These are the lecture notes for CSC349A Numerical Analysis taught by Rich Little in the Spring of 2018. They roughly correspond to the material covered in each lecture in the classroom but the actual classroom presentation might deviate significantly from them depending on the flow of the course delivery. They are provided as a reference to the instructor as well as supporting material for students who miss the lectures. They are simply notes to support the lecture so the text is not detailed and they are not thoroughly checked. Use at your own risk.

1 Quick review of previous lecture

In the previous lecture we covered the logistics of the course, had a discussion about why plagiarism/copying is stupid in addition to being wrong, and briefly described at a very high level the various topics we will cover in this course. This was followed by a motivating example to understand how numerical methods can be applied to solve engineering problems. Using some basic physics we showed how we can predict the velocity of a free falling parachutist using two approaches. The first was based on using differential calculus to directly find a function that satisfies the differential equation describing the motion. We called this solution exact and analytical. The alternative we discussed was to use Euler's method in which we discretize time and approximate the derivative with a linear slope using a finite difference division. This method is incremental utilizing a time step and using the velocity approximation in the current time step to estimate the velocity in the future time step. Even though it is more computationally intensive and is only an approximation to the exact solution it can be made sufficiently close by selecting an appropriately small time step. In this lecture we will continue exploring this particular problem using MATLAB which is a programming language and environment designed for scientific and engineering applications that has strong support for numerical methods. MATLAB will also be the programming language we will use for code examples in this course.

2 Programming in MATLAB

The numerical approach to solving the differetial equation of the falling parachutist is straightforward but requires a lot of repeated tedious calcula-

tion especially for smaller step sizes. This is the perfect job for a computer to do and we will be learning how to do this using high level programming environments for numerical computations. We will be focusing on MATLAB which is a commercial software package. I strongly encourage you to try out the little program examples on your own and experiment. The best (and probably the only) way to become a good programmer is through a lot of practice directly on a computer.

MATLAB is based on an interpreter meaning that the user can type in a statement, press enter and then immediately see the result of that statement. This is in contrast to compilers that need to have the full source code of a program and then translate it to a binary that can be used to run the code. It is typical to write MATLAB code incrementally and the software provides a prompt every time a statement is executed. For example you can type the following in the command window of matlab to obtain the square root of 2:

```
sqrt(2)

ans =
     1.4142
```

It is straightforward to perform simple calculations, and use variable names for storing numbers. For example:

```
>> 5 + 3

ans =

     8

>> a = 5;
>> a + 3
ans =

     8
>> b = a + 3
b =
     8
```

Notice that in MATLAB ending a statement with a semi-colon makes it "silent" meaning there is no output whereas leaving out the semi-colon results in showing the result of that statement. We can now write an expression to evaluate the analytical solution to the falling parachutist problem. Recall that the analytical solution from differential calculus was:

$$v(t) = \frac{gm}{c}(1 - e^{-\frac{ct}{m}}) \tag{1}$$

So for example if we want to compute the value at $t = 4sec$, with a mass of $68.1kg$ with a drag coefficient of $12.5kg/sec$ and a gravity constant $g = 9.8$ we can use MATLAB to do so as follows:

```
>> g = 9.81;
>> m = 68.1;
>> c = 12.5;
>> t = 4;
>> v = (g * m / c) * (1 - exp(-(c * t/m)))
v = 27.7976
```

It is educational to consider how the same computation would be performed with a classic calculator that only supports binary and unary operations. You would have to perform several intermediate calculations and combine the results to achieve the effect that the single line that calculates the velocity in MATLAB does. MATLAB, or more generally the compiler or interpreter used, translates this high level expression automatically into the appropriate sequence of machine instructions that similarly to a calculator are either binary or unary operators.

