.

Distance is 1 or 2 Hamming units

Third point cannot shorten the distance because the minimum distance via a third distinct point is 2 and we have a distance of only 1 or 2 Hamming units on the baseline. Triangle inequality is true.

Distance is 3 Hamming units

Does a point exist that is one unit away from both base points when the base points are 3 units apart? This means that three out of the five positions differ for the baseline points. Pick one of the two base points, say the one on the left (we can assume that the signals are ordered in ascending order by numeric value). For the third point to be one unit away from this base point, either one or none of the bit positions that define the baseline distance will be changed. Thus at least two out of the three bit positions that define the baseline distance will remain unchanged. Consequently the minimum distance from the third point to the base point on the right is 2 Hamming units because at least two bit positions of the original three defining the baseline distance remain unchanged. Which baseline point we select is arbitrary so the same result obtains for either. Therefore, the triangle inequality is true if the Hamming distance between the two base points is 3.

Distance is 4 Hamming units

In this case 4 bit positions will differ between the two baseline points. We must consider distances of 1 and 2 from one baseline point for the third point. To be 1 Hamming unit away the third point can change at most 1 of the distance defining bits in the selected base point. Then at least 3 distance-defining bit location for the third point remain unchanged. Thus the minimum distance to the other baseline point is 3 Hamming units. Thus the minimum distance via the third point is 4 and the triangle inequality is true.

Consider a third point that is now 2 Hamming distance units away from one base point. By the same reasoning as above, at most two of the distance-defining bit locations in the third point can be changed. Thus at least two remain unchanged and the minimum distance from the other baseline point is 2. We again find that the minimum distance via the third point is $2+2=4$ and the triangle inequality is true.

Distance is 5 Hamming units

In this case 5 bit positions will differ. The bit pattern of one signal will be the bit-wise complement of the other. We have to consider distances of 1, 2, and 3 and the argument need not be stated again. In each of these cases the minimum distance will produce at best equality of distance and so the triangle inequality is true.

We have exhausted all possible baseline distances and third point distances, and in each case the triangle inequality was true. Consequently the Hamming distance between five-bit signals satisfies the five axioms for distance. With a bit more work, an induction proof for longer signals could be created.

Part b:

There are five locations available for bits in the signals. In order to establish a distance of n between two signals, we must select n positions out of five and then make sure that only those n positions differ. Consequently, we need the number of ways that n distinct positions (without repetition) can be selected out of five. This is

$$\binom{5}{n} = \frac{n!}{(5-n)!n!}. \tag{27}$$

The base bit pattern does not enter at all! The number of signals that is exactly a distance of 2 from the given signal, is then

$$\binom{5}{2} = \frac{n!}{(5-2)!n!} = 10. \tag{28}$$

The number of signals that are a Hamming distance of 2 or less from the given signal (or any other signal) is then

$$\binom{5}{2} + \binom{5}{1} = 15. \tag{29}$$

Part c:

$S(a,5)$ is the bit-wise complement of the bit pattern in point a. There is only one such signal for each possible a

Problem 4 : Minimum distance from line to origin

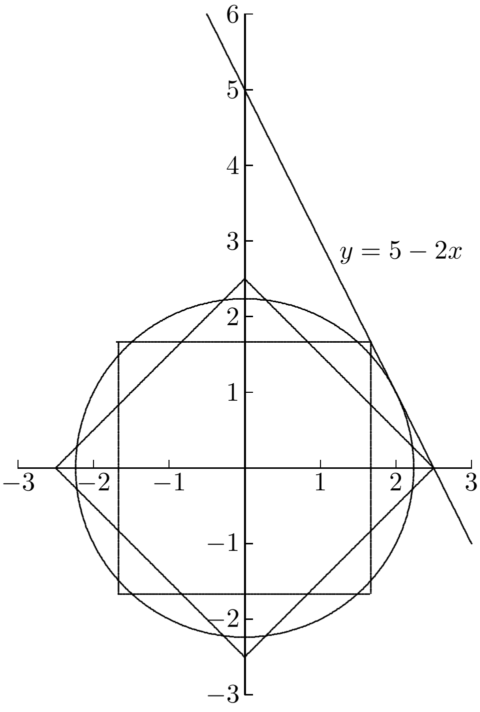
First we note that the points of interest are: $(x, 5-2x)$ for $0 \leq x \leq 2.5$. This is easily derived from the given data. We start with ℓ_1 which requests that we find the minimum value of $|x| + |5-2x|$. We will restrict our selves to the first quadrant where all values are ≥ 0, so we can write $x + 5 - 2x = 5 - x$. The minimum value of $5-x$ while keeping $5-2x \geq 0$ is given when $x = 2.5$ A way to check this result is to draw the diamond shaped sphere for the ℓ_1 distance, shown in Figure 2, page 16, of Sawyer, centered on the origin, as shown in the sketch. A moment's reflection shows that moving the point from $(2.5, 0)$, while keeping the sphere centered on the origin, can only make the sphere larger. Consequently the nearest point, in terms of the ℓ_1 distance is $(2.5, 0)$ and the minimum distance is 2.5

The ℓ_∞ distance is given by $\max(|x|, |5-2x|)$ and we want to find the minimum value of this distance. That is why this distance is sometimes called the minmax distance. The first thing to note is that the coordinates are inversely related, that is, an increase in one, yields a decrease in the other. Thus the minimum value will occur when the two values are the same. Since we are in the first quadrant, we can drop the absolute value signs and we find that $x = 5-2x$ and therefore the point of minimum ℓ_∞ distance is (5/3, 5/3) and the minimum distance is 5/3. Again drawing the sphere for this measure of distance, a square, the sketch, shows that moving the minimum-distance point, always keeping the sphere a square, can only make the sphere larger.

The Euclidean distance is $\ell_2 = \sqrt{x^2 + (5-2x)^2}$. Take the square of this distance and find its minimum using calculus. The derivative of $x^2 + (5-2x)^2$ with respect to x is $2x - 4(5-2x) = 10x - 20$. Setting this to zero and solving for x yields $x = 2$. Thus the minimum point for this distance is $(2, 1)$ and the minimum distance is $\sqrt{5}$.

We can also find this point and distance without using calculus. The sphere for this definition of distance is a circle. This circle will be tangent to the line $y = 5 - 2x$ at the minimum point. If this were not so, we would not have the closest point! A line drawn from the origin, the center of the circle and through the point of tangency is then perpendicular to the tangent line. A line perpendicular to $y = 5 - 2x$ has a slope of $-(1/-2) = 0.5$. This line passes through the origin so its equation is $y = 0.5x$ The intersection of these two lines is the minimum point. That is, solve $0.5x = 5 - 2x$ to obtain the result $x = 2$ and $y = 1$. Again moving the minimum point can only increase the size of the minimum sphere shown on the sketch.

The minimum distances vary inversely with the size of the subscript on the distance definition. That is, ℓ_1 is the largest at 2.5, this is followed by ℓ_2 at $\sqrt{5}$, and ℓ_∞ is the smallest with 5/3.



Sketch for Exercise 4, page 17

Problem 5

Part a:

πR = one-half circumference, so $\pi R/2$ = one-fourth the circumference. Therefore, $S(N, \pi R/2)$ defines all points on the equator.

Part b:

$B(N, \pi R/2)$ represents all points north of the equator. Northern hemisphere excluding points on the equator.

Part c:

$\bar{B}(N, \pi R)$ represents the surface of the earth or globe.

Part d:

$B(N, \pi R)$ represents the surface of the earth excluding the south pole

Part e:

$S(N, \pi R)$ represents the south pole.

4 Page 30 2.8 Axioms of a vector space

Problem 1

Is a vector space: (a, b, c) represents a point in three space.

Problem 2

Is a vector space: $Q(0) = 0$ implies that $c = 0$ in Problem 1. (a, b) represent a point in two space.

Problem 3

Is a vector space: $Q(0) = 0$ and $Q(1) = 0$ implies that $a = -b$. (a) represents a point in one space.

Problem 4

Is not a vector space: The three conditions imply that $a = b = c = 0$, and there is only one vector, the zero vector. Fails V.6 because only $a = 0$ produces a valid result!

Problem 5

Is a vector space.

Problem 6

Is a vector space.

Problem 7

Is a vector space.

Problem 8

Is a vector space.

Problem 9

Is a vector space.

Problem 10

Is not a vector space: Fails V.1 because adding two vectors could yield a vector with a bound larger than 100. This does not apply in Problem 9 because there each function has its own bound.

Problem 11

Is a vector space, in fact, a very important one!

Problem 12

Is a vector space. Addition of two functions as well as multiplication of a function by a real number, maintains $f(0.5) = 0$ in the result of the operation.

Problem 13

Is not a vector space. Addition of two functions as well as multiplication of functions by a scalar does not preserve $f(0.5) = 2$. Also has no zero vector. Fails V.1, V.4, and V.6.

Problem 14

Is not a vector space: We have a second order differential equation with two boundary conditions which define a unique solution! Consequently we have only one point in the space. Fails V.4 (unless the solution is trivial), and V.6 because only $a = 1$ will create a result that is in the solution space!

5 Page 36 Question above last paragraph

M being additive requires that $M0 = 0$ from the definition of additive given in equation 3 on page 36. No additional assumption is required.

6 Page 37 Question in last paragraph

Going through the process used by Sawyer, but with the initial vector ω_0 in place of the zero vector, yields an additional term in the final bounding series: $(k^{n+p} - k^n)\|\omega_0\|$. This term can be made as small as we like by making n large enough because $k < 1$. Consequently the effect of the initial vector eventually becomes nil.

Another way to argue is to observe that good first approximations reduce the number of iterations in an iterative computation. If we happened to pick, for example, the limit point as our start point, the iteration process would be quite short. Thus the zero vector is convenient but not always a good starting point.

7 Page 43 Exercise at top of page.

The following python program computes estimates of the maximum value of the function, the argument at the maximum, the estimated integral using the compound midpoint rule, and the value of the function when $x = 1$, for values of $n = 5, 10, 15 \ldots 80$.

```
from  math import *
m =1000
offset = 1.0/(2*m)

for n in xrange(5,81,5):
   nsqr= n*n
   maxg = 0.0
   sum = 0.0
   for i in xrange(m) :
      x=1.*i/m + offset
      g= (nsqr*x)**n*exp(-nsqr*x)
      if g > maxg:
          maxg = g
          xatmaxg = x
      maxg = max(g, maxg)
#     print "i= %4d" % i," x= %12.6f" % x, " g=",g
      sum += g
   gatone = (nsqr)**n*exp(-nsqr)
   print "n=",n, " xatmaxg=",xatmaxg," max g=",maxg, "integral=", sum/m
   print "      log10 maxg=",log10(maxg)," log10 integral=",log10(sum/m)
   print "gatone=",gatone
```

The following table gives some of the values and shows that the maximum of the function increases rapidly and gets ever closer to an argument of zero. At the same time the function value at the upper limit rapidly becomes small and soon is smaller than the smallest value represented in the 64-bit IEEE floating point representation! The value at the origin is zero always. The integral grows rapidly as well and values of n much larger than 80 soon result in an overflow exception.

n	(x at max g)	(max g)	(g(1))	(integral)
5	0.2005	21	0.0001	4.8
10	0.1005	453943	4×10^{-24}	36288
15	0.0665	1.3×10^{11}	4×10^{-63}	5.9×10^{9}
20	0.0505	2.2×10^{17}	2×10^{-122}	6.1×10^{15}
25	0.0405	1.2×10^{24}	3×10^{-202}	2.5×10^{22}
30	0.0335	1.9×10^{31}	0.0	2.9×10^{29}

8 Page 45 Iteration and contraction mappings

Problem 1

Part a: Function is strictly increasing with minimum of 4 and maximum of 7 on the interval: $||f|| = 7$.

Part b: Function has extremum at $x = 0.5$, that is, $f'(0.5) = 0$, with a value of -0.25. Therefore, $\|f\| = 0.25$.

Part c: Function changes sign in the interval and is strictly increasing. $f(0) = -3$ and $f(1) = 2$ and therefore $\|f\| = 3$.

Part d: $f \geq 0$ on given interval and is strictly increasing. Therefore $\|f\| = |f(1)| = 2$.

Part e: $|\cos \pi x|$ for all x, and thus $\|f\| = 1$.

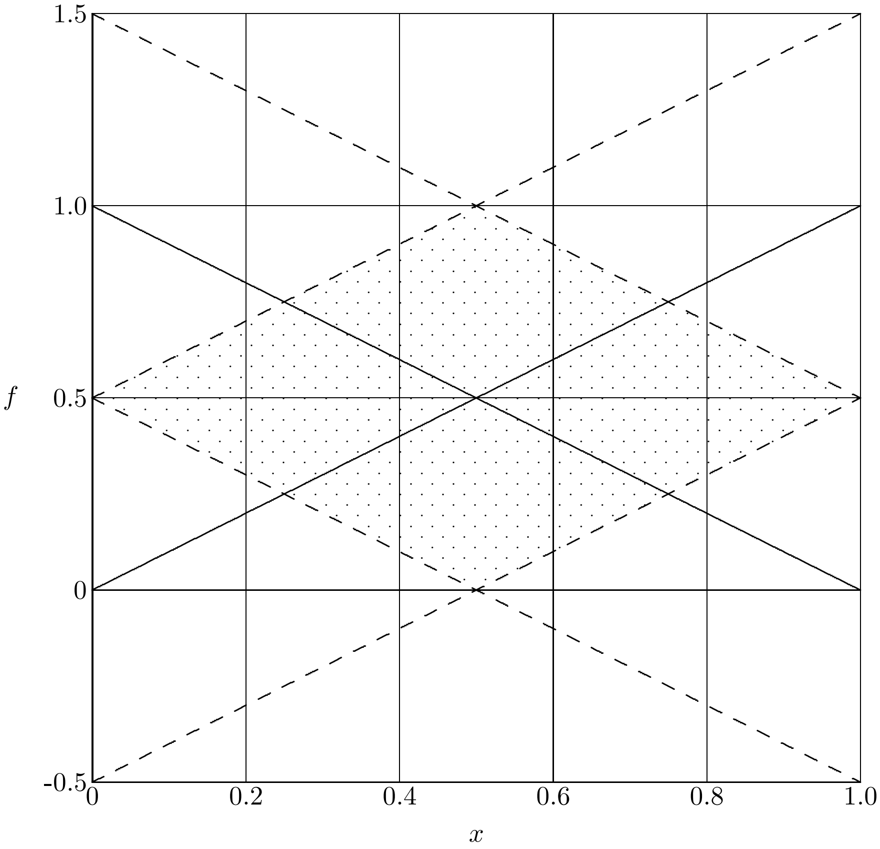
Part f: Maximum value is 0.05 at $x = 0.5$ and minimum value of -0.2 is at the ends of the interval. Therefore, $\|f\| = 0.2$.

