Emerald Genius Questions

William Casazza

31 December 2015

1 Qualitative

1.1 Volume Check Calibration [60 points]

1) (10 points) The function I chose to fit the data is

$$f(V) = \frac{40.33}{1 + e^{0.3258(V-6.489)}}$$

Although the data given appeared to fit a cubic equation better with an R-squared value of 0.96 Vs. the 0.94 value given, the character of the data, with clear upper and lower asymptotic behavior, appeared to be more characteristic of a negative logistic curve. Thus using MATLAB's standard curve-fitting tool, I was able to find the coefficients to the general-form negative logistic function: $f(V) = \frac{-L}{1+e^{-r(V-V_0)}}$.

2) (20 Points)

$$f(3) = \frac{40.33}{1 + e^{0.3258(3-6.489)}} = 30.54$$

Thus, the function predicts a distance of 30.54 mm. The root of the sum of squares of errors is used to combine independent errors (i.e. different measurement type, or model error). The Root-Mean Square Error (or Deviation) for the model is 2.237mm, the given the accuracy of the liquid and distance measurements are $\pm 0.15\mu L$ (or 5%) and 2% respectively. The error propagated by measurement is thus $\sqrt{0.05^2 + 0.02^2} * 32.1703mm = 1.732mm$, taking the sum of the square errors of measurement and that of the model, and then taking the square root, we get an overall error of $\sqrt{(1.732mm)^2 + (4.526mm)^2} = \pm 4.846mm$.

3) (20 Points)

$$f(7.2) = \frac{40.33}{1 + e^{0.3258(7.2-6.489)}} = 17.84$$

Thus, the function predicts a distance of 17.84mm. Calculating total error as in part (2): the measurement error is $\sqrt{0.021^2 + 0.02^2} * 16.004 = 0.4622$. It follows that the overall error is: $\sqrt{(0.4622mm)^2 + (2.237mm)^2} = \pm 2.248mm$.

4) (10 Points) Letting the function $f(V)$ represent the height of the well at volume V, an equation resulting in the radius of the well at any given volume is represented by the expression

$$r(V) = \sqrt{\frac{V}{h(V)\pi}}$$

Plotting this along with it's corresponding range of volumes yields the cross-section in figure 1.

