Numerical Analysis – Lab 2

1 The bisection method

Fooled math professor[1]**.** Once in a bank a salesperson introduced me an insurance product. If one deposits an amount v on the same day in each of the following 5 years, then an annually compound interest rate of $R = 3.5\%$ will be guaranteed through one's whole life, and one may withdraw the total amount including the interests after 20 years since one's first deposit. I was fooled and signed the contract. But later I found that the interests will start to be counted only when the total amount $5v$ has been deposited. So what is the effective interest rate r if I withdraw all after 20 years? An equation can be found as follows

$$v\sum_{k=0}^{4}(1+r)^k(1+r)^{16} = 5v(1+R)^{16}.$$

Apparently, the variable v can be removed from the equation. Let us try to solve the equation using the bisection method. First, we need to find an interval that contains the root r. It is obvious from the contex that $0 < r < R$. To use the bisection method, we also need

$$f(x) = \sum_{k=0}^{4}(1+x)^k(1+x)^{16} - 5(1+R)^{16}$$

to have opposite signs at the two ends of $[0, R]$, which is again obvious from the real world context. Moreover, f is continuous on $[0, R]$. So the bisection method applies. Note that one may calculate the sum and get

$$f(x) = \frac{1}{x}((1+x)^5 - 1)(1+x)^{16} - 5(1+R)^{16},$$

However, the sum formula introduces for f the singularity at $x = 0$ which one must be careful of. A small experiment

```
>> f = inline('((1+x)^5-1)*(1+x)^16/x-5*1.035^16')
>> f(0.035)
>> f(0.0001)
```

shows $f(0.0001) < 0$ so we can do the bisection on $[0.0001, 0.035]$:

```
>> bisect(f,0.0001,0.035,1e-14,100)
```

The bisection method converges after 42 iterations and gives $r \approx 0.0310$. Now, another question comes to me: at least how many years does it take since my first deposit to give an effective interest rate greater than my mortgage rate $\geq 3.25\%$? Can you help me to find out the answer (optional)? (Hint: something similar to the bisection method but now for the function of an integer variable can be used.)

How much daily exercise is optimal[2]**.** In his John von Neumann lecture at the annual SIAM meeting, Joe Keller asked the following question and proposed a very simple model to answer it: suppose at birth, every human being is given a fixed number of heartbeats, and once these heartbeats are used up, life ends. How should one optimally use these heartbeats to have as long a life as possible? A first immediate idea is to stay in bed and rest, so the heart rate stays low, and one uses the heartbeats as economically as possible. Another idea, however, comes from the fact that a well-trained heart beats much more slowly when the person is at rest than the heart of an untrained person. So exercise could increase the lifespan. Unfortunately, during exercise, the heart beats faster, so that one uses up the heartbeats faster, in the hope to gain them back during rest. So is there an optimum?

[1] A true story.

[2] The material is taken from the book *Scientific Computing – An Introduction using Maple and Matlab*.

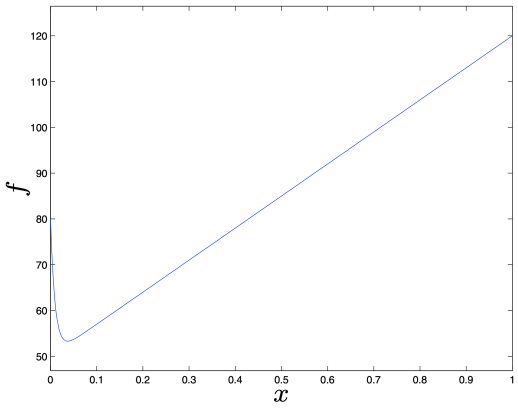
Suppose that the untrained heart beats 80 times a minute when a person is at rest, and that during exercise, it beats 120 times per minute. If a person spends a fraction x of his time exercising, then this person uses on average

$$f(x) := 120x + g(x)(1-x)$$

heartbeats per minute, where the unknown function g should be close to 80 for x small, meaning the person hardly does any exercise, and probably around 50 for x approaching 1, when the person is extremely well trained. Since it is known that a little exercise every day decreases the resting heart rate considerably, a simple model for g would be exponential decay, i.e.

$$g(x) := 50 + 30\mathrm{e}^{-100x},$$

where the choice -100 is quite arbitrary here, and should be much more carefully researched with the help of a medical doctor.