Part g: Maximum value is 0.15 at $x - 0.5$ and minimum value of -0.1 at the ends of the interval. Therefore,

$\|f\| = 0.15$.
Part h: Consider $x^2 - x$. Extreme value in interval is -0.25 at $x = 0.5$. Consequently, $\|f\| = 0.35^{10} \approx 9.5 \times 10^{-7}$.

Problem 2

The shaded region shows the region in which the graph of f must lie. If open balls are used then the function becomes undefined at $x = 0$ and $x = 1$. We could however extend the definition of the function at those points in terms of limits.



Sketch for Exercise 2, page 45

Problem 3

Note that $g(x) \geq 0$ for all x in the given interval and that it is zero ($x = 0$ and $x = 1$. We also have $g'(x) = 9x^8(1-x)^{11} - 11x^9(1-x)^{10}$ and this is again zero at not only $x = 0$ and $x = 1$ but also $x = 9/20 = 0.45$, as division by $x^8(1-x)^{19}$ shows. Consequently the function reaches its maximum value in $[0, 1]$ at $x = 9/20$. The function f is a constant and thus the best fit horizontal line to g in $[0, 1]$ is $f = \max g/2 \approx 5.2 \times 10^{-7}$. Adding any slope to this line can only increase the maximum of the absolute value of its deviation from g. The reduction of the deviation at one point by adding a slope causes an increase at another point. A theorem in approximation theory shows that this horizontal line is the best fit first order polynomial to g in $[0, 1]$ in that there are three equal alternating sign deviations from it.

9 Page 49 Exercise near bottom of page

We develop the iteration first and then show convergence of the series we develop. The iteration yields the following sequence of functions:

$$g_0(x) = 0 \tag{30}$$

$$g_1(x) = v_0 \tag{31}$$

$$g_2(x) = v_0 + \int_0^x K(x,y) g_1(y)\, dy \tag{32}$$

$$g_3(x) = v_0 + \int_0^x K(x,y) g_2(y)\, dy \tag{33}$$

$$\vdots$$

$$g_{n-1}(x) = v_0 + \int_0^x K(x,y) g_{n-2}(y)\, dy \tag{34}$$

$$g_n(x) = v_0 + \int_0^x K(x,y) g_{n-1}(y)\, dy \tag{35}$$

Now let $p_n(x) = g_n(x) - g_{n-1}(x)$ for $n = 1, \ldots$. This gives another sequence of equations:

$$p_1(x) = g_1(x) - g_0(x) = v_0 \tag{36}$$

$$p_2(x) = g_2(x) - g_1(x) = \int_0^x K(x,y) p_1(y)\, dy \tag{37}$$

$$\vdots$$

$$p_n(x) = g_n(x) - g_{n-1}(x) = \int_0^x K(x,y) p_{n-1}(y)\, dy \tag{38}$$

For Eq. 36 we have $p_1(x) \le \|v_0\|$ from the given conditions. Substituting this outcome into Eq. 37 gives

$$p_2(x) \le \int_0^x K(x,y) \|p_1\|\, dy \le M\|v_0\|x \tag{39}$$

because M is an upper bound for the kernel function, K, and $x \ge 0$. In the same manner it is clear that

$$p_3(x) \le \int_0^x K(x,y) \|p_2\|\, dy \le M^2\|v_0\|x^2/2 \tag{40}$$

and

$$p_n(x) \le \int_0^x K(x,y) \|p_{n-1}\|\, dy \le M^{n-1}\|v_0\|x^{n-1}/(n-1)! \tag{41}$$

This establishes the first required result using the notation p_i instead of v_i. Note that we have $v_{i-1} = p_i$. From the definition of p_i we can write: $g_1 = p_1$, $g_2 = p_1 + p_2$, …, $g_n = p_1 + p_2 + s + p_n$. Equations 39-41 show that the sum

$$\sum_{1}^{n} \|p_i\| \leq \sum_{i}^{n} M^{i-1} \|v_0\| x^{i-1}/(i-1)!$$

where the right-hand series converges to $\|v_0\| e^x$ for all x. This latter series consists of absolute values so the theorem referenced near the middle of page 48 establishes that the iteration for g converge also.

Now if the above sequence of operations is repeated, starting with $g_0 = w_0$ where w_0 is some other continuous function, then the general term for the norm of the p_i becomes

$$p_n(x) \leq M^{n-1} \|v_0 - w_0\| \frac{x^{n-1}}{(n-1)!} + M^n \|w_0\| \frac{x^n}{n!} \tag{42}$$

The $\sum p_i$ is again convergent because the two series defined by Eq. 42 are each convergent for all x. Equation 42 suggests the useful simplification of using $w_0 = v_0$ which is the starting value used in discussing the iteration method of solution of Volterra integral equations.

10 Page 62 Exercises on linear functions

Depending on the nature of the function, we pick points in the linear space to check if $L(x+y) = L(x) + L(y)$ and that $L(kx) = kL(x)$, where k is a scalar appropriate to the linear space. If both relationships are true, the function is linear.

Problem 1

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

Assume two points in the vector space, u and v. Then $(u, y, x) \to u$, $(v, y, x) \to v$, and $(u + v, y, x) \to u + v$. Therefore, $L(u + v) = L(u) + L(v)$. Also $(ku, y, z) \to ku$ and then $L(ku) = kL(u)$. Therefore we have a linear operator.

Problem 2

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

$L(u) = (u, u, u)$, $L(v) = (v, v, v)$ and $L(u + v) = (u + v, u + v, u + v)$. Therefore $L(u + v) = L(u) + L(v)$. Also $L(ku) = (ku, ku, ku) = kL(u)$. We have a linear operator.

Problem 3

$\mathbb{R} \to \mathbb{R}$, $x \to x + 1$

$L(u) = u + 1$, $L(v) = v + 1$, and $L(u + v) = u + v + 1$. We have then that $L(u + v) \neq L(u) + L(v)$ and this operator is not linear.

Problem 4

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

$L(u_1, v_1) = (-v_1, u_1)$, $L(u_2, v_2) = (-v_2, u - 2)$, and $L(u_1 + u_2, v_1 + v_2) = (-v_1 - v_2, u_1 + u_2) = L(u_1, v_1) + L(u_2, v_2)$ Also $L(ku, kv) = (-kv, ku) = kL(u, v)$. We have a linear operator.

Problem 5

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

Need two continuous functions: $L(f) = f(0)$, $L(g) = g(0)$, and, $L(f+g) = f(0) + g(0) = L(f) + L(g)$. L(kf) = k(f(0)=kL(f). We have a linear operator.

Problem 6

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

$L(u_1, v_1) = \sqrt{u_1^2 + v_1^2}$, $L(u_2, v_2) = \sqrt{u_2^2 + v_2^2}$, and $L(u_1 + u_2, v_1 + v_2) = \sqrt{(u_1 + u_2)^2 + (v_1 + v_2)^2} \neq L(u_1, v_1) + L(u_2, v_2)$. Operator is not linear.

Problem 7

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

$L(g) = \|g\|$, $L(h) = \|h\|$, and $L(g + h) = \|g + h\| \leq L(g) + L(h)$. Operator is not linear.

Problem 8

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

$\|(x, y, z)\|_\infty = max[|x|, |y|, |z|]$. $L(a, b, c) = \max[|a|, |b|, |c|]$, $L(d, e, f) = max[|d|, |e|, |f|]$, and $L(a + d, b + e, c + f) = \max[|a + d|, |b + e|, |c + f|] \leq L(a, b, c) + L(d, e, f)$. Operator is not linear.

Problem 9

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

$\|(x, y, z)\|_1 = |x| + |y| + |z|$. Just as in problem 8, the sum of two numbers allows cancellation if they are of different signs. Consequently we can only say that $L(a + d, b + e, c + f) \leq L(a, b, c) + L(d, e, f)$. Operator is not linear.

Problem 10

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

$L(a, b, c) = a+b+c$, $L(d, e, f) = d+e+f$, and $L(a+d, b+e, c+f) = a+d+b+e+c+f = L(a, b, c)+L(d, e, f)$. also $L(ka, kb, kc) = ka + kb + kc = k(a + b + c) = kL(a, b, c)$. We have a linear operator.

Problem 11

$C[0,1] \to C[0,1]$, $f \to g$, $g(x) = [f(x)]^2$

$L[u(x)] = [u(x)]^2$, $L[v(x)] = [v(x)]^2$, and $L[u(x) + v(x)] = [u(x), +v(x)]^2 \neq L[u(x)] + L[v(x)]$. Operator is not linear.

Problem 12

$C[0,1] \to C[0,1]$, $f \to g$, $g(x) = [f(x^2)]$

$L[u(x)] = u(x^2)$, $L[v(x)] = v(x^2)$, let $h(x) = u(x) + v(x)$, then $L[h(x)] = h(x^2) = u(x^2) + v(x^2) = L[u(x)] + L[v(x)]$. Also $L[ku(x)] = ku(x^2) = kL[u(x)]$. We have a linear operator.

Problem 13

$C[0,1] \to \mathbb{R}^6$, $f \to v$, $v = [f(0), f(0.2), f(0.4), f(0.6), f(0.8), f(1)]$

$$L(f) = [f(0), f(0.2), f(0.4), f(0.6), f(0.8), f(1)],$$

$$L(g) = [g(0), g(0.2), g(0.4), g(0.6), g(0.8), g(1)],$$

and

$$L(f+g) = [f(0)+g(0), f(0.2)+g(0.2), f(0.4)+g(0.4), f(0.6)+g(0.6), f(0.8)+g(0.8), f(1)+g(1)] = L(f)+L(g).$$

Also,

$$L(kf) = [kf(0), kf(0.2), kf(0.4), kf(0.6), kf(0.8), kf(1)] == kL(f).$$

We have a linear operator.

Problem 14

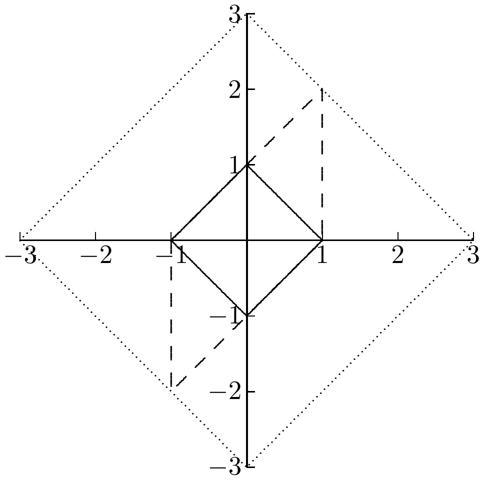
$C[0,1] \to \mathbb{R}$, $f \to \int_0^1 f(x)\,dx - 0.5[f(0) + f(1)]$.
$L(g) = \int_0^1 g(x)\,dx - 0.5[g(0)+g(1)]$, $L(h) = \int_0^1 h(x)\,dx - 0.5[h(0)+h(1)]$, and $L(g+h) = \int_0^1 [g(x)+h(x)]\,dx - 0.5[h(0) + g(0) + h(1) + g(1)] = L(g) + L(h)$. Also, $L(kg) = \int_0^1 kg(x)\,dx - 0.5[kg(0) + kg(1)] = kL(g)$. We have a linear operator.

Problem 15

$f \to g$, $g(x) = f(x+h) - f(x)$
$L(u) = u(x+h) - u(x)$, $L(v) = v(x+h) - v(x)$, and $L(u+v) = u(x)+v(x)-u(x+h)-v(x+h) = L(u)+L(v)$. Also, $L(ku) = ku(x+h) - ku(x) = kL(u)$. We have a linear operator.

11 Page 67: Excercise just below page middle

The sketch for the ℓ_1 norm shows the original unit sphere, a tilted square in Sawyer, the transformed shape, and the sphere that just encloses the transformed shape. The tilted square with solid lines is the original unit sphere, and the dashed tilted parallelogram is the transformed unit sphere. Note that two sides of the orignal unit sphere are not changed by transformation matrix. The points $(1,0)$ and $(-1,0)$ are not changed by the transform. The dotted tilted square is then the sphere that represents the $\|A\|$ which shows a value of 3.



Sketch for Exercise page 67: ℓ_1 case

Sawyer does not discuss the ℓ_2 induced operator norm for a matrix operator. It is somewhat more complex in computation than the other two norms. However, based on various sources, it proves not that difficult to derive. We want to be able to compute its value to use as a check on the computations used to draw the figure below. Let u be a vector on the unit sphere. Thus we have that $\|u\| = 1$. The matrix operator is then used to compute $x = Au$. We premultiply this equation with x^T, that is the transpose of x, which converts a column vector into a row vector. This then give us

$$f(u) = x^T x = u^T A^T A u$$

because $(Au)^T = u^T A^T$. Note that $x^T x$ is the inner product between two vectors and gives the sum of squares of the elements of x. This is just the value we want to make as large as possible while holding $\|u\| = 1$. We can impose this constraint using a Lagrange multiplier, λ, to yield the function we wish to maximize

$$H(u, \lambda) = u^T B u + \lambda(u^T u - 1)$$

where $B = A^T A$. Now take the derivative of H with respect to u and set it to zero to define equations that must be satisfied for an extreme to be present. We get

$$\frac{\partial H}{\partial u} = 2Bu - 2\lambda u = (B - \lambda I)u = 0$$

For this equation to have a solution other than the null vector, a vector of all zero elements, it must be true that

$$|B - \lambda I| = 0$$

which is just the defining equation for the eigenvalues of the matrix B. The matrix B by the nature of its definition is positive semi-definite. This means that its eigenvalues are zero or greater. Which of the eigenvalues should be pick? Premultiply the derivative of H by u^T to yield

$$u^T(B - \lambda I)u = u^T B u - \lambda u^T u = u^T B u - \lambda = 0$$

where we have used the requirement that $u^T u = 1$. The solution to this equation is that $\lambda = u^T B u$ implying that we should pick the largest eigenvalue because our goal is to maximize $u^T B u$.

In the current case

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

and

$$A^T = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

Then

$$A^T A = B = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$$

The eigenvalues for a 2×2 matrix are given by

$$\lambda_1 = \frac{1}{2}\left[b_{11} + b_{22} + \sqrt{(b_{11} + b_{22})^2 - 4(b_{11}b_{22} - b_{12}b_{21})}\right]$$

$$\lambda_2 = \frac{1}{2}\left[b_{11} + b_{22} - \sqrt{(b_{11} + b_{22})^2 - 4(b_{11}b_{22} - b_{12}b_{21})}\right]$$

From these equations and B we find that

$$\lambda_1 = 3 + \sqrt{5}$$

and

$$\lambda_2 = 3 - \sqrt{5}$$

Note that the sum of the eigenvalues is the same as the sum of the diagonal elements of B, called the trace of B, and this is true in general. Also the product of the eigenvalues gives the determinant of B, again a general relationship. The first of these facts will be used shortly to help provide a quick estimate of an upper bound on the norm of B.