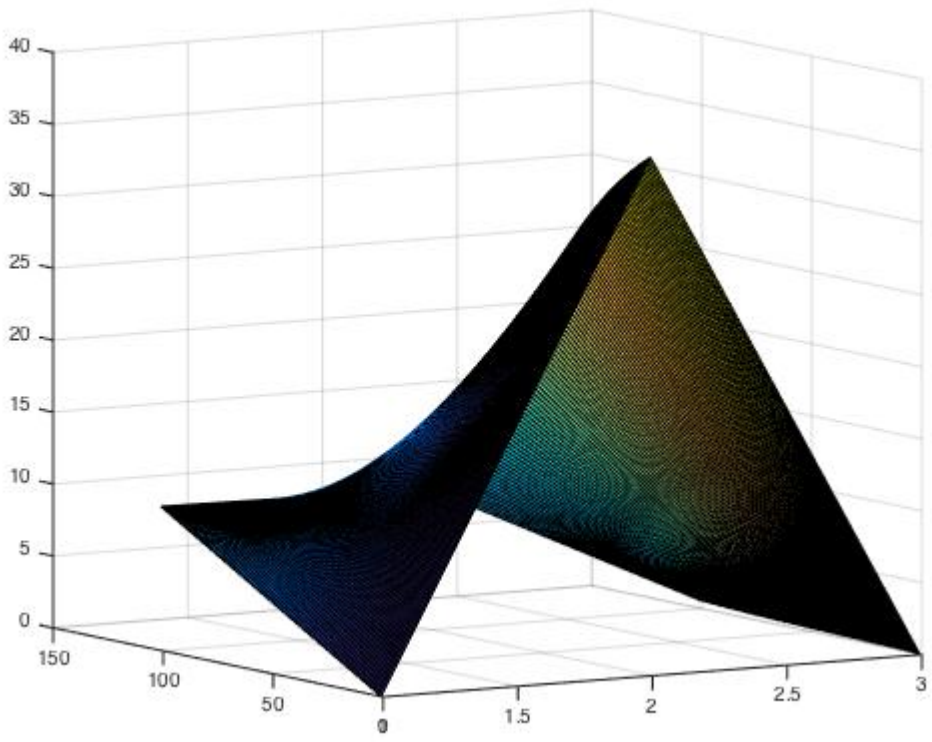

Figure 1: Cross Section of Well Shape

1.2 Vial Experiment [60 points]

1) (15 Points) When the volume of the water is less than around 5.37725 mL (i.e. the height of the water is less than $\sqrt{2}$ cm, the length of a bottom edge), the Volume of the water as a function of height is $V = \frac{h^2}{\tan(180-135-\tan^{-1}(\frac{0.5}{10}))} + 2\sqrt{2}h$. As a function of Volume, this becomes roughly $h = -2.05367 + 0.000116035\sqrt{1.07855 * 10^8 V + 3.13248 * 10^8}$. Above that volume (a water height above $\sqrt{2}$, or a volume of 5.37725 mL, the height as a function of volume becomes:

$$h = 3.04494 * 10^{-6}\sqrt{1.56624 * 10^{11}V - 6.26495 * 10^{11}}$$

2) (45 Points) Reusing some geometry from the previous part of the question, we get that the level of the water will reach the laser when its height is at approximately 3.08051 cm. This is a volume of approximately 10.5347 mL.Using an exponential fit on the cooling data (with an r-squared of 0.9792) we get the following expression for temperature T at time t: $T = 48.36e^{-0.1227t}$. The rate of change in temperature in degrees Celsius per hour is $\frac{dT}{dt} = -5.93377e^{-0.1227t} + 20$. The exposed surface area A_s in the part of the vial with which we are concerned is (in cm^2) $A_s = 2 * \frac{h}{\tan(45-\tan^{-1}(\frac{0.5}{10}))}$ (the rate for which is also solved). This rate solves to around 6.9377 mL/hour until the liquid hits a temperature of 20 Celsius at 7.195 hours, at which it will evaporate at 7.9969 mL/hour. At 7.195 hours, however, more than enough liquid will be lost. Thus we only need to solve the following equation for time t: $hours = \frac{40mL-10.5347mL}{6.9377mL/hour} = 4.2471hours = 254.8277minutes$. Thus the experiment must be completed in approximately **254.8277 minutes**.

1.3 Roomba [60 Points]

1) For convenience, the speed of the roomba in ft/s is 0.82 ft/s. Given that the Roomba has no recollection of its previous path, it is entirely possible for it to start at one wall, turn 180 degrees, and then move toward another wall to turn about again and oscillate in this fashion for 10 minutes. A diagram of this in path can be seen in figure 2. The following additional assumptions about the roomba are made:

   - The roomba cleans the entire area underneath itself.
   - The roomba turns in an arc along its circumference ($C = d\pi = \pi ft$) at the same speed at which it moves, making it rotate $\frac{2\pi rad}{\pi ft} * 0.82\frac{ft}{s} = 1.64\frac{rad}{s}$

   With this path and under these assumptions, the roomba would clean $10ft^2$, minus the parts the roomba cannot cover due to its shape $4(0.25ft^2 - 0.25(\pi(0.25ft)^2 = \frac{4-\pi}{4}ft^2$. The total area of the floor is $0.5*(10+6)*4+10*8 = 112ft^2$ thus the **total percentage of the room cleaned is:** $\frac{10-\frac{4-\pi}{4}}{112} * 100 = 8.737\%$

2) To determine how long on average the roomba would take to clean a room, the following assumptions are established in addition to those in part (1)L The degrees turned at each hit is taken from a uniformly random distribution. 95% of the area of the floor is calculated to be $106.4ft^2$. The roomba will only significantly overlap with it's last

trajectory within a 90 degree turn, thus this probability is around less than 50%. The amount of overlap unaccounted for by that region is for the most part negligible outside of a 45 degree turn clockwise or counter clockwise from the opposite of its original trajectory. For this reason, I will simply assume that the average amount of overlap for each new line is approximately $1ft^2$ for every path made. In addition, the roomba is only likely to be programmed to turn within $\leq$180 degrees of the direction it came from, since any more would be better accomplished from turning the other direction. This makes the maximum turning time around 1.916 seconds. We also approximate the area of the roomba to a square foot.

Thus, with these restrictions and assumptions, the amount cleaned at time t is $f(t) = 0.25\pi\frac{ft^2}{s}(t) - n(1 + turningtime)$ where n is the number of lines the roomba has traveled. The longest path traveled is along the longest diagonal at $\sqrt{244}$, and the shortest at 1 ft between corners, making the average amount cleaned at each path roughly 8.310 ft, in 6.814 seconds. The average turning time is roughly half the maximum at 0.9578 secondsThus making the previous equation $f(t) = 0.25\pi(t) - \frac{t}{6.814}(1 + 0.9578)$. When $f(t) = 106.4$, 95% of the room will be cleaned on average. This is at roughly 213.641 seconds.

