Assignment 1 Design Document

Program Requirements and Strategy

We are required to write a bash script that creates three GNU plots of three separate sets of data. Each set contains data derived from a collatz sequence with a given starting value that our program must provide. The algorithm for creating the collatz sequence for any given starting value is already implemented in a C program which is provided to us. Our three sets of data must contain respectively the lengths of each collatz sequence generated, their maximum value, and the frequency of each length. Clearly, the third data set cannot be obtained until the first has. Data must be appended to the first two data sets with every collatz sequence generated, and a collatz sequence must be generated for every starting value in the range 2 to 10,000.

The task of our program is twofold. The first part of our program must run the given file to build a collatz sequence with the starting value as input across every value in the specified range. Then, it must determine the size of each sequence and store that value with the input value of the corresponding in a data file exclusively for sequence lengths. Next, it must do the same for the maximum values for each sequence. In the interest of efficiency, it would be advisable to avoid running the program multiple times. Hence the output of a collatz run will be stored in a temporary file, and bash commands used to extract the desired information from this data. After the for loop completes, the completed data file for sequence lengths can be used to build the data file for sequence length frequencies.

The second part of our program creates the GNU plots. Example code for doing so has been provided in the assignment document, which covers most of the bases for what we must implement, except setting the style of the plot, fixing the range of either axis, and various other formatting facilities. See the pseudocode for further details.

Pseudocode

# Comments are in blue (they need not reflect comments in the actual code)

* Footnotes are in red

```
Use Makefile to build collatz from scratch and compile it.
Remove the data file for lengths*
Remove the data file for maximums*
Remove the data file for frequencies*

For every value of i where i ranges from 2 to 10,000:
    # This command creates the temporary file, tmp_file
    Run the given collatz file with the starting value i and store the output in tmp_file

    Determine the length of the temporary file and store it in the variable len
    # This command creates the data file for lengths, len_collatz.dat
    Append the i and len values to a data file for lengths and separate them by a space

    Sort the temporary file numerically and store the final value in the variable max
    # This command creates the data file for maximums, max_collatz.dat
    Append the i and max values to a data file for maximums and separate them by a space

# This command creates the data file for frequencies of lengths for all sequences generated
```

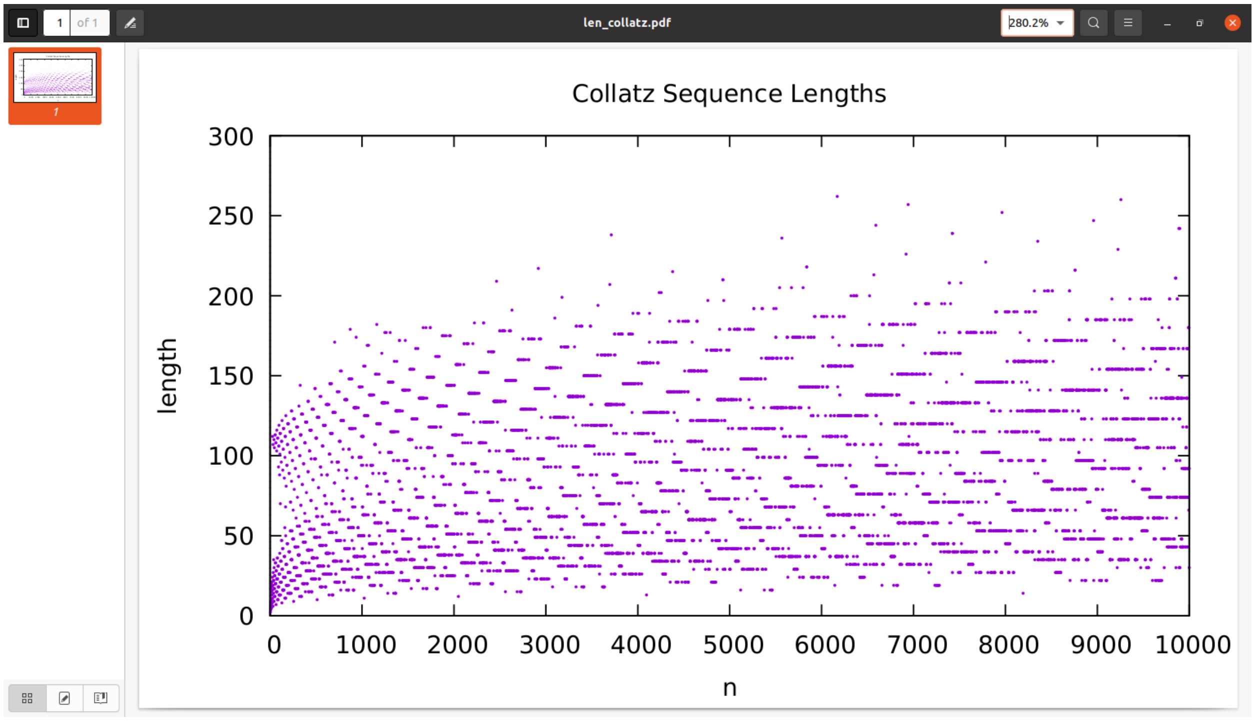
```
Attain the frequencies of lengths from len_collatz.dat and store the data in freq_collatz.dat **

# Now we begin with the GNU plots
# The first plot is for sequence lengths
gnuplot <<END
        Set the terminal to output to pdf
        Provide the file name
        Provide a title
        Name the x-axis
        Name the y-axis
        Move the axes to the center of the chart
        Create the plot with dots and without the line title
END

# The second plot is for sequence maximums
gnuplot <<END
        Set the terminal to output to pdf
        Provide a file name
        Provide a title
        Name the x-axis
        Name the y-axis
        Limit the range of the y-axis to go no higher than 100,000
        Move the axes to the center of the chart
```