So we find that

$$\|A\| = \sqrt{3 + \sqrt{5}}$$

Computing the eigenvalues for a 2×2 matrix is trivial. Doing so for an $n \times n$ matrix where n might be much larger than 2 is not so trivial. However, a simpler estimate of a bound for the norm found by the following argument: the eigenvalues of B are zero or positive. Therefore the sum of the eigenvalues will be equal to or larger than the largest eigenvalue. That is, for the general case of an $n \times n$ matrix, $\lambda_{max} \leq \sum_{i=1}^{n} \lambda_i$. The trace of B gives the sum of all the eigenvalues and the nature of the construction of B from A shows that the trace of B is the same as the sum of the squared elements in A. Thus an easily computed upper bound for the ℓ_2 norm of A is given by

$$\|A\| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$$

In our current case, this gives an upper bound of $\sqrt{1+5} = \sqrt{6} \approx 2.45$ which is not much large than the exact value of $\sqrt{3+\sqrt{5}} \approx 2.29$.

Finally, if we compute the eigenvector that corresponds to the largest eigenvalue, we can find the value of u that causes Au to meet the norm we have computed. The eigenvector can only be found up to a scale factor. Thus we scale the elements so the sum of their squares is 1, which we must have for u to be on the unit sphere. We then get the pair of equations

$$\begin{bmatrix} 1 - (3 + \sqrt{5}) & 1 \\ 1 & 5 - ((3 - \sqrt{5}) \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

These two equations are proportional to each other, so pick the first equation and find that $u_2 = (2 + \sqrt{5})u_1$. Using this result and requiring the sum of the squares of the elements to be 1, finds that

$$u_1 = \sqrt{\frac{5 - 2\sqrt{5}}{10}}$$

$$u_2 = \sqrt{\frac{5 + 2\sqrt{5}}{10}}$$

Therefore we find that the point on the boundary of the unit circle as transformed by A is

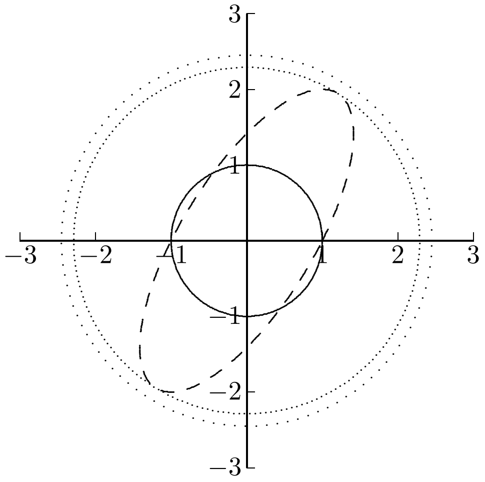
$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} u_1 + u_2 \\ 2u_2 \end{bmatrix}$$

We then find that

$$x \approx \begin{bmatrix} 1.203002 \\ 1.946498 \end{bmatrix}$$

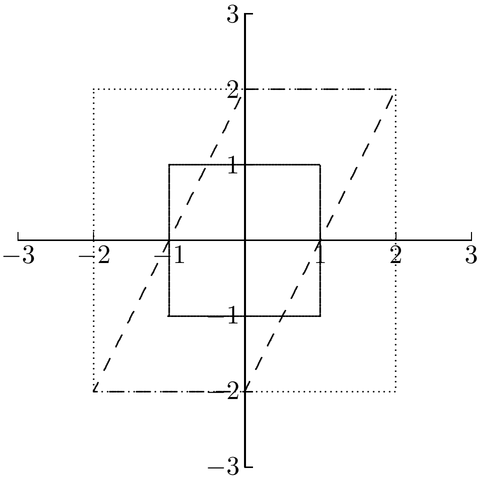
The sum of these two values squared is about 5.236068 and the square root of this last value is about 2.288245 which is a close approximation to $\sqrt{3 + \sqrt{5}}$.

The sketch for the ℓ_2 norm shows the same three items as for the ℓ_1 norm. Here of course the spheres are circles. We computed 41 points on the unit sphere using Python and the NumPy package to compute the transformed shape. The largest sphere shown is for the estimated upper bound on $\|A\|$ computed above. In this case it is quite close and is of course much easier to compute than the exact bound.



Sketch for Exercise page 67: ℓ_2 case

The sketch for the ℓ_∞ norm shows the original unit sphere, the transformed shape, and the sphere that just encloses the transformed shape. The square with solid lines is the original unit sphere, and the dashed parallelogram is the transformed unit sphere. The dotted square is then the sphere that represents the $\|A\|$ which shows a value of 2.



Sketch for Exercise page 67: ℓ_∞ case

12 Page 68: Exercises on operator norms

Problem 1

The maximum value for $|w|$ occurs at that point where all terms are positive, that is, $(1, -1, 1)$. $\|f\| = 9$.

Problem 2

$\|f\|_\infty = |a| + |b| + |c|$

Problem 3

$\|f\|_\infty = \sum_1^n |a_i|$

Problem 4

$x = (1, -1, 1)$ and $x = (-1, 1, -1)$ provide the same extreme for $\|w\|_\infty$. The value of $\|f\|$ is 9.

Problem 5

Part a: No change in the result.
Part b: $x = (1, 1, -1)$ and $\|f\| = 10$

Problem 6

The norm of the transform would be $\max \sum_s |a_{rs}|$ for $1 \leq r \leq m$. That is, we fine the norm for each row in the matrix of coefficients, and then take the largest one to be the norm for the operator.

Problem 7

The maximum value is obtained with $x(0, 0, 1)$ and the value is 4.

Problem 8

$\|f\|_1 = \max[|a|, |b|, |c|]$

Problem 9

$\|f\|_1 = \max[|a_r|; r = 1, \ldots, n]$

Problem 10

Maximum value of $|w_1| + |w_2|$ is 10. The value of x at this maximum is: $(1, 0)$ or $(-1, 0)$. $\|f\| = 10$.

Problem 11

Part a: x is $(0, 1)$ or $(0, -1)$, $\|f\| = 11$.
Part b: x is any one of the four corner points, $\|f\| = 10$.
Part c: x is any one of the four corner points, $\|f\| = 10$.

Problem 12

The norm would be the maximum of the norms of the column vectors of the matrix $a_r s$, that is,

$$\max_{0 \leq s \leq m} \sum_{r=1}^{n} |a_r s|$$

.

Problem 13

The kernel $3x^2 + 2x + 1$ is always > 0, thus the maximum value for c is given by replacing $f(x)$ by its maximum value of 1 and evaluating the integral. This gives $\max[c] = 3$.

Problem 14

The kernel $\sin x$, changes sign at π. Thus we set $f(x) = 1$ for $0 \leq x \leq \pi$ and $f(x) = -1$ for $\pi < x \leq 2\pi$. To make f continuous we can "smooth" the discontinuity by introducing a small region, centered on $x = \pi$ of length ϵ such that $f(x) = 1$ for $0 \leq x \leq \pi - \epsilon/2$, $f(x) = -2(x - \pi)/\epsilon$ for $\pi - \epsilon/2 \leq x \leq \pi + \epsilon/2$, and $f(x) = -1$ for $\pi + \epsilon \leq x \leq 2\pi$. That is, we use a steeply inclined line segment to move from a function value of 1 at $x = \pi - \epsilon/2$, to a function value of -1 at $x = \pi + \epsilon/2$. Thus we can maintain a continuous function so long as $\epsilon > 0$. If we use this continuous function we find that the supremum of the integral is the same as the maximum we obtain with a discontinuous function, which is $2 \int_0^{\pi} \sin x \, dx = 2 \cdot 2 = 4$

Problem 15

Seems clear that $\int_a^b |\phi(x)| \, dx$ would give the same results for the two cases. The goal is to avoid any cancellation of "area" caused by sign changes in the integrand. Taking the absolute value of the kernel function accomplishes this.

Problem 16

The kernel function is positive because both x and y are in the interval $[0, 1]$. Thus $f(x) = 1$ for $0 \leq x \leq 1$ is the function that makes g as large as possible. However, we must also choose the value of x to give the maximum as well. Thus use a two step process: first, set f to its maximum value and evaluate the integral, and second, choose x to get the larger value for $\|g\|$. The integral evaluates to: $x^2 + x + 1$, and this has its maximum value at $x = 1$. Consequently, $\|T\| = 3$.

Problem 17

Part i: In this case we have to be even more clever because the integrand changes sign within the interval of integration but the point of the change is a parameter and not a fixed point. The trick is to notice that the kernel function, $k - y$, changes sign at $y = k$. Thus we break the interval of integration into two parts: $0 \leq y \leq k$ and $k \leq y \leq 1$:

$$\int_0^k (k - y)\, dy - \int_k^1 (k - y)\, dy$$

where we have already set f to is maximum value of 1. The integrand in each of these integrals is of constant sign with the first being positive and the second being negative. Each of the integrals is easily evaluated to give $k^2 - k + 0.5$. The first derivative of the this result is $2k - 1$ and setting this to zero and solving yields $k = 0.5$ as an extreme point. At that point we find that $/sup|c| = 0.25$.
Part ii: The pattern for the definition of T is such that its norm would be the same as the supremum of c found in part i, that is, $\|T\| = 0.25$.

Problem 18

As Sawyer suggests by the note, we can approximate the kernel function, $K(x, y)$ by a matrix of values. We can create a two dimensional grid on $[0, 1]$ in x and y that creates a square matrix with each coordinate point representing the same area. For example, divide each dimension into n equal intervals. Then define argument values for the kernel function at the midpoints of these intervals. Let $[x_i; i = 1, \ldots, n]$ be the sequence for x and $[y_j; j - 1, \ldots, n]$ be the sequence for y. Then we can approximate the integral with the sum

$$\hat{g}(x_i) = \sum_1^n k_{ij} f(y_j) \frac{1}{n}$$

where $k_{ij} = K(x_i, y_j)$. This approximation can be made as accurate as needed for continuous K and f by choosing n large enough. This has transformed the integral into a sum that is analogous to that in problem 6 above. Thus what we want for the norm of the operator, T, is the maximum value of the norms of the rows in the coefficient matrix k,

$$\max_{1 \leq i \leq n} \sum_{j=1}^n |k_{ij}|$$

This summation is an approximation to the integral

$$\|T\| = \sup_{0 \leq x \leq 1} \int_0^1 |K(x, y)|\, dy$$

If K changes sign in the interval, we break up the interval of integration to make the integrand of constant sign in each subinterval.

13 Page 73: Exercises:The space of bounded linear operators

Problem 1

First off, the paperback printing of the book appears to have a typographical error. Since the operator T is given, it must be that, "Find T and ..." should have been "Find $\|T\|$ and ...". The kernel of the operator in this case is just t which does not change sign in the interval of integration. Consequently

$$\|T\| = \sup_{0\leq x\leq 1} \int_0^x |t|\,dt \leq \sup_{0\leq x\leq 1} \int_0^x t\,dt$$

$$\|T\| = \sup_{0\leq x\leq 1} \left.\frac{t^2}{2}\right|_0^x = \frac{1}{2}$$

The norm of the operator is less than 1 so the series converges.

We start with $f_0(x) = 0$, so $f_1(x) = 1$, and then we get the following values for the partial series:

$$f_2(x) = 1 + \int_0^x t\,dt = 1 + \frac{x^2}{2}$$

$$f_3(x) = 1 + \int_0^x t\left(1 + \frac{t^2}{2}\right)\,dt = 1 + \frac{x^2}{2} + \frac{x^4}{24}$$

$$f_4(x) = 1 + \int_0^x \left(t + \frac{t^3}{2} + \frac{t^5}{24}\right)\,dt = 1 + \frac{x^2}{2} + \frac{x^4}{24} + \frac{x^6}{246}$$

The general term becomes

$$\frac{x^{2n}}{2\cdot 4\cdots 2n}$$

Let $y = x^2/2$ which means that $x^2 = 2y$ and substitute into the n-th term to yield

$$\frac{(2y)^n}{2\cdot 4\cdots 2n} = \frac{2^n y^n}{2^n n!} = \frac{y^n}{n!}$$

noting that there are n even terms in the denominator of the n-th term. The nature of the series is simplified and its sum becomes obvious, that is,

$$\lim_{n\to\infty} 1 + y + \frac{y^2}{2!} + \frac{y^3}{3!} + \cdots + \frac{y^n}{n!} = e^y$$

Consequently, $f(x) = e^{x^2/2}$.

Now substitute into the integral equation

$$e^{x^2/2} = 1 + \int_0^x te^{t^2/2}\,dt$$

Let $u = t^2/2$, then $du = tdt$. Also, $t = 0 \Rightarrow u = 0$, and $t = x \Rightarrow u = x^2/2$. Apply these substitutions to the left-hand side (LHS) of the integral equation, to yield

$$1 + \int_0^{x^2/2} e^u\,du = 1 + \left.e^u\right|_0^{x^2/2} = 1 + e^{x^2/2} - 1 = e^{x^2/2}$$

Yes, it is a solution!

Problem 2

Compute several powers of M to detect possible patterns in the elements of the resulting powers. Then from those patterns infer what the sum of the series are for each element to find N.

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

$$M^3 = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} ab & 0 \\ 0 & ab \end{bmatrix} = \begin{bmatrix} 0 & a^2b \\ ab^2 & 0 \end{bmatrix}$$

$$M^4 = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & a^2b \\ ab^2 & 0 \end{bmatrix} = \begin{bmatrix} a^2b^2 & 0 \\ 0 & a^2b^2 \end{bmatrix}$$

$$M^5 = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} a^2b^2 & 0 \\ 0 & a^2b^2 \end{bmatrix} = \begin{bmatrix} 0 & a^3b^2 \\ a^2b^3 & 0 \end{bmatrix}$$

$$M^6 = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & a^3b^2 \\ a^2b^3 & 0 \end{bmatrix} = \begin{bmatrix} a^3b^3 & 0 \\ 0 & a^3b^3 \end{bmatrix}$$

Now collect the terms that form the series for each of the elements in N and also impose the requirement that $ab < 1$, so that the series converge.

$$n_{1,1} = 1 + ab + (ab)^2 + (ab)^3 + \cdots = \frac{1}{1 - ab}$$

$$n_{1,2} = 0 + a + a^2b + a^3b^2 + \cdots = a(1 + ab + (ab)^2 + s) = \frac{a}{1 - ab}$$

$$n_{2,1} = 0 + b + ab^2 + a^2b^3 + \cdots = b(1 + ab + (ab)^2 + s) = \frac{b}{1 - ab}$$

$$n_{2,2} = 1 + ab + (ab)^2 + (ab)^3 + \cdots = \frac{1}{1 - ab}$$

We then have that

$$N = \frac{1}{1 - ab} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix}$$

Check if $(I - M)N = I$:

$$(I - M)N = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \right) \frac{1}{1 - ab} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} = \frac{1}{1 - ab} \begin{bmatrix} 1 & -a \\ -b & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

The relationship is verified.
The eigenvalues of M are defined by $|M - \lambda I| = 0$, yielding

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

The determinant of this matrix is $\lambda^2 - ab$ and this is zero if $\lambda = \sqrt{ab}$.