If we wanted to let's say compute the velocity at $t = 6$ seconds we could change t to be 6 and then retype the expression corresponding to the analytical solution. This is tedious, time consuming and practically impossible if we wanted to compute this for many values of t. In the command-line window, it is possible to cycle through the previous commands using the **Up** and **Down** arrow keys but that only saves a little bit of typing.

```
>> t = 6;
>> v = (g * m / c) * (1 - exp(-(c * t/m)))

v =

    35.6781
```

A fundamental abstraction in computer programming is the concept of a function or procedure that takes some input arguments, performs some computation using them and returns the result. In MATLAB functions are defined in separate files ending with the extension .m. Any text editor can be used to create these files which need to be named with the same name as the defined function. In MATLAB there is a built-in editor window that can be used. We can define a function that abstracts the analytical solution to the velocity problem. In honor of a recent movie that is somewhat related to our problem we will name our function *skyfall* and define it in a file named *skyfall.m* the contents of which are:

```
function [v] = skyfall(g,m,c,t)
        v= (g * m / c) * (1 - exp(- c * t / m));
end
```

As you can see from the source code, the function is named skyfall and takes as arguments the parameters g, m, c, t and returns the value v. We can now call this function from the command window (assuming that the file skyfall.m is in the current directory or a directlory in the MATLAB path).

```
>> skyfall(g,m,c,4)
ans = 27.7976
>> skyfall(g,m,c,6)
ans = 35.6781
>> skyfall(g,m,c,20)
ans = 52.0848
```

Computer programs need to be correct and perform what they are supposed to do from the computer perspective. They also need to be readable and understandable by other programmers as most programming is done in teams. Comments are lines that start with special characters (in MATLAB %) that are completely ignored by the compiler but are used to provide information that can help other programmers understand your code. Here is the *skyfall* function with added comments:

```
% returns the velocity of a free falling parachutists
% based on the analytical solution described in our course notes
% g is the gravity constant
% m is the mass of 007
```

```
% c is the drag coefficient caused by his expensive outfit
% t is the time in seconds
function [v] = skyfall(g,m,c,t)
        v= (g * m / c) * (1 - exp(- c * t / m));
end
```

We can see that we are calling the *skyfall* function several times with different arguments that progress regularly. Loop constructs provide a structured way of expressing repetition in programming languages. In MATLAB there is a special syntax for creating sequences of numbers that works as follows:

```
>> t = 1:10
t =

    1    2    3    4    5    6    7    8    9   10
>> t = 1:2:10
t =

    1    3    5    7    9
>> t = 1:0.5:10
t =
t =

 Columns 1 through 7:

     1.0000     1.5000     2.0000     2.5000     3.0000     3.5000     4.0000

 Columns 8 through 14:

     4.5000     5.0000     5.5000     6.0000     6.5000     7.0000     7.5000

 Columns 15 through 19:

     8.0000     8.5000     9.0000     9.5000    10.0000
```

As is evident from the example the syntax is *start:step:end*. Until this example the code examples have been simple enough that they can easily be

translated in any programming language. This syntax is more specific to MATLAB and is based on the fact that MATLAB has very strong support for matrices. The results that you see are essentially row vectors or matrices of dimension 1 by 10 for the first case. There is special syntax for creating matrices and accessing their elements which should be evident from the following examples.

```
>> x = [1, 2, 3]; % a row vector
>> x
x =

   1   2   3
x = [1; 2; 3]; % a column vector
>> x
x =

   1
   2
   3
>> x = [1,2; 3,4; 5,6] % a 3 by 2 matrix
x =

   1   2
   3   4
   5   6
>> x(1,2)     % accessing a single element at position 1,2
ans = 2
>> x(2,1)
ans = 3
>> x(:,1)      % the first column
ans =

   1
   3
   5
>> x(1,:)      % the first row
ans =
```

```
   1   2
```

We can now write a loop to iterate over values of t and compute the associated velocities using the function *skyfall* we defined.

```
>> g = 9.81;
>> m = 68.1;
>> c = 12.5;
>> for t = 0:2:12;
>      skyfall(g,m,c,t)
> end
ans = 0
ans =  16.4217
ans =  27.7976
ans =  35.6781
ans =  41.1372
ans =  44.9189
ans =  47.5387
```

