University of Glasgow | School of Computing Science

DayOfWeek DayOfMonth Month 2018

XX.XX am/pm – XX.XX am/pm

(Duration: 1.5 hours)

EXAMINATION FOR

DEGREES OF MSci, MEng, BEng, BSc, MA and MA (Social Sciences)

Data Fundamentals (H)

Answer 2 of 3 Questions

This examination paper is worth a total of 50 marks

For examinations of at least 2 hours duration, no candidate shall be allowed to leave the examination room within the first hour or the last half-hour of the examination

INSTRUCTIONS TO INVIGILATORS

Please collect all exam question papers and return to School together with exam answer scripts

ANSWERS ARE HIGHLIGHTED. REMOVE FROM FINAL PAPER.

1.

(a) You are presented with this graph, representing the flow of oil in **kL/s** through a high-pressure oil pipeline as a function of the pressure of the pipeline, in **mPa**. The theoretical maximum flow is also shown.

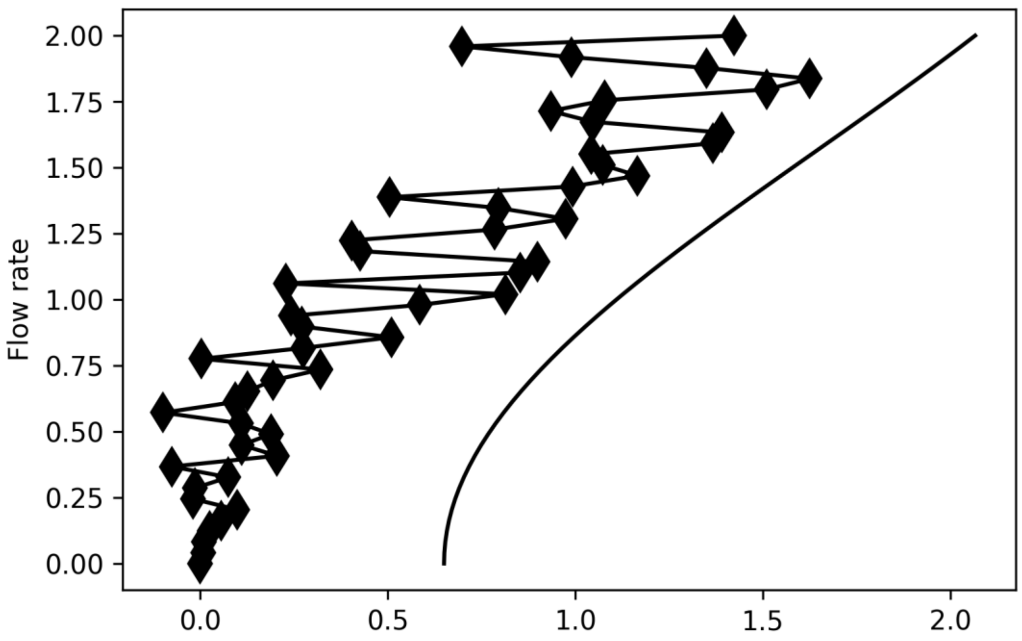

Figure 1: A plot of the flow of oil as pipeline pressure changes, in comparison to the maximum theoretical flow.

Provide a detailed criticism of this graph as an accurate scientific visualisation. You should list any omitted details as well as any objectively poor choices.

[6]

Any of the following are acceptable, [#1] each, up to 6 marks

This graph has the axes the wrong way around (dependent variable should be on y)

Units are not provided for either axis

No axis label on x axis

No title

No legend

Markers are excessively sized

Colouring/styling of lines is indistinct

No error bars

(b) In the terminology of the Layered Grammar of Graphics, this is a *layered* figure with *line geoms* and *point geoms*.

(i) Explain the distinction between faceting and layering in presenting multiple views of a single dataset, and the implications on *coords* used.

[3]

A faceted approach shows two views of the same datasets on *different* coordinate systems (coords) [#1]. A layered approach shows two views of the same dataset on the same set of coords [#1], and thus the same mapping and scaling of attributes must be applied to the data attributes [#1]

(ii) The visualisation director wants to show the vibration level on the pipeline as well as the flow rate at each observation point marked on the graph. Suggest two distinct ways the point geoms could be modified to display this.

[2]

The point geoms could have vibration mapped to colour using a colour map, or their size mapped to vibration.

(c) The mapping of data attributes (pressure, flow) to visual coordinates (x,y) involves a scaling operation. The graphics layout engine uses a linear transform to map from data attributes to visual attributes. This is represented as a 2x2 matrix A.

(i) If this transformation is *purely* scaling, what can you deduce about the structure of A?