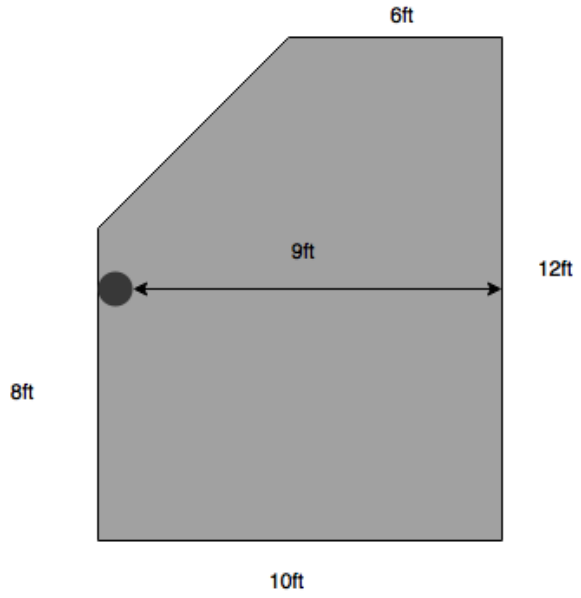

Figure 2: Possible Roomba Path

2 Qualitative

2.1 Lost in the Subway [50 Points]

1) Since the graph of the subway station only has two vertices (cities) with an odd out degree, it is proven that the graph has a eulerian circuit. In other words, there is a path through

the graph that visits each edge exactly once, with the odd degree vertices beginning and ending the path (i.e. San Francisco and Denver). The algorithm for this was taken from "http://www.graph-magics.com/articles/euler.php", and my implementation in python 2.7 can be found in the file emeraldSubwayQuestion.py (the txt file used is also attached). To find the device, the following path can be traveled in 1243 minutes of riding time:

['San Francisco', 'Great Falls', 'Minneapolis', 'Medford', 'Omaha', 'Denver', 'Great Falls', 'Salt Lake City', 'Denver', 'Dallas', 'Albuquerque', 'Phoenix', 'Los Angeles', 'Chicago', 'Minneapolis', 'Detroit', 'Chicago', 'Omaha', 'Minneapolis', 'Memphis', 'Jacksonville', 'Phoenix', 'Seattle', 'Medford', 'San Francisco', 'Los Angeles', 'New York', 'Detroit', 'Washington', 'Memphis', 'Dallas', 'Omaha', 'Great Falls', 'Seattle', 'Jacksonville', 'Charlotte', 'Chicago', 'Memphis', 'Charlotte', 'Washington', 'New York', 'Boston', 'San Francisco', 'Salt Lake City', 'Albuquerque', 'Denver']

In addition to the 195 minutes already spent, this is a total of 1438 minutes spent total, leaving 2 minutes to spare.

2.2 Reverse Engineering [60 Points]

1) For this question (and subsequent questions) I consulted a friend in Electrical and Computer Engineering and Carnegie Mellon named Niteesh Sundaram so as to learn about shift registers. We can represent the input to D as a concatenated bitstring output $B\hat{C}$, where B is the string stored in B, C in C, and $\hat{}$is concatenation, and B and C are changed at the same clocktime. The total possible number of strings for $B\hat{C}$ is 2^6 since there are 6 spots that hold 1 of 2 possible values. Since A can be thought of as putting out 2 inputs in parallel in terms of B-C clocktime, we can think of the system as using 2-bit commands to simultaneously control the output of B and C individually each clock step. It is possible to then generate 2^6 strings with 2^6 2-bit commands, thus the minimum length input required is $2^7 - 2$, counting the string 00 as an input.

2) The following algorithm was done out by hand to generate the input string to give all possible outputs without repeating any output:

   - get all bitstrings where either C or B is set to 0 with the set of input commands: '0010101000100000' (read from the machine left to right in 2-bitwords, where the first character is read in to B, and the second is read in to C)
   - consider the input 2-bit string into A as composed of 1 input to B and one to C, offset C's commands by giving it a 1, while giving B another 0
   - Feed in the string '1110100' to each individual shifter until B and C read the same output for 7 outputs (since the string of all 0's is taken care of at initialization)

   Using the attached script "emeraldRevEngineer.py", one shortest possible set of commands read in left to right as explained is:"00101010001000000111111001100001011110110010010101101

2.3 International Encoding of Amino Acids [35 Points]

1) (15 points) Since the frequencies of each amino acid is known, you can use the Huffman coding algorithm to create a binary tree based off of the frequencies, prioritizing a grouping of higher frequency amino acids at the least depth in the Huffman tree (using a priority queue in the implementation), and the lowest frequencies the at the largest depth. The code for each amino acid is then obtained by assigning one symbol to each branch taken in the Huffman tree (i.e. left is symbol 1, right is symbol 2). The encoded string is then decoded by traversing the tree according to the directions specified in the code, restarting from the root after a leaf node is reached.