From the graph, there is an optimal choice of x that minimizes the average use of heartbeats, and which leads to the longest life possible. From calculus, we know that we need to set the derivative to zero to find the minimum. It is easy to find

$$f'(x) = 30\mathrm{e}^{-100x}(100x - 101) + 70.$$

```
>> fp = inline('30*exp(-100*x)*(100*x-101) + 70')
>> fp(0)
>> fp(1)
>> x=bisect(fp,0,1,1e-14,100)
>> 24*60*x
>> f = inline('120*x+(50+30*exp(-100*x))*(1-x)')
>> f(x)
```

The optimal time of exercises each day is $24 \times 60 \times 3.73\% \approx 53.7$ minutes.

Golden Section for Optimization (optional). Suppose f on $[a, b]$ attains the unique local minimum at $x_{\min}$, and f is strictly decreasing on $[a, x_{\min}]$ and strictly increasing on $[x_{\min}, b]$. Is it possible to find such the minimum point $x_{\min}$ without using the derivative[3] of f? In the spirit of the bisection method, let us pick a point p_1 on (a, b). Then at least one of $[a, p_1]$ and $[p_1, b]$ must contain the minimum point $x_{\min}$, but which subinterval? Note that if f is not monotone on a subinterval then the subinterval must contain $x_{\min}$. To check the monotonicity, we need three points. So let us pick one more point $q_1 \in (p_1, b)$. If $f(p_1) < f(q_1)$ then $x_{\min} \in [a, q_1]$. If $f(p_1) > f(q_1)$ then $x_{\min} \in [p_1, b]$. Then on the selected subinterval we can repeat the procedure. To ensure that $[a, q_1]$ and $[p_1, b]$ have the same length, we let $p_1 = ta + (1-t)b$, $q_1 = (1-t)a + tb$. To reuse the points, we require if $[a, q_1]$ is selected then $p_1 = q_2$; if $[p_1, b]$ is selected then $q_1 = p_2$. By symmetry, we need only

$$ta + (1-t)b = (1-t)a + t((1-t)a + tb) \quad \text{i.e. } (1 - t - t^2)(a - b) = 0.$$

[3] which may be non-existent or expensive to compute

Solution of the equation is $t = \frac{-1+\sqrt{5}}{2} \approx 0.618$. We write the following function:

```
function p = golden_section(f,a,b,tol,maxIt)
t = (-1+sqrt(5))/2;
p = t*a + (1-t)*b; q = (1-t)*a + t*b;
fp = f(p); fq = f(q);
disp('Bisection Methods')
disp('-----------------------------------------------------------------')
disp(' n     a_n             b_n             p_n             f(p_n)')
disp('-----------------------------------------------------------------')
formatSpec = '%2d    % .9f    % .9f    % .9f    % .9f \n';
for n = 1:maxIt
    fprintf(formatSpec,[n,a,b,p,f(p)])
    if abs(b-a)<tol
       break;
    else
        if fp<fq
            b = q; q = p; p = t*a + (1-t)*b; fq = fp; fp = f(p);
        else
            a = p; p = q; q = (1-t)*a + t*b; fp = fq; fq = f(q);
        end
    end
end
```

and use it to solve the optimal exercise problem:

```
>> f = inline('120*x+(50+30*exp(-100*x))*(1-x)')
>> golden_section(f,0,1,1e-10,100)
```

which gave the same solution as before.

2 Fixed point iterations

How to calculate a square-root. A cuneiform tablet dating from around 1750 B.C. shows that the Babylonians knew $\sqrt{2} \approx 1.41421296$. Their method is unknown. But a method appears in Book 1 of *Metrica*, written by Heron of Alexandria in the first century A.D., to calculate $\sqrt{720}$.