Now instead of having to retype the function call several times we can easily express different iterations. For example by changing $t = 0 : 2 : 12$ to $t = 0 : 1 : 12$ we can compute the velocity at every second instead of every two seconds. This version relies on the missing semicolon when calling skyfall in order to output the answer. It would be nice to have all the velocities in a vector so that we could do further operations with them, perhaps average them or plot them. We can easily modify our code to achieve this.

```
>> g = 9.81;
>> m = 68.1;
>> c = 12.5;
>> velocities = zeros(1,11); % create vector of velocities
>> times = zeros(1, 11); % create vector of times
>> for t = 0:1:10;
>      velocities(t+1) = skyfall(g,m,c,t);
>      times(t+1) = t;
> end
>> velocities
```

```
velocities =

 Columns 1 through 7:

    0.00000    8.95318   16.40498   22.60717   27.76929   32.06577   35.64175

 Columns 8 through 11:

   38.61807   41.09528   43.15708   44.87314
>> plot(times, velocities);      % plot the velocities against time
>> xlabel('time in seconds');    % add labels
>> ylabel('velocity in meters/second');
```

One important difference between MATLAB and many other programming languages is that the indexing of vectors and arrays is done starting with 1 rather than 0. That is why we have $t + 1$ in the line assigning to the current return value of *skyfall* to the corresponding entry of the row vector of velocities. In general this is always a tricky part especially when porting MATLAB code to C/C++ or vice versa and one has to be careful. Another possibility is to encapsulate the iteration in the function itself. Here is a version of *skyfall.m* that does that.

```
% returns the velocities of a free falling parachutists
% based on the analytical solution described in our course notes
% g is the gravity constant
% m is the mass of 007
% c is the drag coefficient caused by his expensive outfit
% t0 is the starting time
% tn is the ending time
% h is the time step size

function [times, velocities] = skyfall_with_loop(g,m,c,t0,h,tn)
        n = tn - t0 / h;
        times = zeros(1,n);
        velocities = zeros(1,n);
        i = 1;                        % iteration
        for t=t0:h:tn;
                times(i) = t;
```

```
                velocities(i) = (g * m / c) * (1 - exp(- c * t / m));
                i = i+1;
        end
endfunction
```

Notice the use of a separate variable i which counts iterations so that the velocity vector gets filled by one value at a time independently of the step size. What would happen if i was replaced by t ? If you can't figure it out, try it out and see what happens ? Can you explain it ? What about if i is replaced by $t + 1$ would that work ? If not is there a particular step size for which it would ?

Now, you may have noticed that in this particular case all this extra work of looping is unnecessary. For counting loops the matrix itself gives us a natural way to loop. Go back to our original *skyfall* function:

```
>> g = 9.81;
>> m = 68.1;
>> c = 12.5;
>> t = 0:2:12;
>> v =  skyfall(g,m,c,t)

v =

         0   16.4217   27.7976   35.6781   41.1372   44.9189   47.5387

>> plot(t, v)
```

We have now covered all the concepts we need for writing the numerical iterative method for solving the parachutist problem using Euler's method. Here is the code which should be understandable based on what we have covered.

```
function skyfall_euler(m,c,t0,v0,tn,n)
        % print headings and initial conditions
        fprintf('values of t approximations v(t)\n')
        fprintf('%8.3f',t0),fprintf('%19.4f\n',v0)
        % initialize gravitational constant, compute step size h
        g=9.81;
```

```
        h=(tn-t0)/n;
        % set t,v to the initial values
        t=t0;
        v=v0;
        % compute v(t) over n time steps using Eulers method
        for i=1:n
                v=v+(g-c/m*v)*h;
                t=t+h;
                fprintf('%8.3f',t),fprintf('%19.4f\n',v)
        end
end
```

Notice the two statements inside the loop that update respectively the current value of velocity based on the previous value of velocity and the corresponding time incremented by the time step size h. Also notice the use of *fprintf* which enables nicer formatting of floating point nubmers when they are printed on the screen.

There are many more commands, built-in functions, and syntax in MATLAB that you will gradually learn as we go over more examples and you get more practice working on problems on your own. Here are a few more commands that you might find useful. The effect of the commands should be self evident. It is possible to get help about built-in function using *help* as shown below.

```
>> x = pi;
>> format long
>> x
x =

   3.141592653589793
>> format short
>> x
x =

   3.141592653589793
>> A = [1, 2; 3 4]

A =
```

```
     1     2
     3     4
>>
ans =

   -2.0000    1.0000
    1.5000   -0.5000
>> help inv
 INV    Matrix inverse.
    INV(X) is the inverse of the square matrix X.
    A warning message is printed if X is badly scaled or
    nearly singular.

    See also slash, pinv, cond, condest, lsqnonneg, lscov.
>> x = [0: 0.01: 1.5];
>> y = exp(x) - 4 * sin(x);  % notice the vector syntax (no loop)
>> plot(x,y);
```