2) Using a binary Huffman implementation found at "http://planetcalc.com/2481/", the encoding in figure 3 saves a total of 15.04% fewer symbols than the original encoding.

3) Knowing the frequencies of the amino acids, one could simply use 4-ary Huffman encoding to represent the string of amino acids with less symbols. The only difference is that the data structure used to store a character's encoding is a 4-ary tree instead of a binary tree, i.e. the each non-leaf has 4 children, and each branch encodes one of 4 symbols in the symbol string encoding the amino acid ultimately at a leaf.

4) Using a javascript implementation of n-ary Huffman coding at

   "https://www.npmjs.com/package/n-ary-huffman"

   (implemented in "emeraldNary.js"), encoding with the alphabet of 4 symbols '0','1','2', and '3' (See figure 4) the data was able to be represented using 28.033% less symbols.

2.4 Anagrams in a Book [55 Points]

1) The longest word in the book (i.e. excluding the gutenberg project information at the end) is "characteristically".

2) A formatted padded version of the text can be found in the file "paddedKaramazov.txt"

3 Biology

3.1 Jurassic World [50 points]

A. (10 Points) To determine the organism the from which the mosquito last drew blood, first one would need to use a thin needle and syringe to extract blood from the stomach of the mosquito. Then use a scaled down version of the protocol found at:

   "http://www.genome.ou.edu/protocol

   _book/protocol_partIII.html#III.H"

   to extract genomic DNA from the blood. In addition, one would need to extract genomic DNA from the fly as well. The genome can then be sequenced in fragments, and pieced bac together using standard or next gen sequencing method. Using a BLAST-like

Symbol	Encoding
Serine	1111
Leucine	1110
Glycine	1100
Alanine	1011
Lysine	1010
Valine	1001
Threonine	0111
Aspartic Acid	0101
Glutamic Acid	0100
Proline	0011
Asparagine	0010
Arginine	0001
Phenylalanine	0000
Isoleucine	11011
Glutamine	11010
Tyrosine	10001
Cystine	10000
Histidine	01100
Methionine	011011
Tryptophan	0110101
Stop Codons	0110100

Figure 3: Huffman encoding of the amino acids using a binary alphabet

algorithm on a database of creatures known to exist at the time at which the mosquito is dated, one can look for sequence motifs in the mosquito and blood derived DNA, looking for matches to sequences characteristic to genes found within specific groups of organisms.

B. (15 Points) In Jurassic Park, it is suggested that the gaps in a genome could be filled in with the DNA of another species, however, the likelihood of this producing a viable genome is very low, given the complexity of gene expression. Thus, the only real hope would be to extract as many DNA samples as possible, ligating necessary DNA fragments back together with a varied array of primers over the course of multiple PCR runs. It is possible, albeit unlikely, that the DNA from one blood sample from a mosquito could contain the whole genome yet in different fragments. It is also possible that DNA that is more evolutionarily close to that of the Dinosaur DNA found could be used in the manner described to reconstruct the genome, granted the sequence used does not interfere destructively with regular gene expression.

C. (15 Points) In order to clone a baby dinosaur, one would need to find an organism that lays a egg very similar in composition to what dinosaurs are thought to have laid as seen in fossilized evidence. The most closely related bird to the dinosaur may suffice. The unfertilized oocyte can then be injected with the replicated dinosaur genome packed into a nuclear structure if possible. This process is called enucleation (see Wikipedia "Dolly (sheep)"). The egg is then given direct current to stimulat cell divisions which go on to

Alanine	"03"
Arginine	"30"
Asparagine	"23"
Aspartic Acid	"20"
Cystine	"000"
Glutamic Acid	"21"
Glutamine	"33"
Glycine	"10"
Histidine	"003"
Isoleucine	"32"
Leucine	"02"
Lysine	"11"
Methionine	"0020"
Phenylalanine	"31"
Proline	"22"
Serine	"01"
Threonine	"13"
Tryptophan	"0021"
Tyrosine	"001"
Valine	"12"
Stop Codons	"0022"

Figure 4: 4-ary Huffman encoding of the amino acids

form a blastocyst. if incubated properly, there is a small chance this fertilized egg could develop into a dinosaur.