[2]

A must be diagonal if it is a pure scaling matrix, with 0 in the off diagonal entries [#2]

(ii) Given an (x,y) *visual coordinate* as a vector **v,** and a general transformation matrix A that might include rotations as well as scaling, explain how the corresponding value in terms of the original data values, as a vector **y**, could be computed.

[2]

This could be computed by inverting A, computing **y** = A^{-1}**v** [#2]

(d) In an exploratory drilling project, the data for the expected density of oil deposits on the seabed is to be visualised as a 2D image, using a colourmap to map density at each spatial location to a specific colour.

The drilling company are currently using a "rainbow" mapping that maps each density to a different hue, with constant brightness. The visualisation director has heard that "*perceptually uniform monotonic brightness*" colour maps are better.

Explain carefully what this term means in terms of how data attributes are mapped to colours, and give one reason why a "rainbow" map can be a misleading way to visualise data.

[4]

Perceptually uniform means that equal steps in the data attributes appear to be equal steps in brightness [#1] *Monotonic brightness* means that the data attributes are mapped to colours which increase in apparent brightness as the attribute increases in value.[#1] Other attributes, such as hue can also be mapped to value but do not have to be [#1] A rainbow map runs the risk of introducing "false contours" [#1] into a plot where boundaries that do not exist may be hallucinated.

(e) The visualisation director is concerned that flow rate graph of (a) is considered to be "too spiky" and is masking the underlying trend. The director intends to use a *linear filter* to smooth out the data by convolving the data with a smoothing kernel.

(i) Explain what steps must be taken for such a filter to be meaningfully applied.

[3]

The data must be reinterpolated [#2] onto a regular grid / evenly spaced values [#1] before convolution is applied, or the results will be meaningless.

(ii) Give an example of a simple linear filter that might be applied.

[1]

A moving average [#1] is a linear filtering operation (with a constant kernel).

(iii) Suggest an alternative to linear filtering that might provide better results in this case, justifying your suggestion.

[2]

A nonlinear filter like a median filter [#1] would be better at rejecting spikes [#1] than a linear filter.

(Alternatively, linear/polynomial regression could be applied here, which would be a reasonable alternative suggestion).

2.

(a) IEEE754 defines the machine precision ϵ for **float64** to be $2^{-53} \approx 1.11\text{e-}16$. What does this imply in terms of the accuracy of real numbers approximated as floating point numbers in **float64**? Be precise, and give an equation to support your statement. *You should not perform any calculation.*

[4]

Machine precision defines the *relative* error [#1] between a real number and its floating point representation [#1]. It is defined for a real number x as

$$\epsilon \leq \frac{|x| - \text{float}(x)}{|x|}$$

[#1]
The IEEE754 guarantee ensures that the relative error in storing or computations with a floating point number in the representable range will not exceed 2^{-53} [#1]

(b) If an IEEE754 operation cannot be completed within machine precision, this will cause an *inexact* floating point exception. Describe two other floating point exceptions, and give an example of a numerical operation that would trigger this exception.

[6]

Any two of the following: [#1] for name, [#1] for operation

- **Invalid operation**, caused by 0/0, sqrt(-1.0), inf-inf, or any other invalid operation
- **Division by zero**, caused by x/0 for x not zero, nan or inf
- **Overflow**, caused by a computation having an result (absolutely) larger than the largest representable value
- **Underflow**, caused by a computation having a result (absolutely) smaller than the smallest representable value

(c) You have been asked to develop a model for modeling the flow of crude oil through a pipeline. Your client wants to maximise the flow rate of oil through the pipeline. The diameter of the pipe, the inner diameter of each of the 18 flow valves on the pipeline and the pipeline pressure can all be adjusted. The pipe cannot be more than 900mm in diameter or it will not fit onto the existing framework. The pipe pressure may not exceed 7.0MPa without risking rupture.

This is an optimisation problem. State the objective function, constraints and parameters in this model, and state a **simple** algorithm a junior developer could use to optimise this process. You may assume a simple mathematical model of the pipeline is available that estimates the flow rate for given pipe configuration. You do not have to describe the operation of the algorithm. You may disregard efficiency.

[5]

The objective function is the flow rate [#1] as a function of the parameters: pipe diameter, valve diameter and pipeline pressure [#1]. The maximum pipe diameter and pipe pressure is are constraints [#1].

Any zeroth-order algorithm would be acceptable, including random search, grid search, or any heuristic local search algorithm. [#2]

(d) **Gradient descent** could accelerate this optimisation. Give a brief outline of the operation of the gradient descent algorithm, including the update equation, and discuss in what circumstances gradient descent would fail to converge to an optimal solution. You may ignore the role of constraints in the solution.

[5]

Gradient descent steps along the locally steepest direction [#1] around the current parameter setting using the gradient vector computed at that point [#1]. It requires that the gradient of the objective function been evaluable at any point.