The concepts of a function, the ability to write algebraic expressions that get translated into machine language, and iteration are fundamental to computer programming and today we take them for granted. It is important to realize that when they were introduced they were considered very advanced and radical ideas that were met with resistance from the human programmers whose job was to do some of these processes directly in machine language. *Fortran* (an acronym for Formula Translator) is considered the grand father of all programming languages and was a language designed for scientific computing and consequently numerical methods. A significant amount of still running code is written in Fortran and several of the libraries that are used almost everywhere for numerical methods are still only available in Fortran.

3 Approximation and Roundoff Errors

There are several types of different erros that can arise in engineering scientific applications and it is important to understand both their sources and effects when studying numerical methods.

- **Formulation or modeling error** arises because of incomplete mathematical models. For example the simplified parachutist model we described will never exactly predict the motion of a real parachutist.
- **Data uncertainty/inherent erorr** Noisy data due to inexact measurements or observation.
- **Truncation error (Chapter 4)** results from using inexact approximations instead of an exact mathematical procedure such as the errors between the Euler method numerical procedure for predicting velocity versus the analytical solution obtained through differential calculus.
- **Roundoff error (Chapter 3)** is due to the fixed, finitite precision representations used to represent real/complex numbers in computers.

We start by discussing different ways of measuring error. If p denotes the true (exact) value of some quantity, and p^* denotes some approximation to p then the **absolute true error** is defined as:

$$|E_t| = |p - p^*| \tag{2}$$

The absolute error only makes sense if you have a sense of the magnitude of p, the quantity you are approximating. For example consider $p = 1234321$ and $p* = 1234000$, the $|E_t| = 321$ seems large, although $p*$ is quite accurate and agrees with p to 4 significant digits. In constrast, if $p = 0.001234$ and $p* = 0.001111$, then $|E_t| = 0.000123$ seems small, although $p*$ is not very accurate and agrees with p to only 1 significant digit.

A better approach is to introduce the concept of relative error, in which the magnitude of the exact value p is used to "normalize" the error. More specifically the **relative error** is defined as:

$$|\varepsilon_t| = \frac{|p - p^*|}{|p|} = |1 - \frac{p^*}{p}| \tag{3}$$

The *significant digits* of a number are those that can be used with confidence. It is a discrete way of measuring error in approximations in contrast to the relative error which is a continuous way of measuring error in approximation. The relative error also indicates the number of correct significant digits in an approximation p^*. For example for $p = \pi = 3.14159265\ldots$:

Frequently it is useful to relate the relative error to the number of correct significant digits. This can be done with heuristics like the following:

approximations	number of correct significant digits	relative error
3.1	2	0.013
3.14	3	0.00051
3.141	4	0.00019
3.1415	5	0.000029

$$|\varepsilon_t| < 5 \times 10^{-n} \tag{4}$$

in which case $p*$ approximates p to **n significant digits**. Note that this is not an exact relationship and is a conservative estimate. It is possible that the approximation has more than n correct significant digits but not less.

In order to calculate the relative or absolute error of a particular quantity we need to know the value of the approximation p^* as well as the true value p. In many cases we don't know the true value. In fact numerical methods in engineering are used when we don't know the true value otherwise we could just simply use it instead of computing a numeric approximation. The main reason these error measures are useful is that they can give us good estimates of how well a particular method works for the case where we have exact results increasing our confidence that they will work well for the cases for which there are no exact results.

There are many algorithms in numerical methods that operate in an iterative function in which an estimate of solution is used to compute a better estimate and the process is repeated until some convergence criterion is met. In this cases it is common to use our best current estimate of the true value (i.e the previous estimate) to compute the error compared to the current estimate.

