University of Glasgow | School of Computing Science

DayOfWeek DayOfMonth Month 2018

XX.XX am/pm – XX.XX am/pm

(Duration: 1.5 hours)

SPECIMEN 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

This a *specimen* paper. It has not been through the vetting process and errors may still be present.

1.

(a) Given a hypothesis *H* and some observed data *D*, Explain how Bayes' rule relates prior belief, posterior belief, likelihood*,* and evidence, giving appropriate equations.

[3]

(b) In the fabrication of semiconductors, an industrial manufacturer has a process that produces silicon wafers. Sometimes, the wafers come out defective, and have to be discarded.

(i) Every wafer has a production cost *c*. Every correct wafer can be sold for *v;* defective wafers cannot be sold and bring in no money. Assuming the probability of a defective wafer is written *P(D=1),* where D is a random variable that may take on the values 0 and 1, write down the expression for the expected **profit** of manufacturing a single wafer.

[6]

(ii) The defect rate in one production line is 1:100000. The fabrication company learns of a new device that can predict whether the wafer will be defective after only a few seconds with a reliability of 99%. This will allow the fabricator to pull the wafer and abort the production process before the more expensive lithography process begins.

Explain how Bayes' rule would help the fabricator decide if this device is a worthwhile investment and give your recommendation based on these figures. You do **not** have to provide an exact calculation but you should give approximate figures. State any assumptions you make.

[6]

(c) The production system is to be optimized to minimize the number of defects.

(i) The fabrication process has adjustable temperature, dopant bias and slice thickness. The temperature cannot exceed 900C. State the objective function, parameters and constraints in this optimization problem.

[3]

(ii) The fabricator is considering optimizing the defect rate using **simulated annealing**. Outline how the simulated annealing metaheuristic works, and describe in what situations in might be a better choice than random local search as applied in stochastic hillclimbing.

[3]

(iii) In the context of testing optimization algorithms, what relevant diagnostic could be visualised to evaluate how well the optimization is going?

[2]

(iv) After discussion with the plant engineers, it is decided that the process is to be optimized to minimize defects *and* keep the temperature of the baking process as small as possible. Describe how this relates to the concept of Pareto optimality and discuss why a Pareto optimization approach could be superior to a convex sum approach.

[3]

2. You are building a suite of scientific software which has to eventually be GPU accelerated. This means that the code has to be vectorised.

(a) In writing software for a vectorised architecture, what programming pattern should be **avoided** in order to take advantage of hardware vectorisation support?

[2]

(b) In NumPy, write a *vectorised* function which computes and returns the result of the equation below. **x** and **y** are two 1D vectors with the same length N.

$$z = \sum_{i=0}^{N-1} \mathbf{x}_i^{\frac{i}{k}} \mathbf{y}_i^{\frac{i}{k}}$$

The function should have the signature:

```
def compute(x,y,k):
    ...
    return z
```

[6]

(c) The software will do computations on arrays of IEEE754 float64 numbers.

(i) If the exponent of a float64 has bit pattern 11111111111 (all ones) what three distinct values could this represent, and how are they distinguished?

[3]

(ii) Explain the role of the implied leading 1 in the mantissa, and state why this is a useful approach in binary floating point representations, but not in decimal.

[3]

(iii) You are asked to perform a floating point calculation of the form *((a - b) + c) / d*. Give three different sources of floating point inaccuracy that might occur in this computation, and properties of *a,b,c* and *d* that would trigger these problems. Assume that *a,b,c,d* are all nonzero finite IEEE754 float64s.

[3]

(d) You are asked to help optimise some vectorised GPU code, which uses a matrix multiplication and inversion to compute the equation:

$$\mathbf{y} = \mathrm{A}^{-1}\mathbf{x}$$

At the moment, the implementation uses the SVD**.** However, you have learned that A is in fact a **square diagonal** matrix. Suggest a much faster approach, and write vectorised NumPy code that computes the value of **y** from A and **x**. The diagonal of a matrix A is returned by `np.diag(A)`

[4]

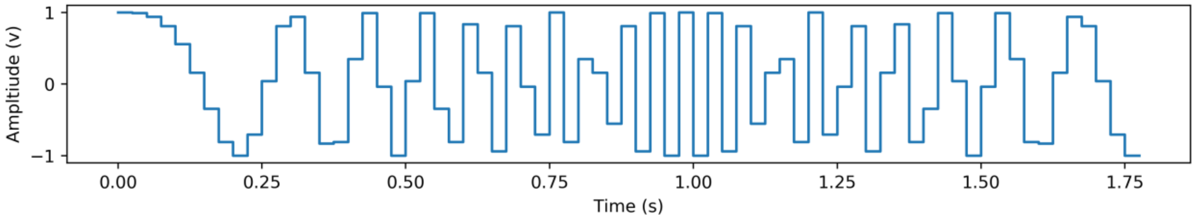
(e) Part of the vectorised software being developed will involve probabilistic calculations; in particular the calculating of the likelihood of multiple simultaneous system failures, each of which is very rare. After calculation, a computed probability has to be shown on an alphanumeric display so an operator can make a judgement as to whether to take action. Recommend appropriate choices for computation and display of these probabilities, justifying your choice.

[4]

3. (a) Define the Nyquist limit, specifying both its definition and its relation to the phenomenon of aliasing.

[3]

(b) You are involved in building a test system for a low-bandwidth seismic monitoring device. You are presented with the following graph, showing a plot of a regularly sampled signal from the device. This shows the measurement of a test signal which continuously increased in frequency, starting 0Hz and increasing by 20Hz every second. Estimate the sampling rate of the seismic device, and justify your reasoning.



[5]

(c) You are asked to produce a figure showing an ultrawideband seismic device's sensitivity across a range of frequencies. The device has three different settings: A, B, C, which influence the sensitivity response. For each setting, the response at each test frequency is measured 100 times to provide an estimate the uncertainty.

The caption for the figure is:

> *"Figure 1 shows the sensitivity of the prototype UWB seismic device as the input frequency, in Hz, is varied from 0.001Hz to 10kHz. The sensitivity in the settings A, B, and C are shown. Measured sensitivity ranges from 0 to 500 V/mm s^{-2}"*

Sketch an outline of an appropriate figure, using *layering*. This figure will be reproduced in black and white. You do not have any knowledge of the data, so you may assume any reasonable form for the graph. Ensure all details are present.

[8]

(d) You are asked to detect portions of a seismic signal where the signal appears to "repeat" itself; that is there a short portion of activity which is very similar to one that has been sensed before. These repetitions are not regularly spaced, but could occur anywhere in the signal. Describe in high-level terms how you would build an algorithm to detect these repetitions, using the following ideas: *sliding window; the infinity norm; argmin.*

Do not write code. Explain, briefly, in pseudo-code or words how you would go about this, referring explicitly to the named terms.

[6]

(e) (i) If the *norm* of a vector captures a notion of "length", what geometric notion does the *inner product* capture?

[1]

(ii) A seismic measurement of a suspected nuclear blast might be represented as a 10,000 dimensional vector which could be compared against a catalogue of other (equally sized) vectors of known blasts. Give an argument as to why the standard Euclidean norm of the difference of two vectors $||\mathbf{x}\text{-}\mathbf{y}||_2$ might **not** be used as a measure of similarity in this application.

[2]