D. (5 points) In the attempts made to clone extinct animals, all resulting clones have died. his suggests that it is possible that a high level of compatibility between the genome of the clone and it's surrogate environment must be achieved in order to produce a successful offspring, barring new scientific discoveries. Thus it would seem that close to 100% of the genome is probably a minimum requirement for cloning, and the identity must be incredibly high to that of the surrogate mother.

E. (5 Points) Based on a study at "http://www.livescience.com/23861-fossil-dna-half-life.html", the half life of DNA in Bone is somewhere around 521 years. This would give around 27 days, or perhaps several months is the oldest age of blood meal required to clone an organism in this age.

3.2 Directed Evolution [60 Points]

1) (30 points) After consulting an article on : "http://bitesizebio.com/8252/what-can-nmr-do-for-you-part-one/", quick way to tell whether or not a protein is folded is to carry out Nuclear Magnetic Resonance Spectroscopy on a high concentration sample of the protein, which works by reading response of the magnetic field of hydrogen nuclei (protons) to an electromagnetic field. Thus after obtaining said sample through protein extraction (either through gradated centrifugation or Ni-NTA bead extraction if the proteins contain a his tag), the sample can be loaded into an NMR machine using a standard NMR protocol depending on the machine. Since an unfolded protein has its hydrogens in an environment

with less restriction, the chemical shift (broadness) of a peak on an NMR readout will be larger. Thus more restricted protons in a folded protein will have narrower, more distinct peaks. A sample of heat denatured protein can be used as a reference for unfolded protein in the assay.

2) (30 Points) If the epitope is caught in an intramembrane domain, a binder must be designed to out-compete the array of proteins in the extracellular space between yeast in vitro. Thus in a new round of mutagenic proteins, the yeast surface should be modified to include several more extracellular proteins, as extracted and identified from a-non detergent based dissolution of a yeast colony, as detergent would lkely break open the membrane of the yeast, muddling access to specific intracellular proteins. Once this new yeast display is created, the binders can be allowed to undergo directed evolution in vitro with the surface until binding is reached. This will hopefully result in a more successful in vivo test.

4 Powers of Estimation

1) According to information from IDT at:
"https://www.idtdna.com/pages/docs/educational-resources/molecular-facts-and-figures.pdf?sfvrsn=4"

The mass of DNA in the human Genome, which is not far off in magnitude from that of most single and multicellular organisms, is roughly $3.59x10^{-12}g$. The number of eukaryotic cells in a human body is stated as $1.0x10^{14}$. There are billions of humans and other human sized organisms, in addition to bacteria and other small microorganisms with genome sizes within a few orders of magnitude of the size of the human genome. Thus taking say $1.0X10^{30}$ as the number of eukaryotic cells in multicellular organisms of human sized as containing around 30% of the earth's DNA mass, we have around $1.96x10^{1}9$ grams of DNA on earth, or $2.638x10^{16}$ pounds of DNA on earth.

4) The Current Life expectancy in the united states is tabulated from the US census, and according to the Social Security Administration's website it is currently 76.18 years for males, and 80.95 years for females. The combined average of the two (assuming a roughly equal male and female population in america), is 78.57 years. According to data on mortality from the Center for Disease Control's website, the number of deaths from viral infection is on the order of 55,000, out of a total 2,515,458 at the time of the census, for a total of 2.186% of all death. The census predicts that the current US population is 322,762,018. Based on some rough estimation and the calculation of life expectancy as described on wolfram alpha, this would make the revised life expectancy at close to the current number, at around 78.58 years on average for both genders.

5 Physical Chemistry

5.1 Kinetics [60 Points]

A. (10 Points) Vague attempt made in "emeraldKinetics.m"

5.2 ThermoDynamics [40 Points]

1) (5 Points) According to wikipedia and the Integrated DNA Technologies (IDT) website, the average extinction coefficient of single stranded DNA is $0.027(\mu g/ml)^{-1}cm^{-1}$.

2) (5 Points) Using the equation found at "http://www4.ncsu.edu/ franzen/public_html/ CH433/lecture/Kinetics_Second_Order.pdf"

   The expected equilibrium constant of hybridization is k = 180277.56 L/(mol*s).