Problem 3

Find the norm for the operator $\|M\|$. Again we take $f(y)| \leq 1$. With $f(y)$ fixed at its maximum value of 1, we have

$$\|M\| = \sup_{0\leq x\leq 1}\left|\int_0^x yf(y)\,dy + \int_x^1 xf(y)\,dy\right| \leq \sup_{0\leq x\leq 1}\left|\int_0^x y\,dy + \int_x^1 x\,dy\right| \leq \sup_{0\leq x\leq 1}\left|\int_0^x y\,dy\right| + \sup_{0\leq x\leq 1}\left|\int_x^1 x\,dy\right|$$

Evaluate the last two integrals and we get

$$\|M\| = \sup_{0\leq x\leq 1}\frac{x^2}{2} + \sup_{0\leq x\leq 1}|x - x^2|$$

The first of these terms has a maximum value of 1/2 at $x = 1$ and the second a maximum value of 1/4 at $x = 1/2$ for a final result of $\|M\| = 0.75$. However, this result is not the best estimate of the operator norm. We note that neither integrand is ever negative so that we can find the operator norm from

$$\sup_{0\leq x\leq 1}\left|\int_0^x y\,dy + \int_x^1 x\,dy\right| = \sup_{0\leq x\leq 1}\left|\frac{x^2}{2} + x - x^2\right| = \sup_{0\leq x\leq 1}\left|x - \frac{x^2}{2}\right|$$

and the last term has a maximum value of 1/2 at $x = 1$. In either case the proposed iteration will converge because $\|M\| < 1$.

Several iterations of this operator were computed using the symbolic manipulation features of Mathcad. The results were converted to Latex and then edited to create the following sequence of results. Note that the order of terms is as Mathcad gave them. One of the challenges of automated symbolic manipulation programs is how terms are grouped and in what order they are given.

For completeness we start with the null function,

$$f_0\,(x) = 0$$

which gives the first approximation easily

$$f_1(x) = 1$$

We then start evaluating the integrals symbolically. Some of the numerical coefficients have been converted to factorial form to shorten the equations and to also show the pattern of the evolving series more clearly.

$$f_2(x) = 1 + \int_0^x yf_1\,(y)\,dy + x\int_x^1 f_1\,(y)\,dy = 1 - \frac{1}{2!}x^2 + x$$

$$f_3(x) = 1 + \int_0^x yf_2(y)\,dy + x\int_x^1 f_2(y)\,dy = 1 + \frac{1}{4!}x^4 - \frac{1}{3!}x^3 - \frac{1}{2!}x^2 + \frac{4}{3}x$$

$$f_4(x) = 1 + \int_0^x yf_3(y)\,dy + x\int_x^1 f_3(y)\,dy = 1 - \frac{1}{6!}x^6 + \frac{1}{5!}x^5 + \frac{1}{4!}x^4 - \frac{2}{9}x^3 - \frac{1}{2!}x^2 + \frac{22}{15}x$$

We omit the restatement of the integrals to allow the following results to fit on one line.

$$f_5(x) = 1 + \frac{1}{8!}x^8 - \frac{1}{7!}x^7 - \frac{1}{6!}x^6 + \frac{1}{90}x^5 + \frac{1}{4!}x^4 - \frac{11}{45}x^3 - \frac{1}{21}x^2 + \frac{479}{315}x$$

$$f_6(x) = 1 - \frac{1}{10!}x^{10} - \frac{479}{1890}x^3 + \frac{1}{9!}x^9 - \frac{1}{2!}x^2 + \frac{1}{8!}x^8 - \frac{1}{3780}x^7 - \frac{1}{6!}x^6 + \frac{11}{900}x^5 + \frac{1}{4!}x^4 + \frac{4373}{2835}x$$

This process can be continued as far as one's patience permits. However, there are already signs of some pattern appearing in the results. In particular the even powers have a simple pattern in g_6, that is

$$1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \frac{1}{8!}x^8 - \frac{1}{10!}x^{10}$$

where the order of the powers has been rearranged. This expression contains the first few terms of the series expansion of $\cos x$. The series made up of the odd powers must then represent some other series expansion of a function. However, the pattern is not that clear and even division by the factor on the first power of x does not provide that much insight.

These iterations of the operator are approximating the solution to the integral equation

$$f(x) = 1 + \int_0^x yf(y)\,dy + x\int_0^1 f(y)\,dy$$

and the method outlined by Sawyer on page 46 provide an estimate of how close a given approximation might be to the limit point of the iteration. For example, we can say that $\|f_6 - f_\infty\| \leq 2^{-7} + 2^{-8} + 2^{-9} + \dots$ and this gives the estimate $\|f_6 - f_\infty\| \leq 2^{-6} \approx 0.016$.

Hints of the solution are in the series evolving as more terms are added. Clearly one term in the solution is $\cos x$. Taking the derivative of both sides of the integral equation with respect to x yields

$$f'(x) = \int_x^1 f(y)\,dy$$

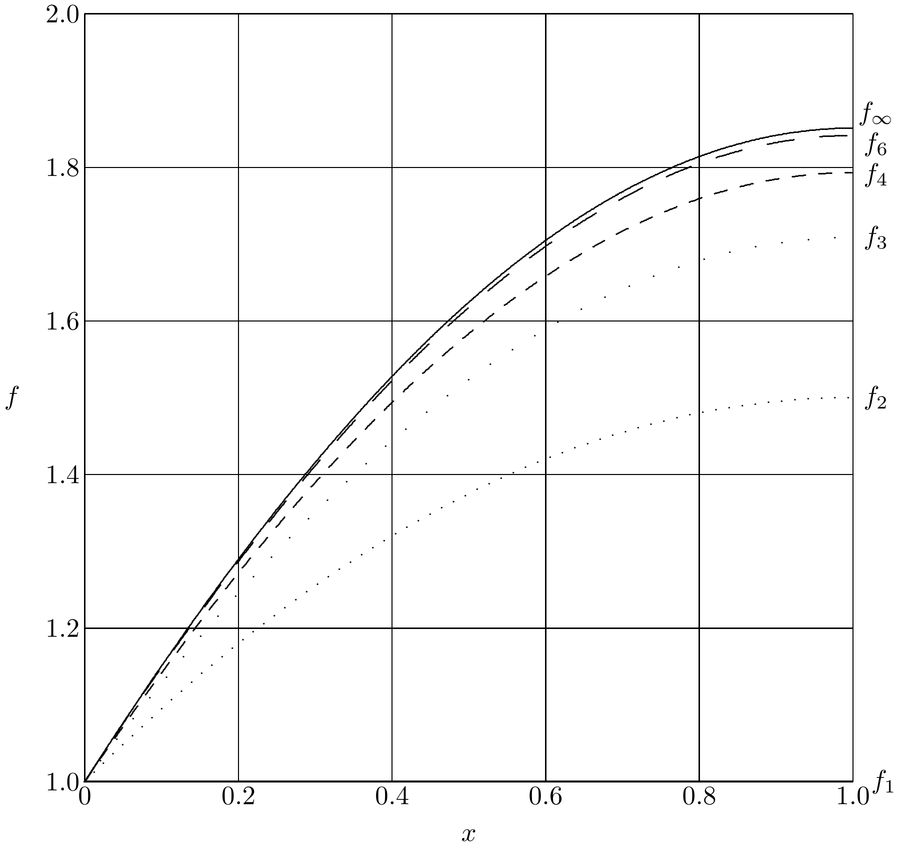
which show that $f'(x) = 0$ when $x = 1$. When $x = 0$, the integral equation shows that $f(x) = 1$. Taking the derivative of the first derivative gives

$$f''(x) = -f(x)$$

a second order ordinary differential equation. A solution for this equation is then of the form $a\cos x + b\sin x$ where a and b are determined by the end conditions. The condition that $f(x) = 1$ when $x = 0$ shows that $a = 1$. The derivative condition at the end point shows that $-\sin 1 + b\cos 1 = 0$ and we have that $b = \tan 1$. Thus the solution to the integral equation is $f(x) = \cos x + \tan 1 \sin x$. Substitution into the integral equation shows that this is the solution.

The distance between the solution and f_6, that is $\|f_6 - f\|$ is computed at $x = 1$ to be approximately $0.0095 < 0.016$. As is expected, the bounds on the solution are often quite wide. Iterations not shown here continued until f_9 and the estimated bound there was about 0.002 and the actual bound was about 0.00063.

The sketch shows some of these results in graphic form. Clearly the approximation improves rapidly as the iterations progress. This is a reflection of the size of the operator norm. Had it been closer to 1, the convergence would have been slower.



Sketch for Exercise 3, page 73

14 Page 88 Exercises: A classical theory of iteration

Problem 1

We find that $f'(x) = 0.5(1 - a/x^2)$. What condition must exist between a and x such that $|f'(x)| < 1$ (Note that there is an error in the paperback printing in that the absolute value is not used for f'.)? This inequality is really two: the first being that $-1 < 0.5(1-a/x^2)$, and the second being that $0.5(1-a/x^2) < 1$. The first inequality yields the requirement that $a < 3x^2$ and the second that $a > -x^2$. This last requirement really translates to $a > 0$ since the square of a real number cannot be negative and we are working with real numbers! Yes, a few trials with various numbers makes clear that the iteration converges for any given $a > 0$ and an initial $x > 0$. If the initial estimate is negative, then the iteration converges to the negative square root of a. If one tries to find the square root of a negative number, the iteration either fails to converge or terminates with an attempted division by zero.

The iteration is derived from Newton's method applied to $F(x) = x^2 - a = 0$ and the requirement that $|f'(x)| < 1$ no longer applies. Other criteria define the conditions for convergence and the square root example is one in which the interval of convergence is truly broad!

Problem 2

We compute $x_1 = \sin 2 + 0.5x \approx 1.90930$. Also $|2 - 1.90930|/0.5 \approx 0.1814$. Therefore all iterates will be within this distance of 2. To estimate the number of iterates required to have the result be within 1×10^{-9}

of the limit, find the value n such that $0.5^n(0.1814) = 1 \times 10^{-9}$. This gives an $n \approx 27$. The following table shows the iterates with the ninth decimal position becoming constant at about iteration 11. The table also shows that the derivative is always less than one in absolute value.

n	x_n	$f'(x_n)$
0	2.00000000000	0.08385
1	1.90929742683	0.08385
2	1.89790218350	0.16793
3	1.89592729887	0.17870
4	1.89557254719	0.18057
5	1.89550843107	0.18090
6	1.89549683031	0.18096
7	1.89549473093	0.18098
8	1.89549435099	0.18098
9	1.89549428223	0.18098
10	1.89549426978	0.18098
11	1.89549426753	0.18098
12	1.89549426712	0.18098
13	1.89549426705	0.18098

Problem 3

Part a: We have that $f(x) = \tan x$, and then $f'(x) = \sec^2 x$. At the starting point, $x = 4.5$ we find that $f'(4.5) \approx 22.5$. This is not a good method to find the root!

Part b: Here, $f(x) = \pi + \tan^{-1} x$ and $f'(x) = 1/(1 + x^2)$. Thus, $f'(4.5) \approx 0.047$. This equation is a good choice for finding the root. Four iterations get close to a precision of 1×10^{-4}!

15 Page 92: Exercises: Differentiation and integration

Problem 1

Part a: If $f(x, y) = \begin{bmatrix} -y \\ x \end{bmatrix}$ then $f'(x, y) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

Part b: If $f(x, y) = \begin{bmatrix} x + y \\ x - y \end{bmatrix}$ then $f'(x, y) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$

Part c: If $f(x, y) = \begin{bmatrix} x + y \\ x + y \end{bmatrix}$ then $f'(x, y) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Part d: If $f(x, y) = \begin{bmatrix} x^2 - y^2 \\ 2xy \end{bmatrix}$ then $f'(x, y) = \begin{bmatrix} 2x & -2y \\ 2y & 2x \end{bmatrix}$

Part e: If $f(x, y) = \begin{bmatrix} e^x \cos y \\ e^x siny \end{bmatrix}$ then $f'(x, y) = \begin{bmatrix} e^x \cos y & -e^x \sin y \\ e^x \sin y & e^x \cos y \end{bmatrix}$

Problem 2

Part a: If $f(x, y) = x^2 + y^2$ then $f'(x, y) = [\,2x \quad 2y\,]$

Part b: If $f(t) = \begin{bmatrix} a \cos t \\ b \sin t \end{bmatrix}$ then $f'(t) = \begin{bmatrix} -a \sin t \\ b \cos t \end{bmatrix}$

6.4: Some worked examples with the Frechet derivative

Exercise 1 - computing the Frechet derivative

For each of the given functionals we desire to compute the Frechet derivative at the "point" u_0, where in these examples this "point" is really a function. This derivative is computed by assuming a perturbation $h(x)$ to our base point u_0 and computing the linear operator "M" that expresses the change in the functional F in terms of the change in the input $h(x)$ i.e.

$$\Delta F \equiv F(u_0 + h) - F(u_0) = Mh + e(h)\,,$$

where $e(h)$ is an "error term" such that as $||h|| \to 0$, we have

$$\frac{||e(h)||}{||h||} \to 0\,.$$

If such an operator "M" exists then F is said to be Frechet-differentiable at u_0 and we define $F'(u_0) = M$. We now compute the Frechet derivative of the given functionals.