The update equation is

$$\theta_n = \theta_{n-1} - \delta \nabla L(\theta)$$

[#1]

Gradient descent will not converge to the optimal solution in the presence of local minima [#1] (alternatively could state that the step size might be set poorly, or that the gradient is not available)

(e) The client you are developing for declines to use automatic differentiation to implement gradient descent, as it is incompatible with their software stack. They opt to use **finite differences** instead. This is an approximation based on the definition of differentiation:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x-h)}{2h}$$

Explain how finite differences works and criticise this choice for this problem, both in terms of floating point numerical stability and computational efficiency.

[5]

Finite differences approximates the function evaluated at two closely spaced points [#1]. This is only accurate for small h [#1]. But this is numerically unstable for small h as it has both subtraction of two very similar values f(x+h) and f(x-h) [#1] as well as division by a very small value 2h [#1].

Furthermore, for a multidimensional case this must be evaluated in each direction [#1] around the current parameter estimate, which is extremely inefficient.

3.

(a) 0.05% of oil wells in a company's portfolio are known to be **high capacity deposits** – but which ones is unknown. A new test is devised to determine **high capacity** status. If the well *is* **high capacity**, the test comes back positive 99% of the time. If the well is *not* **high capacity**, the test still comes back with a false positive 10% of the time.

The test is run on a Well Q and comes back positive. The executive in charge decides to list the Well Q as a high capacity resource in the next year's financial report, saying it is 90% likely that it is **high capacity** given the test result.

(i) Explain clearly why this is not correct and explain, *giving a formula*, how to correctly compute the probability of Well Q being high-capacity. You **do not** need to perform the calculation.

This is the conditional probability fallacy (or similar wording) [#1] P(A|B) != P(B|A) [#1] The correct answer is given by Bayes' Rule [#1]

P(Q|T) = P(T|Q)P(Q)/P(T) =

P(T|Q)P(Q) / (P(T|Q)P(Q) + P(~T|Q)P(~Q))

[#2] for basic Bayes' rule [#1] for integrating over evidence

[6]

(ii) If 10 random wells are chosen randomly from the companies portfolio, what is the **log probability** of them *all* being high-capacity? You do not need to calculate the answer numerically; provide a formula.

[2]

log P(all Q) = 10 log(0.0005)
or
log P(all Q) = 10 log(Qany)
or any similar expression

[#1] for probability of independent events multiplying
[#1] for log probability changing to summation

(b) The flow of oil among a network of pumping stations is modeled using a linear model in discrete time, with daily time units. This model collects the current oil level at each station on each day t a vector $\mathbf{x_t}$ of length N. The flow is modeled using a $N \times N$ square matrix A, which does not change over time. The predictive model is stated in the form:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t$$

(i) Explain how to compute what the oil level at each station was exactly one week ago, from a vector of current oil levels $\mathbf{x}_t$.

[3]

This can be computed as $x_{t-7} = A^{-7}x_t$ [#2] , assuming that A is nonsingular. [#1]

(ii) State conditions under which the estimation of previous oil levels in (i) is possible, *in terms of eigenvalues of the matrix* A.

[2]

A matrix is nonsingular if $\det(A) \neq 0$, and since $\det(A) = \prod_i \lambda_i$ this is equivalent to having no zero eigenvalues. [#2]

(iii) The computation in (i) might be *theoretically* possible but numerically unstable in floating point computations. State how the *SVD* would factorise A and discuss how this would help determine whether numerical instability in the flow calculation is likely to be a problem.

[5]

The SVD of A will factorise it into three matrices $U\Sigma V^T$ (or $U\Sigma V$ is also acceptable) [#2]. Σ is a diagonal matrix of singular values [#1]. The condition number is the ratio between the smallest and largest nonzero entries in Σ [#1]. If the condition number is large, instability is likely [#1].

(c) The pipeline flow routing is to be maximised algorithmically. There are 230 continuously varying parameters that can be adjusted to configure flow between stations. Small changes to these parameters have a small effect on the flow.

(i) Two options proposed for the search are *hill climbing* and *grid search*. Recommend one of these two approaches, justifying your answer.

[3]

This problem has many dimensions [#1]. Grid search is likely to be able to explore only a very small part of the space [#1]. Hill-climbing algorithms will be able to take advantage of locality of the problem. [#1]

(ii) The original optimisation problem in (i) only maximised total flow. The operations manager is concerned that the financial cost of the routing should be taken into account. Suggest an approach that could combine these two objective functions to form a single objective function amenable to standard optimisers, and discuss any additional input that would required.

[4]

One approach is to form a convex sum of sub-objective functions [#2], where $L(\theta) = \lambda_1 L(\theta_1) + \lambda_2 L(\theta_2)$. [#1] The weighting of the two factors λ_1, λ_2 must be set to trade off these requirements. [#1]