In this case of iterative algorithms we can use the following approximation to the relative errror:

$$|\varepsilon_a| = \frac{|p_i - p_{i-1}|}{p_i} \tag{5}$$

Similarly to the imprecise conservative estimation of significant digits above we can use the following relation to estimate the number of significant digits for the relative error in iterative algorithms:

$$|\varepsilon_a| < 0.5 \times 10^{-n} \tag{6}$$

i	p_i	$\lvert\varepsilon_t\rvert = \frac{\lvert e^{0.5} - p_i\rvert}{\lvert e^{0.5}\rvert}$	$\lvert\varepsilon_a\rvert = \frac{\lvert p_i - p_{i-1}\rvert}{\lvert p_i\rvert}$
1	1	0.393	
2	1.5	0.0902	0.333
3	1.625	0.0144	0.0769
4	1.645833	0.00175	0.0127
5	1.6484375	0.000172	0.00158
6	1.6486979	0.0000142	0.000158

4 Example of Error Estimate in Iterative Methods

These concepts will hopefully become clearer through an example (Example 3.2 pages 58-59 of the 6th edition and 61-62 of the 7th edition).

Mathematicians like to represent functions by infinite series. For example the exponential function can be computed using:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} \tag{7}$$

As more terms are added to the series the approximation becomes better and better. This is called a McLaurin series expansion. Let $p_1 = 1$, $p_2 = 1 + x$, $p_3 = 1 + x + \frac{x^2}{2!}$ and so on.

We can use this series approximation at different levels of precision to compute the value of e^x for $x = 0.5$. We can use the true value $p = e^{0.5} = 1.648721..$ to compute the true relative error $|\varepsilon_t|$.

Note that the value of $|\varepsilon_t|$, which can only be computed if we know the true (exact) answer p, can be used to estimate the number of correct significant digits in each approximation. For example: For p_3 we have $0.0144 < 5 \times 10^{-2}$ so the approximation $p_3 = 1.625$ has at least 2 correct significant digits. For p_5 we have $0.000172 < 5 \times 10^{-4}$ and therefore $p_5 = 1.6484375$ has at least 4 correct significant digits.

In practice, if a sequence of approximations to some unknown value is computed using an iterative algorithm, (we will use several such algorithms during this course), then the exact relative error $|\varepsilon_t|$ can not be computed. However, the relative error in each approximation p_i can be approximated by $|\varepsilon_a|$ and can be used to estimate conservatively the number of correct significant digits. For examples for i=6 in the table above we have $|\varepsilon_a| = 0.000158 < 0.5 \times 10^{-3}$ implying that $p_6 = 1.6486979$ has at least 3

correct significant digits (notice that in fact it has 4 which means this is a conservative estimate).

Armed with the syntax and concepts we learned for programming in MATLAB it is straightforward to write a MATLAB function to compute the table above showing how the true and approximate relative error decrease as more terms are added to the series. Here is the corresponding code of the function *mclaurin_exponential*:

```
function [ em ] = mclaurin_exponential( x, n )
% mclaurin_exponential:
%     Prints the true relative error and the approximate
%     relative error when approximating e with a
%     McLaurin series. The code corresponds to example
%     3.2 of the textbook and also is covered in handout 2.

    em = 1;          % the approximation
    e = exp(1);      % the true value
    et = abs(e^(x) - em) / e^(x);   % true rel error
    ea = 0;                              % approx rel error
    prev_em = 0;
    i = 1;
    fprintf('i \t pi \t\t true error \t approximation error\n');
    fprintf('%d \t %4.6f \t %4.6f \t %4.6f\n', i, em, et, ea);

    for i = 1:n
        prev_em = em;               % store prev approx
        em = em + (x^i / factorial(i));
        et = abs(e^(x) - em) / e^(x);   % true rel error
        ea = abs((em - prev_em) / em);      % approx rel error
        fprintf('%d \t %4.6f \t %4.6f \t %4.6f\n', i+1, em, et, ea);
    end
end
```

To produce the table we simply provide the value of x and the number of series terms to use for the approximation (or number of iterations) n.

```
>> mclaurin_exponential(0.5, 6)
i   pi true error   approximation error
```

```
1   1.000000   0.393469   0.000000
2   1.500000   0.090204   0.333333
3   1.625000   0.014388   0.076923
4   1.645833   0.001752   0.012658
5   1.648438   0.000172   0.001580
6   1.648698   0.000014   0.000158
7   1.648720   0.000001   0.000013

ans =

   1.648719618055555
```