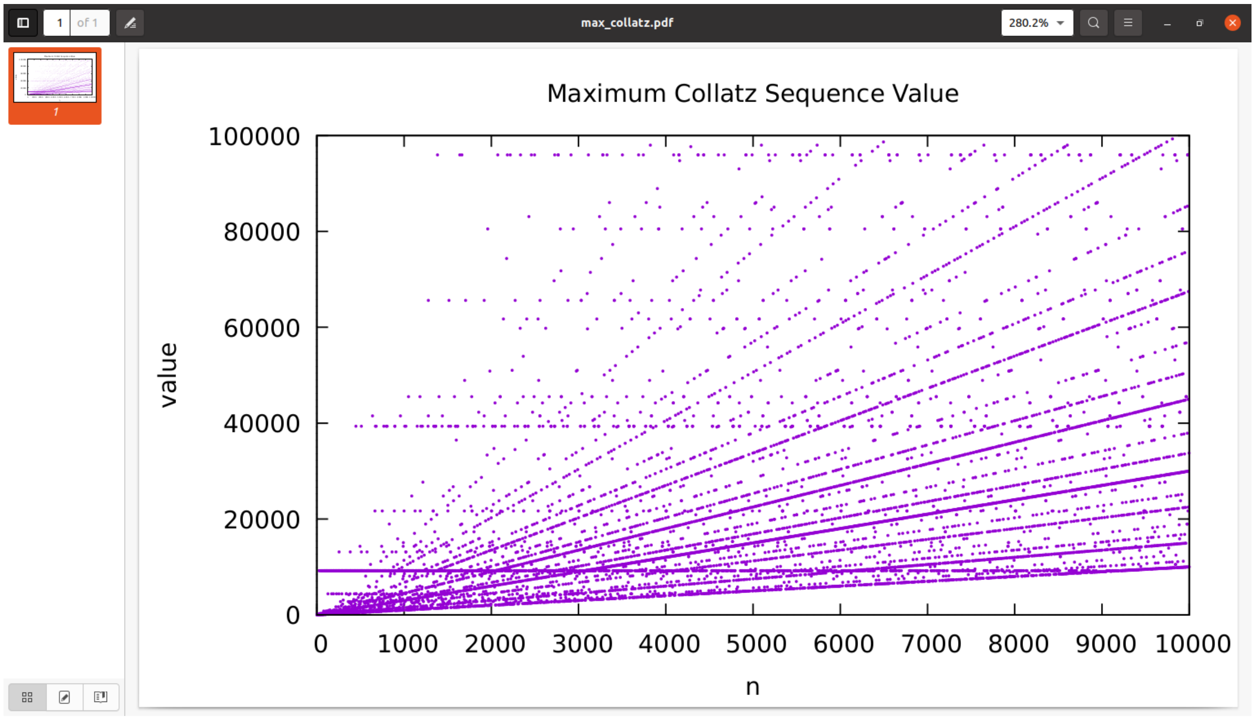
```
    Create the plot with dots and without the line title
END


# The third plot is for sequence length frequencies
gnuplot <<END
    Set the terminal to output to pdf
    Provide a file name
    Provide a title
    Name the x-axis
    Name the y-axis
    Limit the range of the x-axis to go no farther than 225
    Move the axes to the center of the chart
    Specify the style of the plot to be a histogram
    Space the ‘tics’ on the x-axis by 25 from 0 to 225
    Create the plot without the line title
END
```

