

## Computer Science 1510 Assignment #7

---

- This assignment requires electronic submission of your source code files. Follow the directions under “Submission Details for All Assignments” on the “Links” tab on the course webpage to submit your assignment.
  - Your submission should also include postscript files of your graphs for Question 2(d), as well as one or more plain text files (which can be created using the editor that you use to create your source code files) containing your answers to the questions related to the behaviour of your code.
  - It is not necessary to submit hard (printed) copies of your assignment.
  - Be sure to include sufficient comments in your code, and labels in your output.
- 

1. Given a function  $f(x)$ , the derivative,  $f'(x)$  is often expensive or impossible to compute analytically. One method of approximating the derivative is to use a finite difference,

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

where  $h$  is small. Write a C program to approximate  $f'(x)$  at  $x = 2$  for the function  $f(x) = x^2 - 3x + 2$ , declaring your variables as `floats`. Your program should ask the user for a value for  $x$ , a starting value for  $h$ , the number of approximations to be computed, and the factor by which  $h$  should decrease between consecutive approximations. You should include a separate function to obtain  $f(x)$  values for a given value of  $x$ . Test your code starting with  $h = 1$  and decreasing  $h$  by a factor of 10 on each iteration. Compare your results to the exact value. What value of  $h$  gave the most accurate result? What happened to your approximations when  $h$  was smaller than this value? Try changing your variables to type `double` and repeat the experiment.

2. Suppose that you have data collected from an experiment consisting of ordered pairs of points  $(x, y)$ , and you want to compute a “line of best fit.” You would need to compute values for the slope  $m$ , and the  $y$ -intercept  $b$  of a line  $y = mx + b$  that in some way best fits the data. In the method of least squares, the best fit is obtained by minimizing the sum of the squares of the deviations of the observed  $y$ -values ( $y_i$ ) from the predicted  $y$ -values ( $mx_i + b$ ),

$$\sum_{i=1}^n [y_i - (mx_i + b)]^2$$

where  $n$  is the number of data points. Minimizing this function gives,

$$m = \frac{(\sum xy) - (\sum x)\bar{y}}{(\sum x^2) - (\sum x)\bar{x}} \quad b = \bar{y} - m\bar{x}$$

where  $\bar{x}$  and  $\bar{y}$  denote the mean of the  $x$  and  $y$  values respectively.

For some data, a line may not be such a good fit, or in some cases the data may be very noisy and hence difficult to fit well. A standard measure of the goodness of a particular fit is called a correlation coefficient, given by,

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}.$$

A value of  $|r|$  near 1 is indicative of a good fit, while a value near 0 indicates a poor fit (ie. the data does not exhibit a linear relationship).

- (a) Write a C program to read an unknown number of  $(x, y)$  pairs from a file (ie. read to the end of the file), compute and display the line of best fit using linear least squares, and compute and display the correlation coefficient for your fit. You should have the user input the name of the file containing the data. Note that at least two points are required to define a line, therefore your code should ensure that the data file contains at least two points.
- (b) Test your code on each of the following data sets.

$x$	$y$			$x$	$y$
0.25	0.036			1	0.395780
0.375	0.037			2	0.283273
0.5	0.039			3	0.382400
0.75	0.042	$x$	$y$	4	0.357503
1.0	0.046	5	0.99999	5	0.390247
1.25	0.050	15	0.99913	6	0.390247
1.5	0.053	25	0.99707	7	0.429491
2.0	0.058	35	0.99406	8	0.372722
2.5	0.065	45	0.99024	9	0.383922
3.0	0.073	55	0.98573	10	0.303953
3.5	0.078				
4.0	0.085				
4.5	0.093				
5.0	0.102				

Did you obtain a good fit in each case? (If a correlation coefficient of  $|r| < 0.3$  is obtained, you can assume that the fit is poor).

- (c) Once you have obtained a line of best fit for a data set, provided that it is a good fit, you can obtain a reasonable approximation of the value of the underlying function  $y = f(x)$  using this line. Add to your C program to have it ask the user for a value of  $x$  and compute and display an approximate value of  $y$ . This process of approximating the value of a function using an equation obtained from fitting data is called *interpolation* and *extrapolation*, depending on whether the requested point is within the existing data bounds, or outside it, respectively.

- (d) Use a plotting package (such as Grace or gnuplot) to create graphs of each data set (using small filled circles for the data) along with your calculated line of best fit. Save the graphs as postscript files to include with your assignment submission. Refer to the documentation for the package of your choice for help with creating the graphs and files.