Part (a): For the functional $F(f) = f(x)^2$ we have the following dropping the x dependence on f and h for clarity

$$\Delta F \equiv (f+h)^2 - f^2 = f^2 + 2fh + h^2 - f^2 = 2fh + h^2\,.$$

Dropping terms that vanish faster than linear as $h \to 0$ we can see that our Frechet derivative for this functional is given by

$$F'(f)h \cdot (x) = 2f(x)h(x)\,.$$

Part (b): For the functional $F(f) = f(x)^n$ we have the following

$$\begin{aligned}
\Delta F &\equiv (f+h)^n - f^n \\
&= \sum_{k=0}^{n} \binom{n}{k} f^k h^{n-k} - f^n \\
&= \sum_{k=0}^{n-1} \binom{n}{k} f^k h^{n-k} \\
&= nf^{n-1}h + \frac{n(n-1)}{2} f^{n-2}h^2 + \cdots + nfh^{n-1} + h^n\,.
\end{aligned}$$

From which we can see (by dropping terms sub-dominant to h as $h \to 0$) that our Frechet derivative is

$$F'(f)h \cdot (x) = nf(x)^{n-1}h(x)\,.$$

Part (c): For the functional $F(f) = x^3(f(x))^2$ we have the following

$$\begin{aligned}
x^3(f(x) + h(x))^2 - x^3 f(x)^2 &= x^3(f(x)^2 + 2f(x)h(x) + h(x)^2) - x^3 f(x)^2 \\
&= x^3(2h(x)f(x) + h(x)^2)\,.
\end{aligned}$$

From which we see that x^3 effectively acts like a constant and our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = 2x^3 f(x)h(x)\,.$$

Part (d): For the functional $F(f) = \sin(f(x))$ we have the following (dropping the x dependence)

$$\begin{aligned}
\sin(f+h) - \sin(f) &= \sin(f)\cos(h) + \cos(f)\sin(h) - \sin(f) \\
&= \cos(f)\sin(h) + \sin(f)(\cos(h) - 1)
\end{aligned}$$

Now since both

$$\cos(h) - 1 \to 0$$

and

$$\frac{\sin(h)}{h} \to 1$$

as $||h|| \to 0$ our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = \cos(f(x))h(x)\,.$$

Part (e): For the functional $F(f) = \int_0^x t(f(t))^2 dt$ we have the following

$$\begin{aligned}
\int_0^x t(f(t)+h(t))^2 dt - \int_0^x t(f(t))^2 dt &= \int_0^x t\left((f(t)+h(t))^2 - (f(t))^2\right) dt \\
&= \int_0^x t\left(f^2 + 2fh + h^2 - f^2\right) dt \\
&= \int_0^x t\left(2fh + h^2\right) dt
\end{aligned}$$

From which we see that our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = 2\int_0^x tf(t)h(t)dt\,.$$

Part (f): For the functional $F(f) = f'(x)$ we have the following

$$f'(x) + h'(x) - f'(x) = h'(x)\,.$$

Thus our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = h'(x)\,.$$

Part (g): For the functional $F(f) = f(x)f'(x)$ we have the following

$$\begin{aligned}
(f+h)(f'+h') - f'f &= ff' + fh' + hf' + hh' - ff' \\
&= fh' + hf' + hh'\,.
\end{aligned}$$

Thus our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = f(x)h'(x) + h(x)f'(x) = (f(x)h(x))'\,.$$

Part (h): For the functional $F(f) = x\int_0^1 (f'(t))^2 dt$ we have the following

$$\begin{aligned}
x\int_0^1 (f'(t)+h'(t))^2 dt - x\int_0^1 (f'(t))^2 dt &= x\int_0^1 \left((f'(t)+h'(t))^2 - (f'(t))^2\right) dt \\
&= x\int_0^1 \left(f'^2 + 2f'h' + h'^2 - f'^2\right) dt \\
&= x\int_0^1 \left(2f'h' + h'^2\right) dt\,.
\end{aligned}$$

From which we see that our Frechet derivative of $F(f)$ is given by

$$F'(f)h \cdot (x) = 2x\int_0^1 f'(t)h'(t)dt\,.$$

Part (i): For the functional

$$F(f) = \int_0^x ((f(t))^2 + (f'(t))^2)dt$$

we have the following

$$\begin{aligned}
\Delta F &= \int_0^x \left((f+h)^2 + (f'+h')^2 - f^2 - f'^2\right) dt \\
&= \int_0^x \left(f^2 + 2fh + h^2 + f'^2 + 2f'h' + h'^2 - f^2 - f'^2\right) dt \\
&= 2\int_0^x \left(fh + f'h' + \frac{h^2}{2} + \frac{h'^2}{2}\right) dt\,.
\end{aligned}$$

showing that

$$F'(f)h \cdot (x) = 2\int_0^x (f(t)h(t) + f'(t)h'(t))dt$$

Part (j): For the functional $F(f) = \left(\int_0^x f(t)^2 dt\right)^2$ we have the following

$$\begin{aligned}
\Delta F &= \left(\int_0^x (f+h)^2 dt\right)^2 - \left(\int_0^x f(t)^2 dt\right)^2 \\
&= \left(\int_0^x (f+h)dt + \int_0^x f dt\right)\left(\int_0^x (f+h)dt - \int_0^x f dt\right)
\end{aligned}$$

using the algebraic identity that $a^2 - b^2 = (a-b)(a+b)$. Simplifying the individual expressions above we have ΔF given by

$$\begin{aligned}
\Delta F &= \left(\int_0^x (2f+h)dt\right)\left(\int_0^x h dt\right) \\
&= 2\left(\int_0^x f dt\right)\left(\int_0^x h dt\right) + \left(\int_0^x h dt\right)^2
\end{aligned}$$

showing that

$$F'(f)h \cdot (x) = 2\left(\int_0^x f dt\right)\left(\int_0^x h dt\right)$$

Exercise 2 - convergence of functional iterations

For the mapping S given by $f \to g$ where

$$g(x) = \frac{x}{8} + \int_0^x f(t)^2 dt\,,$$

to compute the derivative we begin with ΔS defined with respect to a perturbation h as

$$\Delta S \equiv S(f+h) - S(f) = \left(\frac{x}{8} + \int_0^x (f+h)^2 dt\right) - \left(\frac{x}{8} - \int_0^x f^2 dt\right) = \int_0^x (2fh + h^2)dt\,.$$

Which means that our Frechet derivative S' is given by

$$S'(f)h \cdot (x) = 2\int_0^x f(t)h(t)dt\,.$$

To compute the norm of this derivative (which is required if we want to prove convergence of functional iterations using the S mapping) we note that

$$||S'(f)||_\infty \leq 2x||f||_\infty \leq 2x||f||_\infty \leq 2(1)\frac{1}{4} = \frac{1}{2}\,.$$

So we expect our iteration to converge since the Frechet derivative is less than one. Now defining the ball centered at $f_0 = 0$ with radius of $\frac{1}{4}$ (in the infinity norm), for every point in this ball we have just proven that $||S'(f)||_\infty \leq 0.5 < 1$ and our iteration will converge to a point in that ball, from Theorem 4 presented in the book. As an aside, if we start our iteration with the zero function i.e. $f_0(x) = 0$ and get for $f_1(x)$, and $f_2(x)$ the following

$$\begin{aligned} f_1(x) &= \frac{x}{8} + \int_0^x 0\,dt = \frac{x}{8}\,. \\ f_2(x) &= \frac{x}{8} + \int_0^x \frac{t}{8} dt = \frac{x}{8} + \frac{x^2}{192}\,. \end{aligned}$$

In addition see the Mathematica notebook `chap_6_sec_4_prob_2.nb` for code to generate the iterated functions $f_n(x)$ and plots of their behavior. Specifically this code generates *analytic* iterations of the functions $f_n(x)$. In addition, to the analytic manipulations, the Matlab code `chap_6_sec_4_prob_2.m` iterates these functional equations *numerically.* Using either method one sees that the iterations converge very quickly, and in the limit produce a limiting function $f_\infty(x)$ that is very linear in appearance.

6.6: The Newton-Raphson method

Here we present a slightly different way of looking at the example discussed during this section of the book and work through the required algebra. We begin by recognizing that the modified Newton-Raphson method seeks to iterate the following expression

$$y_{n+1} = y_n - [f'(y_0)]^{-1} f(y_n)$$

where now $f(\cdot)$ can be a *functional* and the y_n can be *functions* that upon iteration will converge to a root of the functional equation $f(y_\infty) = 0$. We begin with an example where y_n are functions and $f(\cdot)$ is an differential operator. To apply the Newton-Raphson method to the differential equation

$$y'(x) + y(x)^2 = \frac{1}{(x+1)^2} \quad \text{with} \quad y(0) = 1\,.$$

We begin by converting the given problem to that of searching for a functional root. Specifically, we write the above as $f(y) = 0$ by considering our functional operator $f(y)$ as

$$f(y) \equiv y'(x) + y(x)^2 - \frac{1}{(1+x)^2}\,, \tag{43}$$

which would be zero if a solution was found. To perform Newton-Raphson iterations, we need to be able to compute the derivative of this operator. We compute this derivative by looking at $\Delta f \equiv f(y+h) - f(y)$ which in this specific case becomes

$$\begin{aligned} \Delta f &\equiv \left(y' + h' + y^2 + 2yh + h^2 - \frac{1}{(1+x)^2}\right) - \left(y' + y^2 - \frac{1}{(1+x)^2}\right) \\ &= h' + 2hy + h^2\,. \end{aligned}$$

So dropping the nonlinear terms in h, we see that the derivative is defined as $f'(y)h\cdot(x) = h'(x) + 2y(x)h(x) \equiv k(x)$. The modified Newton-Raphson method requires computing this derivative *at* a specific fixed point. Now with our specific fixed point being the solution to the homogeneous equation $y' + y^2 = 0$ with $y(0) = 1$, we have $y_0 = \frac{1}{x+1}$. In this case then k becomes

$$k(x) = h'(x) + \frac{2h(x)}{x+1}$$

which is Equation 7 in the book. Now specifically $k(x)$ is the derivative of our functional $f(y)$ evaluated at $y = y_0$. For our Newton-Raphson iterations we require the inverse of this operator $f(y_0)^{-1}$. Thus we we are looking to invert the above operator i.e. given $k(x)$ (or the derivative) compute $h(x)$ (or the input). Specifically we want to solve for the perturbation $h(x)$ in

$$h'(x) + \frac{2h(x)}{x+1} = k(x)\,.$$

To find this perturbation $h(x)$ by multiplying by $(x+1)^2$ the above becomes

$$(x+1)^2 h'(x) + (x+1)h(x) = (x+1)^2 k(x)\,,$$

where we recognize the left hand side of this expression as the derivative with respect to x of the expression $(x+1)^2 h(x)$. Using this simplification gives

$$\frac{d}{dx}((x+1)^2 h) = (x+1)^2 k(x)$$

which can be integrated and gives

$$(x+1)^2 h(x) = \int_0^x (s+1)^2 k(s) ds\,. \tag{44}$$

This is Sawyer Eq. 8 and represents the *inverse* of our functional operator $[f(y_0)]^{-1}$. We can now evaluate the right hand side of our iteration

$$y_n - [f'(y_0)]^{-1} f(y_n)\,,$$

for which we obtain (by using the definition of $f(y)$ given in Eq. 43) above

$$y_{n+1} = y_n - [f'(y_0)]^{-1} \left(y_n' - y_n^2 - \frac{1}{(1+x)^2} \right)$$

or using the definition of $[f'(y_0)]^{-1}$ that we worked so hard to obtain and given by Eq. 44 we have functional iterations given by

$$y_{n+1}(x) = y_n(x) - \frac{1}{(x+1)^2} \int_0^x \left((s+1)^2 \left(y_n'(s) + y_n(s)^2 - \frac{1}{(1+s)^2} \right) \right) ds\,.$$

In general the operations of integration and differentiation are inverses so we can simplify the above by applying integration by parts to the first term in the above integral. This gives us

$$y_{n+1}(x) = y_n(x) - \frac{1}{(x+1)^2} \left((s+1)^2 y_n(s)\Big|_0^x - \int_0^x 2(s+1) y_n(s) ds + \int_0^x (s+1)^2 y_n(s)^2 ds - x \right)$$

since $y_n(0) = 1$ for all n, we can simplify the first integral and obtain

$$\begin{aligned} y_{n+1}(x) &= y_n(x) - \frac{1}{(x+1)^2} \left((x+1)^2 y_n(x) - 1 - \int_0^x 2(s+1) y_n(s) ds + \int_0^x (s+1)^2 y_n(s)^2 ds - x \right) \\ &= \frac{1}{(x+1)^2} + \frac{2}{(x+1)^2} \int_0^x (s+1) y_n(s) ds - \frac{1}{(x+1)^2} \int_0^x (s+1)^2 y_n(s)^2 ds + \frac{x}{(x+1)^2} \\ &= \frac{1}{x+1} + \frac{2}{(x+1)^2} \int_0^x (s+1) y_n(s) ds - \frac{1}{(x+1)^2} \int_0^x (s+1)^2 y_n(s)^2 ds\,. \end{aligned}$$

This last equation is Sawyer Eq. 10. We can now perform the algebra needed for the functional iterations, with $y_0 = \frac{1}{x+1}$, we have from the above iterations $y_1(x)$ given by

$$\begin{aligned} y_1(x) &= \frac{1}{x+1} + \frac{2}{(x+1)^2} \int_0^x ds - \frac{1}{(x+1)^2} \int_0^x ds \\ &= \frac{1}{x+1} + \frac{2x}{(x+1)^2} - \frac{x}{(x+1)^2} \\ &= \frac{1}{x+1} + \frac{x}{(x+1)^2}\,. \end{aligned}$$

Continuing for $y_2(x)$ we obtain the following

$$y_2(x) \quad = \quad \frac{1}{x+1} + \frac{1}{(x+1)^2}\left(\int_0^x 2\left(1+\frac{s}{s+1}\right)ds - \int_0^x (s+1)^2\left(\frac{1}{s+1}+\frac{s}{(s+1)^2}\right)^2 ds\right)$$

which would need to be evaluated to compute $y_2(x)$. Using Mathematica to evaluate this integral we have for $y_2(x)$ the following

$$\begin{aligned} y_2(x) &= \frac{1}{x+1} + \frac{1}{(x+1)^2}\left(-1+\frac{1}{1+x}+2\log(1+x)\right) \\ &= \frac{1}{x+1} + \frac{2\log(x+1)}{(x+1)^2} - \frac{x}{(x+1)^3}\,. \end{aligned}$$

Which is Sawyer Eq. 13. In the general case rather than perform these integrations by hand we will use Mathematica to evaluate these expressions. In the Mathematica notebook given by `chap_6_sec_6_page_111.nb`, a program is given that generates these iterates for an arbitrary n is provided.