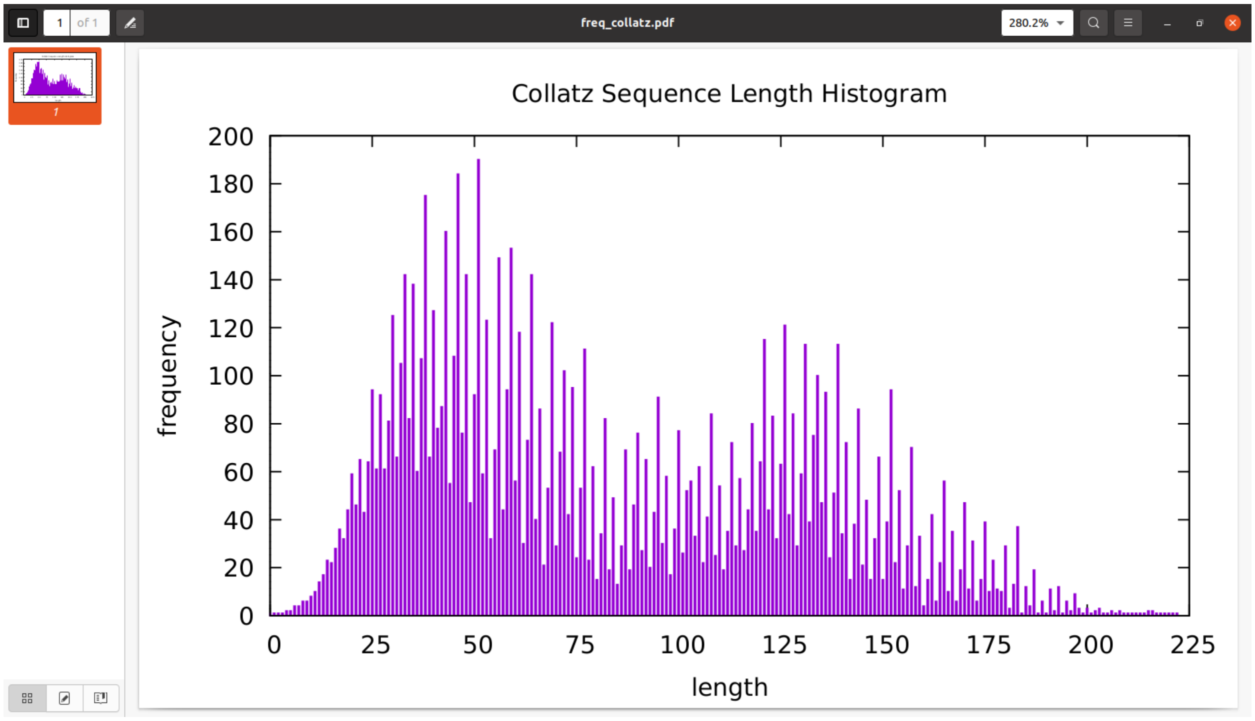
* The reason we delete each data file at the beginning of the program is so that data is not perpetually appended to the existing files across multiple runs of the program. We cannot solve this problem by writing to the data files instead of appending to them, because then, only the length and maximum for the last collatz sequence built under the for loop will be stored in the corresponding data files.

** The data file for length frequencies is built using the data file of the sequence lengths in the following way. First, the second column (the lengths) is isolated and sorted. Whether it is sorted lexically or numerically is irrelevant, because we are only looking to create groupings of equal lengths, each of which will be assigned a count value. Next, in order for each set of values to be assigned to the correct axis, the two columns are switched. After a final numerical sorting, the data are appended to the data file for length frequencies. All of the aforementioned is done using a single command.

Final Expected Output




1
of 1
max_collatz.pdf
280.2%
1
Maximum Collatz Sequence Value
value
0
20000
40000
60000
80000
100000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
n


1
of 1
freq_collatz.pdf
280.2%
1
Collatz Sequence Length Histogram
frequency
0
20
40
60
80
100
120
140
160
180
200
0
25
50
75
100
125
150
175
200
225
length