The cuneiform tablet of the Babylonians

Heron or Hero of Alexandria (about 10 A.D. to 75 A.D.)

Heron was an important geometer and worker in mechanics who invented many machines including a steam turbine. His best known mathematical work is the formula for the area of a triangle in terms of the lengths of its sides.

Suppose we want to calculate $p = \sqrt{c}$ for a given positive number $c > 1$. An initial guess $p_0 = 1$ is obviously too low, and hence c/p_0 is too high. Thus it is reasonable to average the two to get a better guess:

$$p_1 = \frac{1}{2}(p_0 + \frac{c}{p_0}).$$

We can use it as the fixed point iteration

$$p_n = \frac{1}{2}(p_{n-1} + \frac{c}{p_{n-1}}).$$

If $p_n \to x$, then x would satisfy

$$x = \frac{1}{2}(x + \frac{c}{x})$$

which implies $x^2 = c$, i.e. the fixed point is exactly $p = \sqrt{c}$. We use

```
>> g = inline('(x+2/x)/2')
>> fixedpoint(g,1,1e-14,100)
>> sqrt(vpa(2))
```

to get

n	p_n	$g(p_n)$
0	1.000000000000000	1.500000000000000
1	1.500000000000000	1.416666666666667
2	1.416666666666667	1.414215686274510
3	1.414215686274510	1.414213562374690
4	1.414213562374690	1.414213562373095
5	1.414213562373095	1.414213562373095
6	1.414213562373095	1.414213562373095

while the exact value of $\sqrt{2}$ = 1.4142135623730950....

We find p_5 already has the 16 significant digits being correct. It is interesting to observe that p_2 has 3 correct significant digits, p_3 has 6 correct significant digits, and p_4 has 12 correct significant digits. That is, every iteration *doubles* the number of correct significant digits! Let $g(x) = \frac{1}{2}(x + \frac{c}{x})$. We have $g'(x) = \frac{1}{2}(1 - \frac{c}{x^2})$ and $g'(\sqrt{c}) = 0$ at the fixed point $p = \sqrt{c}$ of g. Consequently,

$$\frac{p - p_{n+1}}{p - p_n} = \frac{g(p) - g(p_n)}{p - p_n} = g'(\xi_n) \overset{n}{\to} g'(p) = 0.$$

The error reduction from each iteration is by a factor that tends to zero! We can also calculate $g''(x) = \frac{c}{x^3}$ and $g''(\sqrt{c}) = c^{-1/2}$. Consequently,

$$\frac{p - p_{n+1}}{(p - p_n)^2} = \frac{g(p) - g(p_n)}{(p - p_n)^2} = \frac{1}{2}g''(\eta_n) \overset{n}{\to} \frac{1}{2}g''(p) = \frac{1}{2}c^{-1/2}.$$

If $|p - p_n| \approx 10^{-m}$ then $|p - p_{n+1}| \approx \frac{1}{2}c^{-1/2}10^{-2m}$. That explains our observation.

Periodic points and chaos (optional). The fixed point iteration $p_n = g(p_{n-1})$, $n = 1, 2, \ldots$ does not always lead to a fixed point of g. It can have much more complicated dynamical behaviours. An example is the discrete Logistic equation

$$p_n = rp_{n-1}(1 - \frac{p_{n-1}}{k})$$

where p_n is the population of a species at the time n, $r > 0$ is the growth rate, and k is the carrying capacity of the environment. Suppose $k = 1$. Let $g(x) = rx(1 - x)$. It is easy to find

$$\min_{[0,1]} g(x) = 0, \quad \max_{[0,1]} g(x) = \frac{r}{4}.$$

So g is a continuous map from $[0, 1]$ into $[0, 1]$ when $0 \le r \le 4$, and, by the Brower fixed point theorem, g has at least a fixed point p on $[0, 1]$. Indeed, $x = g(x)$ has only two roots 0 and $1 - \frac{1}{r}$. The question is whether the fixed point iteration converges. We find $g'(x) = r(1 - 2x)$ and $|g'(x)| \le r$ for $x \in [0, 1]$. So if $0 < r < 1$ then the fixed point iterates $\{p_n\}$ converge to the unique (on $[0, 1]$) fixed point $p = 0$ for every $p_0 \in [0, 1]$. This can be verified numerically:

```
>> g = inline('r*x*(1-x)')
>> r = rand(1)
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
```

where the first command returns a function $g(r,x)$, `rand(1)` returns a random 1×1 matrix (number) from the uniform distribution on $[0,1]$, and `@(x)g(r,x)` defines an anonymous function of x using the specified value of the parameter r.