In general the above procedure can be applied either analytically or numerically in an attempt to find the fixed point of the given mapping. It was Kantorovich that proved that the decision as to whether or not the above procedure converges can be answered by considering the convergence of a much simpler function of the real variables. This allow the numerical analysis to study the properties of this one dimensional function and if *it* converges to proceed to the more complicated problem with out worry, since the computed solution is then guaranteed to converge. We next derive the needed results for this problem to prove convergence using Kantorovich's arguments.

$$y_{n+1} = \frac{1}{x+1} + \frac{1}{(x+1)^2}\int_0^x \left(2(s+1)y_n(s) - (s+1)^2 y_n(s)^2\right) dt \equiv S(y)$$

so computing the ΔS, we have

$$\begin{aligned} S(y+h) - S(y) &= \frac{1}{(x+1)^2}\int_0^x \left(2(s+1)(y+h) - (s+1)^2(y+h)^2 - 2(s+1) + (s+1)^2y^2\right) ds \\ &= \frac{1}{(x+1)^2}\int_0^x \left(2(s+1)h(s) - (s+1)^2(y^2+2yh+h^2-y^2\right) ds \\ &= \frac{1}{(x+1)^2}\int_0^x \left(2(s+1)h(s) - 2(s+1)^2(2yh+\frac{h^2}{2}\right) ds \\ &\to \frac{1}{(x+1)^2}\int_0^x \left(2(s+1) - 2(s+1)^2y(s)\right) h(s)ds \equiv S'(y)h\cdot(x) \end{aligned}$$

To determine the convergence properties of this iterative scheme we want to compute the norm of its derivative defined by

$$||S'(y)||_\infty \equiv \mathrm{sup}_{||h||=1}||S'(y)h\cdot(x)||_\infty = \mathrm{sup}_x\left\{\frac{1}{(x+1)^2}\int_0^x \left|2(s+1) - 2(s+1)^2y(s)\right| ds\right\}\,.$$

At this point it is still up to the mathematician to construct a function ϕ that will satisfy the nessesary arguments of Kantrovich's. Once an appropriate ϕ has been specified we are gaurenteed that our functional iterations will converge at least as fast as that of the one dimensional iterations $t \to \phi(t)$. To determine ϕ we will use the *second* Kantorovich requirement that if $||y-y_0|| \le t$ then this implys that $||S'(y)|| \le \phi'(t)$ to determine a differential equation for $\phi(t)$ which we can then solve. To start this process, for this example lets take our initial function to be the solution y_0 to the homogeneous equation. That is $y_0(s) = \frac{1}{1+s}$. Then the general y norm relationship

$$||y-y_0|| \le t\,,$$

is obviously equivalent in the L_∞ norm to the absolute value expression

$$\left| y - \frac{1}{s+1} \right| \le t \,.$$

This provides a bound on the expression involving the unknown $y(s)$ in the expression in $S'(y)$. Specifically, by multiplying the above by $2(s+1)^2$ we obtain

$$\left| 2(s+1) - 2(s+1)^2 y(s) \right| \le t \,,$$

which is exactly the expression containing $y(s)$ found in $||S'(y)||$. Inserting this expression into the form for $||S'(y)||$ and integrating we have

$$\begin{aligned} ||S'(y)||_\infty &\le \sup_x \left\{ \frac{1}{(x+1)^2} \cdot 2t \cdot \frac{((x+1)^3 - 1)}{3} \right\} \\ &\le \sup_x \left\{ \frac{1}{(x+1)^2} \cdot 2t \cdot \frac{(x+1)^3}{3} \right\} = \frac{2t}{3} \sup_x \left\{ x+1 \right\} \,. \end{aligned}$$

The above is linear in x. To develop a bound lets assume that x is bounded i.e. lets take $x \in [0, a]$ we have that $||S'(y)||_\infty$ is bounded above by

$$\frac{2t(a+1)}{3} \,.$$

Since we can *pick* any function $\phi(t)$ as long as it satisfies the inequality that $||S'(y)|| \le \phi'(t)$, lets take the $\phi'(t)$ that *equals* the above, i.e. take ϕ to be the solution to

$$\phi'(t) = \frac{2t(a+1)}{3} \,.$$

The initial condition for this differential equation is given by the *first* Kantorovich requirement that the first jump in y is not larger than the first jump in ϕ. Since our iterations in y start with y_0 our first jump in y is given by $|y_1 - y_0|$, while since our iterations in t start with zero the first jump in ϕ is $\phi(0)$. Thus the "first" Kantorovich requirement is that

$$|y_1 - y_0| \le \phi(0)$$

Again, by taking $\phi(0)$ *equal* to the norm above we certainly have the above to hold. As shown in the book this gives $\phi(0) = \frac{a}{(a+1)^2}$. With this initial condition, we can integrate the differential equation above to obtain ϕ of

$$\phi(t) = \frac{t^2(a+1)}{3} + \frac{a}{(a+1)^2} \,.$$

After all of this work we now can conclude our search. We have found a function who convergence properties (in one dimension) provides qualitative information on the more general mapping $y \to S(y)$.

To determine how well the iterative mapping $y \to S(y)$ performs in practice we need to have access to the exact solution to our original differential equation. In what follows we will derive this solution. As suggested in the book we can solve our original differential equation

$$y'(x) + y(x)^2 = 1/(x+1)^2 \,,$$

by letting $y(x) = z(x)/(1+x)$ and solving for $z(x)$. With this substitution that our derivatives of $y(x)$ in terms of the function of $z(x)$ become

$$\begin{aligned} y(x) &= \frac{z(x)}{1+x} \\ y'(x) &= \frac{z'(x)}{1+x} - \frac{z(x)}{(1+x)^2} \,. \end{aligned}$$

When these are put into our original equation gives

$$\frac{z'(x)}{1+x} - \frac{z(x)}{(1+x)^2} + \frac{z(x)^2}{(1+x)^2} = \frac{1}{(x+1)^2}\,.$$

When we multiply both sides by $(x+1)^2$ we obtain

$$(1+x)z'(x) - z(x) + z(x)^2 = 1$$

or

$$(1+x)\frac{dz}{dx} = z(x) - z(x)^2 + 1$$

or by placing all z terms on the left and all x terms on the right we have

$$\frac{-dz}{z^2 - z - 1} = \frac{dx}{x+1}\,,$$

which is now in a form where both sides can be integrated.

Exercise 1 (a two dimensional example)

For this problem we wish to study the roots of the two dimensional function $f(v) = 0$ determined by

$$f\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} xy + 0.07 \\ x^2 + y^2 - 0.41 \end{array}\right)\,.$$

With a starting location of $(x_0, y_0) = (0.1, -0.6)$. In this exercise we are looking for two things. The first is some practical experience at how such a mapping may behave and the second is a one dimensional function ϕ that provides convergence information on this more complicated mapping. Rather than use the direct functional iteration provided above (see the later part of this problem e.g. Section 6.4 from the book) we will use the modified Newton-Raphson method and from it derive the operator $S(\cdot)$ needed for the application of Kantorovich's arguments to compute our $\phi(t)$ function.

The constant slope or modified Newton-Raphson method for finding the zero of the above mapping gives the following iterations

$$v_{n+1} = v_n - (f'(v_0))^{-1} f(v_n)\,,$$

where v_n is a two dimensional vector of (x_n, y_n). For this problem the derivative of our mapping $f(\cdot)$ is given by

$$f'\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{cc} y & x \\ 2x & 2y \end{array}\right)\,.$$

Evaluating $f'(\cdot)$ at the initial point $(x_0, y_0) = (0.1, -0.6)$ we have that

$$f'\left(\begin{array}{c} x_0 \\ y_0 \end{array}\right) = \left(\begin{array}{cc} -6\cdot 10^{-1} & 10^{-1} \\ 2\cdot 10^{-1} & -12\cdot 10^{-1} \end{array}\right) = \frac{1}{10}\left(\begin{array}{cc} -6 & 1 \\ 2 & -12 \end{array}\right)\,.$$

The inverse of this expression is needed for the modified Newton-Raphson iterations and is given by

$$\left(f'\left(\begin{array}{c} x_0 \\ y_0 \end{array}\right)\right)^{-1} = \frac{10}{6\cdot 12 - 2}\left(\begin{array}{cc} -12 & -1 \\ -2 & -6 \end{array}\right) = \frac{-1}{7}\left(\begin{array}{cc} 12 & 1 \\ 2 & 6 \end{array}\right)\,.$$

With all of these pieces the modified Newton-Raphson iterations then become (and defines our operator $S(\cdot)$)

$$\begin{aligned}
\left(\begin{array}{c} x_{n+1} \\ y_{n+1} \end{array}\right) &= \left(\begin{array}{c} x_n \\ y_n \end{array}\right) + \frac{1}{7}\left(\begin{array}{cc} 12 & 1 \\ 2 & 6 \end{array}\right)\left(\begin{array}{c} x_n y_n + 7\cdot 10^{-2} \\ x_n^2 + y_n^2 - 41\cdot 10^{-2} \end{array}\right) \\
&= \left(\begin{array}{c} x_n \\ y_n \end{array}\right) + \frac{1}{7}\left(\begin{array}{c} 12 x_n y_n + 84\cdot 10^{-2} + x_n^2 + y_n^2 - 41\cdot 10^{-2} \\ 2 x_n y_n + 14\cdot 10^{-2} + 6x_n^2 + 6y_n^2 - 246\cdot 10^{-2} \end{array}\right) \\
&= \left(\begin{array}{c} x_n \\ y_n \end{array}\right) + \frac{1}{7}\left(\begin{array}{c} x_n^2 + y_n^2 + 12 x_n y_n + 43\cdot 10^{-2} \\ 6x_n^2 + 6y_n^2 + 2x_n y_n - 232\cdot 10^{-2} \end{array}\right) \equiv S\left(\begin{array}{c} x_n \\ y_n \end{array}\right)\,.
\end{aligned}$$

With these precalculations done we can now apply the two conditions of Kantrovich's to determine a scalar function ϕ, the behavior of which will give quantitative understanding of the iterates of the modified Newton-Raphson iterates above. I'll begin with Kantrovich's *first* condition that the first jump in y must be smaller than the first jump in ϕ, i.e.

$$||v_1 - v_0|| < \phi(0)$$

Using the provided starting guess, and the above we easily calculate that the first iterate v_1 of the modified Newton-Raphson method is given by

$$\left(\begin{array}{c} x_1 \\ y_1 \end{array} \right) = \left(\begin{array}{c} 0.114 \\ -0.6314 \end{array} \right)$$

Thus we have that (when using the l_∞ norm) our distance $||v_1 - v_0||$ is given by

$$||v_0 - v_1|| = \left|\left| \left(\begin{array}{c} 0.1 \\ -0.6 \end{array} \right) - \left(\begin{array}{c} 0.114 \\ -0.6314 \end{array} \right) \right|\right|_\infty = \left|\left| \left(\begin{array}{c} -0.0114 \\ +0.0314 \end{array} \right) \right|\right|_\infty = 0.0314\,.$$

We can ensure that $\phi(0)$ is less than or equal to this value by taking it *equal* to this or slightly larger. Therefore we can let $\phi(0) = 0.0315$, be our initial condition on our function $\phi(t)$.

To apply Kantrovich's *second* condition, we require that if $||v - v_0|| \le t$, then $||S'(v)|| \le \phi'(t)$. With this definition of the operator S we can compute a derivative (for the purposes of applying Kantorovich's second condition) as follows

$$S' \left(\begin{array}{c} x \\ y \end{array} \right) = I + \frac{1}{7} \left(\begin{array}{cc} 2x + 12y & 2y + 12x \\ 12x + 2y & 12y + 2x \end{array} \right).$$

Here I is the two by two identity matrix. So our derivative of S evaluated at $v \equiv (x, y) = (0.1 + X, -0.6 + Y)$ (after some simplification) is given by

$$S' \left(\begin{array}{c} 0.1 + X \\ -0.6 + Y \end{array} \right) = \frac{1}{7} \left(\begin{array}{cc} 2X + 12Y & 2Y + 12X \\ 12X + 2Y & 12Y + 2X \end{array} \right)$$

Note that in the above we have introduced the *offset* variables X and Y denoting how far from the base point v_0 our function evaluation is taken. The benefit of introducing these variable is that the second Kantorvich's criterion of $||v - v_0|| \le t$ then becomes an explicit relationship in terms of X and Y, for example we see that it is

$$||v - v_0|| = || \left(\begin{array}{c} X \\ Y \end{array} \right) ||_\infty = \max(|X|, |Y|) \le t\,.$$

To see how a condition like this will appear in the norm of the derivative S consider

$$\begin{array}{rcl} ||S' \left(\begin{array}{c} 0.1 + X \\ -0.6 + Y \end{array} \right) ||_\infty & = & \frac{1}{7}(|2X + 12Y| + |2Y + 12X|) \\ & \le & \frac{1}{7}(2|X| + 12|Y| + 2|Y| + 12|X|) \\ & \le & \frac{1}{7}(2t + 12t + 2t + 12t) = 4t\,. \end{array}$$

Since $\max(|X|, |Y|) \le t$ implys both that $|X| \le t$ and $|Y| \le t$. Thus we can gaurentee that $\phi'(t)$ will be larger or equal to this if we take it equal. Thus we have that the function $\phi(t)$ satisfies both of Kantrovich's if we take

$$\phi'(t) = 4t \quad \text{and} \quad \phi(0) = 0.0315\,.$$

Integrating the above we have that $\phi(t) = 0.0315 + 2t^2$ as claimed in the book. We can now use these results to carry out all iterations and compare how well our bounding function ϕ predicts the behavior of the two dimensional iterations defined above. The exact solution to the expression $t = \phi(t)$ is easy to compute with the quadratic equation and provides the limit of the ϕ iterations.

In addition to the modified Newton-Raphson iterations (defined above) in Section 6.4 of the book different iterations seeking the same roots were performed by *directly iterating* the mapping $v_{n+1} = f_{\rm fp}(v_n)$ where $f_{\rm fp}(\cdot)$ is defined as

$$f_{\rm fp}\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} xy + x + 0.07 \\ x^2 + y^2 + y - 0.41 \end{array}\right) .$$

Notice that the difference between the function $f_{\rm fp}$ and f defined earlier is that the first component of the output of the $f_{\rm fp}$ function has an additional x while the second component of the ouput of our $f_{\rm fp}$ function has a additional y. To evaluate the efficiency of this mapping and the modified Newton-Raphson mapping we recognize that the exact solutions of both iterative scheme is given by the solution to the system

$$\begin{array}{rcl} xy &=& -0.07 \\ x^2 + y^2 &=& 0.41 \,, \end{array}$$

which upon eliminating y gives a quartic in x of

$$x^4 - 0.41x^2 + 0.0049 = 0 .$$

In `chap_6_sec_6_prob_1.m` we tabulate the error versus iteration number for both two dimensional iterations and our one dimensional qualitative iteration. We note that the iterations of the ϕ function in the form of $t_{n+1} = \phi(t_n)$ converge *slower* than the modified Newton-Raphson method as would be expected from the discussion in the text.

Exercise 2 (another two dimensional example)

We desire to fine the $v \equiv (x, y)$ such that $f(v) = 0$, where the function $f(\cdot)$ is defined by

$$f\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} x^2 + y^2 - 200 \\ y^3 + xy - x^3 \end{array}\right) .$$

Which we will do by using the modified Newton-Raphson method where we construct a sequence of iterates as

$$v_{n+1} = v_n - [f'(v_0)]^{-1} f(v_n) ,$$

and in the limit $n \to \infty$ we find the solution. Notationally, this right hand side is defined as the function $S(v_n)$, by Kantorovich. The above expression involves the derivative of f, taking this derivative for the specific function considered here we have

$$f'\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{cc} 2x & 2y \\ y - 3x^2 & 3y^2 + x \end{array}\right) .$$

Evaluating the derivative f' at the given center point $(x_0, y_0) = (10, 10)$, we find

$$f'(x_0, y_0) = \left(\begin{array}{cc} 20 & 20 \\ -290 & 310 \end{array}\right) .$$

Computing the matrix inverse of this derivative (as required for the above iterations) we have

$$\begin{array}{rcl} N &\equiv& [f'(x_0, y_0)]^{-1} \\ &=& \dfrac{1}{20 \cdot 310 + 20 \cdot 290} \left(\begin{array}{cc} 310 & -20 \\ 290 & 20 \end{array}\right) \\ &=& \dfrac{1}{12000} \left(\begin{array}{cc} 310 & -20 \\ 290 & 20 \end{array}\right) . \end{array}$$

We now have everything to evaluate the modified Newton-Raphson iterations used to define a numerical iterative scheme that will hopefully step towards our root. From the definition of N given above we see that our function $S(\cdot)$ is defined as in terms of (x, y) as

$$S\left(\begin{array}{c} x \\ y \end{array}\right) = \left(\begin{array}{c} x \\ y \end{array}\right) - N\left(\begin{array}{c} x^2+y^2-200 \\ y^3+xy-x^3 \end{array}\right).$$

The derivative of S is then seen to be

$$S'\left(\begin{array}{c} x \\ y \end{array}\right) = I - N\left(\begin{array}{cc} 2x & 2y \\ y-3x^2 & 3y^2+x \end{array}\right),$$

as claimed in the text. Here I is the two by two identity matrix, and we used the fact that $N = [F'(v_0)]^{-1}$ is a constant matrix (independent of x and y) in taking this derivative. In general, we can evaluate S' at the centering point (x_0, y_0) where we find that

$$S'(v_0) = I - [f'(v_0)]^{-1}f'(v_0) = I - I = 0.$$

Thus the result that $S'(v_0) = 0$ must hold in general. This implys that convergence properties of the iterations $v_{n+1} = S(v_n)$ must be very good since the norm of the zero matrix is certainly less than one, the threshold required for convergence. Now computing $S'(10+X, 10+Y)$ we find that

$$S'\left(\begin{array}{c} 10+X \\ 10+Y \end{array}\right) = \frac{1}{1200}\left(\begin{array}{cc} -182X-6X^2+2Y & 2X+58Y+6Y^2 \\ 62X+6X^2-2Y & -2X-178Y-6Y^2 \end{array}\right).$$

The algebra for this calculation is done in the Mathematica file `chap_6_sec_6_prob_2.nb`. Note that for a simple check on our algebra we must have that $S'(v_0) = 0$, which we can see is true in the above formula by setting both X and Y to zero. To apply Kantorovitchs second criterion (the one on the bound on the derivatives) we need to compute $||S'(v)||$ so that we can select an appropriate bound for $\phi'(t)$, such that if $||v - v_0|| \le t$ then this would imply that $||S'(v)|| < |\phi'(t)|$. We will use the infinity norm which for matrices implys that we sum the absolute value of the elements in each row and then take the maximum over the rows. For the matrix defined by $S'(v)$ above with a centering point of $v_0 = (10, 10)$ we can use the triangle inequality for the absolute value function (i.e. the fact that $|a - b| < |a| + |b|$) to show that

$$\begin{aligned} ||S'\left(\begin{array}{c} 10+X \\ 10+Y \end{array}\right)||_\infty &\le \frac{1}{1200}\max(12t^2+244t, 12t^2+244t) \\ &= \frac{1}{100}t^2 + \frac{61}{300}t. \end{aligned}$$

We can enforce that $\phi'(t)$ is larger than or equal to this value if we take it *equal* to this value. Thus we can set

$$\phi'(t) = \frac{1}{100}t^2 + \frac{61}{300}t,$$

which is a differential equation that could be integrated once we can come up with an initial condition. The the ϕ function is something that the mathematician creates its only requirement is that its derivative be larger than that of the derivative of S. Thus we are at liberty to take any function ϕ that has a derivative *greater* than this value also. Since

$$\frac{61}{300} \approx \frac{1}{5} < \frac{1}{3},$$

we could also take as a definition of our ϕ function a function that satisfies

$$\phi'(t) = \frac{1}{100}t^2 + \frac{1}{3}t.$$

The initial condition for the ϕ function is provided by Kantorovich's first condition which requires that the first step of the functional iterations is *smaller* than the first step of the ϕ iterations. Mathematically this requires

$$||v_1 - v_0|| < |\phi(0)|.$$

To determine the value of $||v_1 - v_0||$ we produce v_1 by one step of our modified Newton-Raphson iterate. When this is done we find for $||v_1 - v_0||_\infty$ that

$$||v_0 - v_1|| = \left|\left| \begin{pmatrix} 10 \\ 10 \end{pmatrix} - \begin{pmatrix} 10.1667 \\ 9.8333 \end{pmatrix} \right|\right|_\infty = \left|\left| \begin{pmatrix} -0.1667 \\ 0.1667 \end{pmatrix} \right|\right|_\infty = 0.1667 < \frac{1}{6}\,.$$

Thus our inequality will be true if we take $\phi(0) = \frac{1}{6}$. Integrating this differential equation for ϕ we find that

$$\phi(t) = \frac{1}{6} + \frac{t^2}{6} + \frac{t^3}{300}\,.$$

Iterations for both the modified Newton-Raphson method and the function ϕ can be found in the Matlab function `chap_6_sec_6_prob_2.m`. There one see the behavior that Kantorovich proved. That the nth link in the chain of v points $||v_{n+1} - v_n||$ never exceeds the nth link in the chain of t point $|t_{n+1} - t_n|$. This is especially apparent in the plots of the nth link in each chain which is produced by the Matlab code.

Exercise 3 (a linear operator plus a nonlinear part)

We desire to compute the v that satisfies $f(v) = 0$ where $f(v)$ can be explicitly decomposed into a linear term and a nonlinear terms as

$$f(v) = Mv + g(v)\,.$$

To use the constant slope Newton-Raphson method one iterates

$$v_{n+1} = v_n - [f'(v_0)]^{-1} f(v_n)\,.$$

Here in this problem we can see that the derivative of $f(v)$ is also given by a linear part plus a nonlinear part as

$$f'(v) = M + g'(v)\,.$$

Since we are told that $g'(v_0) = 0$, we see that $f'(v_0) = M$. So the constant slope iterations then simplify to

$$v_{n+1} = v_n - M^{-1} f(v_n) = v_n - M^{-1}(Mv_n + g(v_n)) = -M^{-1} g(v_n)\,.$$

This right hand side defines the operator $S(v)$, we see that

$$S(v) \equiv -M^{-1} g(v)\,,$$

as claimed in the text. For the specific function suggested

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -37x + 9y + x^5 + y^5 + 25 \\ 4x - 28y + x^3y^3 + 18 \end{pmatrix},$$

we note that this function can be written as a linear part and a nonlinear remainder. This decomposition then defines the linear mapping M and the nonlinear function $g(v)$ introducted above. Specifically for this function we have

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -37 & 9 \\ 4 & -28 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x^5 + y^5 + 25 \\ x^3y^3 + 18 \end{pmatrix} \equiv Mv + g(v)\,.$$

Then our constant slope iterates are given by $v_{n+1} = S(v_n)$ where $S(v) = -M^{-1}g(v)$ is given by

$$\begin{aligned} S(v) &= \frac{-1}{37 \cdot 28 - 36} \begin{pmatrix} -28 & -9 \\ -4 & -37 \end{pmatrix} \begin{pmatrix} x^5 + y^5 + 25 \\ x^3y^3 + 18 \end{pmatrix} \\ &= \frac{1}{1000} \begin{pmatrix} 28 & 9 \\ 4 & 37 \end{pmatrix} \begin{pmatrix} x^5 + y^5 + 25 \\ x^3y^3 + 18 \end{pmatrix}. \end{aligned}$$

With this information we can take a derivative of $S(v)$, which is required for the arguments of Kantrovitch (specifically the second rate of change requirement). We find that

$$
\begin{aligned}
S'(v) &= \frac{1}{1000}\begin{pmatrix} 28 & 9 \\ 4 & 37 \end{pmatrix}\begin{pmatrix} 5x^4 & 5y^4 \\ 3x^2y^3 & 3x^3y^2 \end{pmatrix} \\
&= \frac{1}{1000}\begin{pmatrix} 140x^4 + 27x^2y^3 & 140y^4 + 27x^3y^2 \\ 20x^4 + 111x^2y^3 & 20y^4 + 111x^3y^2 \end{pmatrix} .
\end{aligned}
$$

So the norm of this derivative is given by

$$
||S'(v)||_\infty = \frac{1}{1000}\max\left\{|140x^4 + 27x^2y^3| + |140y^4 + 27x^3y^2|, |20x^4 + 111x^2y^3| + |20y^4 + 111x^3y^2|\right\} .
$$

Since our starting point v_0 is zero, we want to find bounds for this derivative in the case when $||v - v_0||_\infty = ||v||_\infty = \max(|X|, |Y|) \leq t$, we can use the triangle inequality in the above to further bound $S'(v)$. We then have

$$
\begin{aligned}
||S'(v)||_\infty &\leq \frac{1}{1000}\max\left\{140t^4 + 27t^5 + 140t^4 + 27t^5, 20t^4 + 111t^5 + 20t^4 + 111t^5, \right\} \\
&= \frac{1}{1000}\max\left\{280t^4 + 54t^5, 40t^4 + 222t^5\right\} \\
&< \frac{1}{1000}(280t^4 + 222t^5) .
\end{aligned}
$$

Where in obtaining the last inequality we have selected from each polynomial the term with the largest coefficient. Thus we will have $\phi'(t)$ greater than or equal to $||S'(v)||$ is we take $\phi'(t)$ equal to this value. Thus we will solve

$$
\phi'(t) = \frac{1}{1000}(280t^4 + 222t^4) .
$$

The initial condition for the ϕ function is provided by Kantorovich's first condition which requires that the first step of the functional iterations is *smaller* than the first step of the ϕ iterations. Mathematically this requires

$$
||v_1 - v_0|| < |\phi(0)| .
$$

To determine the value of $||v_1 - v_0||$ we produce v_1 by one step of our modified Newton-Raphson iterate. When this is done we find for $||v_1 - v_0||_\infty$ that

$$
||v_0 - v_1|| = \left|\left|\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0.8620 \\ 0.7660 \end{pmatrix}\right|\right|_\infty = \left|\left|\begin{pmatrix} -0.8620 \\ -0.7660 \end{pmatrix}\right|\right|_\infty = 0.8620 .
$$

Thus our inequality will be true if we take $\phi(0) = 0.862$. Integrating this differential equation for ϕ we find that

$$
\phi(t) = 0.862 + \frac{280}{5000}t^5 + \frac{222}{6000}t^6 .
$$

We note that this is the same function ϕ obtained in the text. The required iterations for this problem are carried out in the Matlab file `chap_6_sec_6_prob_3.m`. Again with these iterations we see that the link size in the t steps is always greater than or equal to the link size of the v steps. The text then states that for the relevant values of t an improved bound can be found. This is because since when $t \approx 0.93$ one of the functions in the computation of the norm of $S'(v)$ dominates the other. Specifically for $0 < t \leq 1.3$ we have that

$$
\frac{1}{1000}\max\left\{280t^4 + 54t^5, 40t^4 + 222t^5\right\} = \frac{1}{1000}(280t^4 + 54t^5) ,
$$

And the calculations could be repeated with this for the value of $\phi'(t)$. This would give a function $\phi(t)$ defined as

$$
\phi(t) = 0.862 + \frac{280}{5000}t^5 + \frac{54}{6000}t^6 .
$$

Iterations with this function are also shown in the Matlab code above. There one can see that the better the bound we use, the better the iterations of our ϕ function approximate that of the true iterates $v_{n+1} = S(v_n)$.

Exercise 4 (an example from integral equations)

We desire to study the functional mapping $z = S(y)$ given by

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

This operator S starting with an initial function $y_0(x)$ defines a sequence of iterate functions through $y_{n+1} = S(y_n)$, which specifically in this case is given by

$$y_{n+1}(x) = \int_0^x (y_n(t) + t)^2 dt \,.$$

We will use Kantorovichs arguments to study the convergence properties of this iteration sequence. To do so we now seek a function ϕ whos behavior will bound the qualitative behavior of the more general functional iterations $v_{n+1} = S(v_n)$. To find this function $\phi(t)$, we seek a scalar function that has properties related to S when we perform ϕ iterations $t_{n+1} = \phi(t_n)$, beginning with $t_0 = 0$. The properties that ϕ must have are

- The distance between the starting point y_0, and the first iterate $y_1 \equiv S(y_0)$ of the functional mapping is less than the distance between the start of the ϕ iterations ($t_0 = 0$) and the first step $t_1 = \phi(0)$ or

$$||y_1 - y_0|| \le |\phi(0) - 0| = |\phi(0)| \,.$$

- For all points y less than t in distance from y_0, the rate of change of the functional operator S is less than that of the function ϕ or

$$\text{if} \quad ||y - y_0|| \le t \quad \text{then} \quad ||S'(y)|| \le ||\phi'(t)|| \,.$$

We will begin calculating our ϕ with the first condition discussed above. As suggested in the book let $y_0 = 0$ and then $y_1(t)$ is given by

$$y_1(t) = \int_0^x t^2 dt = \frac{x^3}{3} \,.$$

Then the distance between y_1 and y_0 is then given by

$$||y_1 - y_0||_\infty = ||\frac{1}{3}x^3||_\infty = \frac{a^3}{3}$$

where we have evaluated this norm under the restriction that $x < a$. Using this result we can guarantee the first condition above by defining $\phi(0) = \frac{a^3}{3}$. Now the second condition above requires the derivative of our mapping functional $S(y)$ which we compute by taking the following "forward difference"

$$\begin{aligned} S(y+h) - S(y) &= \int_0^x ((y+h+t)^2 - (y+t)^2) dt \\ &= \int_0^x (2(y+t)h + h^2) dt \\ &\to 2\int_0^x (y+t)\, h\, dt \quad \text{as} \quad h \to 0 \,. \end{aligned}$$

Showing that the derivative of S is given by

$$S'(y)h \cdot (x) = 2\int_0^x (y(t) + t)\, h\, dt \,.$$

So the norm of this operator is given by

$$||S'(y)|| = 2 \sup \left\{ \int_0^x |y(t) + t| dt \right\} \,.$$

Since $y_0 = 0$ the second Kantrovich's condition becomes if $||y||_\infty \leq t$ then this implys $||S'(y)|| \leq \phi'(t)$. Using the fact that $||y||_\infty \leq t$, we have that

$$\begin{aligned} ||S'(y)|| &\leq 2\sup_{x\in[0,a]} \int_0^x (s+t)dt \\ &= 2\sup_{x\in[0,a]}(st + \frac{t^2}{2}|_0^x \\ &= 2\sup_{x\in[0,a]}(sx + \frac{x^2}{2}) \\ &= 2(sa + \frac{a^2}{2}) = 2sa + a^2 \,. \end{aligned}$$

Then we will satisfy Kantrovich's second condition if we take $\phi'(t)$ *equal* to this expression. Thus in summary we have

$$\phi'(t) = 2ta + a^2 \quad \text{with} \quad \phi(0) = \frac{a^3}{3} \,.$$

Integrating this expression we obtain

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

for an expression for the comparison function.