If $r > 1$, then both 0 and $1 - \frac{1}{r}$ are the fixed points of g on $[0,1]$. But $g'(0) = r > 1$ and $g'(1-\frac{1}{r}) = 2 - r$. By the continuity of g', in some neighbourhood of 0 it holds that $|g'(x)| > \frac{1+r}{2} > 1$. Consequently, whenever p_{n-1} goes into the neighbourhood, it leads to $|p_n - 0| > \frac{1+r}{2}|p_{n-1} - 0|$, that is, the fixed point 0 is a repelling point (unstable). Since $g(x) = g(0)$ if and only if $x = 0$ or $x = 1$ but $1 \notin g([0,1])$ if $1 < r < 4$, we expect for $1 < r < 4$ the fixed point iterates $\{p_n\}$ can never converge to 0 for any $p_0 \in (0,1)$. For the other fixed point $p = 1 - \frac{1}{r}$, we have $|g'(p)| = |2 - r| < 1$ if $1 < r < 3$ and p is an attracting point (stable). Moreover, for $1 < r < 3$ we find $1 - \frac{1}{r} > g(x) > x$ for $x \in (0, 1 - \frac{1}{r})$ and $g(x) < x$ for $x \in (1 - \frac{1}{r}, 1)$. So, in that case, we expect the fixed point iterative sequence $\{p_n\}$ converges to $1 - \frac{1}{r}$ for every $p_0 \in (0,1)$. Let us test it:

```
>> r = 1+2*rand(1)
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
>> 1 - 1/r
```

If $r > 3$, then both the fixed points 0 and $1 - \frac{1}{r}$ are unstable! It is very rare for the fixed point iterative sequence $\{p_n\}$ to converge to $1 - \frac{1}{r}$ – if and only if p_0 is on the backward orbit from $p = 1 - \frac{1}{r}$ i.e. $p_0 \in \{p\} \cup g^{-1}(p) \cup g^{-1}(g^{-1}(p)) \cup \dots$ which consists of two subsequences: one converging to 0 and the other to 1. For a general $p_0 \in [0,1]$, where is p_n going? Try

```
>> r = 3.1
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
```

I saw that eventually p_n will take alternatively the two values $p^{(2,1)} = 0.558014125202696$ and $p^{(2,2)} = 0.764566519958594$. That is, $p^{(2,1)} = g(p^{(2,2)})$ and $p^{(2,2)} = g(p^{(2,1)})$. It follows that $p^{(2,1)} = g(g(p^{(2,1)}))$ i.e. $p^{(2,1)}$ is a fixed point of $g \circ g$ (and so is $p^{(2,2)}$). We call a fixed point of $g \circ g$ also as a period 2 point of g. The things change rapidly as we increase r:

```
>> r = 3.5
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
```

gave four period 4 points: 0.500884210307218, 0.874997263602464, 0.382819683017324, 0.826940706591439.

```
>> r = 3.56
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
```

gave eight period 8 points. It is very interesting to find all the critical values of r across which the period doubles. Such critical values are called bifurcation points. Let r_k be the bifurcation point across which the period changes from 2^{k-1} to 2^k. Then the limit exists: $r_k \overset{k}{\to} r_\infty < 4$. For $r_\infty < r < 4$, the fixed point iterative sequence becomes a chaos:

```
>> r = 3.6
>> p0 = rand(1)
>> fixedpoint(@(x)g(r,x), p0,1e-14,200)
```