We now desire to show that this iteration scheme $v_{n+1} = S(v_n)$ for S defined above arises from an iterative procedure for solving a differential equation. To do this consider the differential equation given by

$$\frac{du}{dx} - (x+u)^2 = 0 \quad \text{with} \quad u(0) = 0 \,.$$

The modified Newton-Raphson iteration method for looking for the roots of $F(u) = 0$ is given by the following iterations

$$u_{n+1} = u_n - [F'(u_0)]^{-1} F(u_n) \,.$$

For a functional F defined as $F(u) = \frac{du}{dx} - (x+u)^2$, we needed to compute what $F'(u)$ would be. We again consider a "forward difference" of the functional F to obtain

$$\begin{aligned} F(u+h) - F(u) &= \frac{d(u+h)}{dx} - (x+u+h)^2 - \frac{du}{dx} + (x+u)^2 \\ &= \frac{dh}{dx} - 2(x+u)h - h^2 \,. \end{aligned}$$

Upon taking limits $h \to 0$ we have that our Frechet derivative is given by

$$F'(u)h \cdot (x) = \frac{dh}{dx} - 2(x+u)h \,.$$

If we begin our u iterates with $u_0 = -x$, as suggested in the text, our iterates require finding $[F'(u_0)]^{-1}$. We see that for $F'(u_0)h(x) = \frac{dh}{dx} \equiv k(x)$, we have an inverse given by

$$h(x) = [F'(u_0)]^{-1} k(x) = \int_0^x k(t) dt \,.$$

So that the modified Newton-Raphson method becomes

$$\begin{aligned} u_{n+1} &= u_n - \int_0^x \left(\frac{du_n}{dt} - (t+u_n)^2 \right) dt \\ &= u_n - u_n + u_n(0) + \int_0^x (t+u_n)^2 dt \\ &= \int_0^x (t+u_n)^2 dt \,. \end{aligned}$$

Where we have used the condition that $u_n(0) = 0$ for all n. This expression is the mapping introduced at the start of this problem. We also see that with $u_0 = -x$ we see that $u_1 = 0$ as claimed.

Another way to see the introduction of this differential equation is that if the $y_{n+1} = S(y_n)$ fixed point iteration converges to a limit y, then this limit function must satisfy the following integral equation

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

From which we see that $y(x)$ must satisfy $y(0) = 0$. In addition, taking the derivative of both sides with respect to x reduces to the differential equation considered at the start of this discussion.

Exercise 5 (a more complicated integral equation)

The functional defined in this problem is

$$f(y) = b - y(x) + ax[y(x)]^2 + a\int_1^x [y(s)]^2 ds \,.$$

The fixed slope Newton-Raphson iterates are given by

$$y_{n+1} = y_n - [f'(y_0)]^{-1} f(y_n) \,.$$

which requires $[f'(y_0)]^{-1}$. To evaluate this we require the Frechet derivative of f by computing its forward difference

$$\begin{aligned} f(y+h) - f(y) &= \left(b - (y+h) + ax(y+h)^2 + a\int_1^x (y+h)^2 ds\right) - \left(b - y + axy^2 + a\int_1^x y(s)^2 ds\right) \\ &= (-1 + 2axy)h + 2a\int_1^x hyds + axh^2 + a\int_1^x h^2 ds \\ &\to (-1 + 2axy)h + 2a\int_1^x hyds \quad \text{when} \quad h \to 0 \,. \end{aligned}$$

Therefore our Frechet derivative $f'(y)h \cdot (x)$ is given by

$$f'(y)h \cdot (x) = (-1 + 2axy)h + 2a\int_1^x yhds \,.$$

For a general y it is difficult to invert this expression. If we take $y_0 = 0$, then this derivative simplifies and we have

$$f'(y_0)h \cdot (x) = -h \,,$$

and the corresponding inverse of $[f'(y_0)]^{-1}$ is simply multiplication by negative one. We can now evaluate our Newton-Raphson iterates we have that

$$\begin{aligned} y_{n+1}(x) &= y_n(x) + (b - y_n(x) + ax[y_n(x)]^2 + a\int_1^x [y_n(s)]^2 ds) \\ &= b + ax[y_n(x)]^2 + a\int_1^x [y_n(s)]^2 ds \end{aligned}$$

We can now compute the y_n iterates if desired. To determine quantitatively how well these iterates will converge we can construct a scalar function ϕ, whose $t_{n+1} = \phi(t_n)$ iterates behave in the same way as the functional iterates. To derive this function requires that $||S'(y)|| \leq \phi'(t)$. To compute $S'(y)$ requires the computing the Frechet derivative. Recognizing that our $S(\cdot)$ functional is the same as our $f(\cdot)$ functional but without the term $y(x)$, by dropping the -1 in the expression for $f'(y)h \cdot (x)$ we immediately see that its derivative is given by

$$S'(y)h \cdot (x) = 2axyh + 2a\int_1^x yhds \,,$$

as claimed in the book. Computing the norm of this operator under the condition that

$$||y - y_0|| = ||y|| = \max_x |y(x)| \le t$$

we find that

$$\begin{aligned} ||S'(y)h \cdot (x)||_\infty &\le 2ax|y| + 2a \int_1^x |y| ds \\ &= 2axt + 2a \int_1^x s ds \\ &= 2axt + a(x^2 - 1) \\ &\le 2at \end{aligned}$$

Where the last inequality follows because on the interval $x \in [0, 1]$ the expression $x^2 - 1$ is always less than or equal to zero. We can gaurrentee this if we assign $\phi'(t)$ equal to this value. To determine an initial condition for ϕ we remember that the link in the chain from y_0 to y_1 must be smaller than the corresponding link in the function ϕ i.e

$$||y_1 - y_0|| \le |\phi(0)| \, .$$

Using the initial value of $y_0 = 0$ we can immediately compute that $y_1 = b$. Then the above will hold true if we assign $\phi(0) = b$. Integrating the differential equation of ϕ we obtain

$$\phi(t) = b + at^2 \, ,$$

verifying the claim made in the text. The Mathematical file `chap_6_sec_6_prob_5.nb` contains the algebraic iterates y_0, y_1, y_2, y_3 for this example, numeric iterates could be constructed as they were for problem number 2 in Chapter 6 Section 4. The unusual thing in this case is that the link sizes ($|t_{n+1} - t_n|$) for the ϕ iterations *equal* the link sizes in the functional iterations $||y_{n+1} - y_n||$. This can be clearly seen in the Mathematica plots.

Exercise 6 (an integral equation with a linear plus a nonlinear part)

For the functional mapping $f(y)$ given by $f(y) = Ly - g(y)$, the fact that $S(y) = L^{-1}g(y)$ follows in exactly the same way as the derivation performed in exercise number 3 of this chapter. Now if $h(x)$ satisfies the Fredholm integral equation of the first kind given by

$$h(x) + \int_0^x h(s) ds = k(x) \, .$$

We can explicitly derive a solution $h(x)$ with the following steps. First taking the derivative of this expression with respect to x gives

$$\frac{dh}{dx} + h(x) = \frac{dk}{dx} \, .$$

from which we can see that an integrating factor for this equation is given by e^x and results in

$$\frac{d}{dx}(e^x h(x)) = e^x \frac{dk}{dx} \, .$$

Which can be integrated to produce

$$h(x) = e^{-x} \int_0^x e^s \frac{dk(s)}{ds} ds + h(0) \, .$$

To derive the result found in the book, we integrate this expression by parts obtaining

$$\begin{aligned} h(x) &= h(0) + e^{-x}\left[e^s k(s)|_0^x - \int_0^x e^s k(s) ds \right] \\ &= h(0) + e^{-x}\left[e^x k(x) - k(0) - \int_0^x e^s k(s) ds \right] \\ &= h(0) - e^{-x} k(0) + k(x) - e^{-x} \int_0^x e^s k(s) ds \,. \end{aligned}$$

If we assume that $h(0) = k(0) = 0$, then the above gives

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

as claimed in the book. We note that when we specify $h(0) = k(0) = 0$ in the above we are explicitly considering *only* the inhomogeneous solution of this integral equation and are ignoring the homogeneous solution. Since this is the expression we want to consider when we invert our linear operator L this is a justified simplification. Now considering the example integral equation given as

$$z(x) \equiv f(y) = y(x) + \int_0^x y(s) ds - a[y(x)]^2 - b \,,$$

we see that this expression is a linear term plus a nonlinear term. Specifically, the linear term L and the nonlinear functional $g(y)$ in the definition of $f(y) = Ly - g(y)$ are given by

$$\begin{aligned} Ly &= y(x) + \int_0^x y(s) ds \\ g(y) &= a[y(x)]^2 + b \,. \end{aligned}$$

Now we have that the Frechet derivative of g is given by $g'(y) = 2y$ and as such if we start our iterations with $y_0 = 0$, then $g'(y_0) = 0$. Thus the Kantorovich iteration function is given by $L^{-1}g(y)$, where the operator L^{-1} acting on a function $k(x)$ represents the solution $h(x)$ to the equation $L^{-1}k(x) = h(x)$. Multiplying by L on both sides of that expression we obtain that $h(x)$ must also solve $k(x) = Lh(x)$. Since this is exactly the integral equation we solved above we see that

$$L^{-1}k(x) = k(x) - e^{-x} \int_0^x e^s k(s) ds \,,$$

so the Kantorovich iteration functional $S(\cdot)$ then becomes

$$\begin{aligned} S(y) &= L^{-1} g(y) \\ &= a(y(x))^2 + b - e^{-x} \int_0^x e^s (a y(s)^2 + b) ds \\ &= a(y(x))^2 + b - a e^{-x} \int_0^x e^s y(s)^2 ds - b e^{-x} \int_0^x e^s ds \\ &= a(y(x))^2 + b - a e^{-x} \int_0^x e^s y(s)^2 ds - b e^{-x} e^s|_0^x \\ &= a(y(x))^2 + b e^{-x} - a e^{-x} \int_0^x e^s y(s)^2 ds \,. \end{aligned}$$

We will use Kantorovichs arguments to study the convergence properties of this iteration sequence. To do so we now seek a function ϕ whose behavior will bound the qualitative behavior of the more general functional iterations $y_{n+1} = S(y_n)$. To find this function $\phi(t)$, we seek a scalar function that has properties related to S when we perform ϕ iterations $t_{n+1} = \phi(t_n)$, beginning with $t_0 = 0$. The properties that ϕ must have are

- The distance between the starting point y_0, and the first iterate $y_1 \equiv S(y_0)$ of the functional mapping is less than the distance between the start of the ϕ iterations ($t_0 = 0$) and the first step $t_1 = \phi(0)$ or
$$||y_1 - y_0|| \leq |\phi(0) - 0| = |\phi(0)| \, .$$
- For all points y less than t in distance from y_0, the rate of change of the functional operator S is less than that of the function ϕ or
$$\text{if} \quad ||y - y_0|| \leq t \quad \text{then} \quad ||S'(y)|| \leq ||\phi'(t)|| \, .$$

We will begin calculating our ϕ with the first condition discussed above. As suggested in the book let $y_0 = 0$ and then $y_1(t)$ is given by
$$y_1(t) = be^{-x} \, .$$
Then the distance between y_1 and y_0 is then given by
$$||y_1 - y_0||_\infty = ||be^{-x}||_\infty \leq b \, ,$$
where we have evaluated this norm under the restriction that $x > 0$. Using this result we can guarantee the first condition above by defining $\phi(0) = b$. Now the second condition above requires the derivative of our mapping functional $S(y)$ which we compute by taking the following "forward difference"
$$\begin{aligned}
S(y + h) - S(y) &= a(y + h)^2 - ae^{-x} \int_0^x e^s (y + h)^2 ds - ay^2 + ae^{-x} \int_0^x e^s y^2 ds \\
&= 2ayh + ah^2 - ae^{-x} \int_0^x e^s (2yh + h^2) ds \\
&\to 2ayh - 2ae^{-x} \int_0^x e^s yhds \quad \text{as} \quad h \to 0 \, .
\end{aligned}$$
Showing that the derivative of S is given by
$$S'(y)h \cdot (x) = 2ay(x)h(x) - 2ae^{-x} \int_0^x e^s y(s)h(s)ds \, ,$$
as claimed in the text. So the norm of this operator is given by
$$||S'(y)|| = 2\,a \sup \left\{ \left| y(x)h(x) - 2ae^{-x} \int_0^x e^s y(s)h(s)ds \right| \right\} \, .$$
Since $y_0 = 0$ the second Kantrovich's condition becomes if $||y||_\infty \leq t$ then this implys we seek a function $\phi(t)$ such that $||S'(y)|| \leq \phi'(t)$. Using the fact that $||y||_\infty \leq t$, we have that
$$\begin{aligned}
||S'(y)|| &\leq 2\,a \left(||y|| + ||e^{-x} \int_0^x e^s y(s)ds|| \right) \\
&\leq 2\,a \left(t + e^{-x} \int_0^x e^s ||y(s)||ds \right) \\
&\leq 2\,a\,t \left(1 + e^{-x} \int_0^x e^s ds \right) \\
&= 2\,a\,t(2 - e^{-x}) \\
&\leq 4\,a\,t \, .
\end{aligned}$$
Then we will satisfy Kantrovich's second condition if we take $\phi'(t)$ *equal* to this expression. Thus in summary we have
$$\phi'(t) = 4at \quad \text{with} \quad \phi(0) = b \, .$$
Integrating this expression we obtain
$$\phi(t) = b + 2at^2 \, ,$$
for an expression for the comparison function. Using the values $a = 1$ and $b = 0.09$ we can perform the iterations discussed above. These results can be found in the Mathematica file `chap_6_sec_6_prob_6.nb`.