THE UNIVERSITY OF AUCKLAND

FIRST SEMESTER, 2016
Campus: City

Computer Science

Computational Science

(Time allowed: THREE hours)

NOTE: Attempt *all* questions

Use of calculators is NOT permitted.

Put your answers in the answer boxes provided below each question.

You may use the blank pages at the end of the exam script for scratch work, which will not be marked.

Section:	**A**	**B**	**C**	Total
Possible marks:	20	20	20	60
Awarded marks:				

SURNAME:

FIRSTNAME:

ID:

Section A: Linear Systems, Root Finding, Cellular Automata (20 marks)

1. (a) Define what it means for x to be a root of the function $f(x)$. In what circumstance would you choose to use the bisection method rather than Newton's method to find the roots of a function? [**2** marks]

(b) Apply one iteration of Newton's method to the function $f(x) = 2x^2 - 2x - 2$ starting from $x_0 = 1$. [**1** mark]

2. (a) Suppose that the linear system $\mathbf{Au} = \mathbf{b}$ is over-determined. Write down the associated normal equation which is found by left-multiplying by the transpose of the matrix $\mathbf{A}$ and provide the name of the solution found by solving the normal equation. [**1** mark]

(b) Show how the orthogonalisation of $\mathbf{A}$ into the product of matrices $\mathbf{Q}$ and $\mathbf{R}$ can be used to solve the normal equation. Make sure you state which properties of $\mathbf{Q}$ and $\mathbf{R}$ you are relying on at each step. [**2** marks]

(c) Why should we worry about stability when solving the normal equation and what is the most stable method we discussed in lectures for solving it? [**1** mark]

3. Suppose $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

(a) One of the eigenvalues of $\mathbf{A}^\mathsf{T}\mathbf{A}$ is 6 and corresponds to the eigenvector $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$. What is the eigenvalue of $\mathbf{A}^\mathsf{T}\mathbf{A}$ corresponding to the eigenvector $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$? [**1** mark]

(b) The first two unnormalised eigenvectors of $\mathbf{A}\mathbf{A}^{\mathsf{T}}$, given in order of decreasing eigenvalues, are

$$\begin{bmatrix} 5 \\ 2 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix}.$$

Normalise these eigenvectors. [**1** mark]

(c) What are the eigenvalues corresponding to the normalised eigenvectors in the Question 3b? [**1** mark]

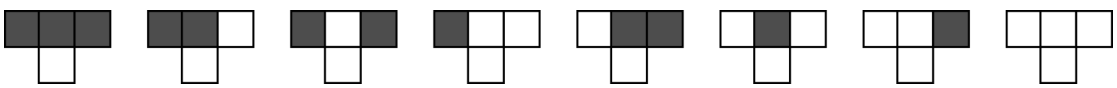
(d) Write down the singular value decomposition for $\mathbf{A}$. [**2** marks]

(e) Write down the pseudo-inverse of $\mathbf{A}$. [**1** mark]

4. Let $\mathbf{A}$ be a $m \times n$ measurement matrix where each column represents the result of an experiment where m measurements are taken, so each row represents a particular measurement taken across n experiments. How can the principal components of $\mathbf{A}$ be found using singular value decomposition? [**2** marks]

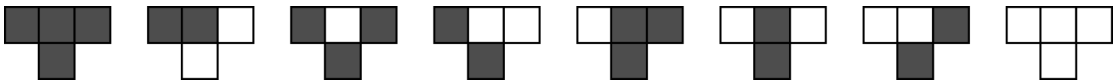
5. (a) Briefly describe each of Wolfram's four classes of cellular automata, using 1-2 sentences each. [**2** marks]

   (b) The figure below indicates a set of update rules for a one-dimensional cellular automaton.

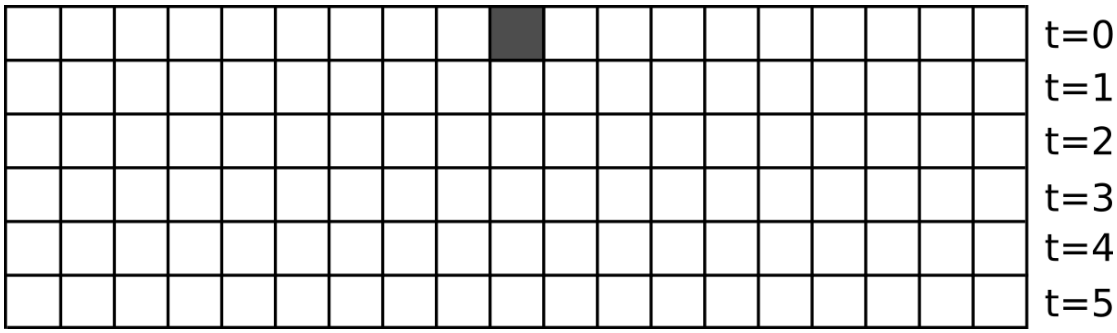


   What Wolfram-class does this CA belong to? Justify your answer in one or two sentences [**1** mark]

(c) The figure below indicates another set of update rules for a one-dimensional cellular automaton



Using these update rules, fill in the grid below, showing how the initial condition indicated by row t=0 evolves over 5 iterations. [**1** mark]



(d) Given the result that you produced above, what Wolfram-class does the CA in Question 5c likely belong to? Justify your answer in one or two sentences. [**1** mark]

ID: _____

Section B: Sampling and Markov models (20 marks)

6. (a) Suppose you have a method `PoissonProcess(`λ`, t)` which returns an array of event times of a Poisson process run at intensity λ for time t. Write a pseudo-code method `randpoiss(`λ`)` for sampling Poisson random variables with parameter λ. [**1** mark]

(b) Given a method `randunif()` that returns a uniform random variate between 0 and 1, write a pseudo-code method, `sample(x,prob)`, which takes the arrays `x` and `prob` of equal length such that the sum of the elements of `prob` is 1 and returns `x[i]` with probability `prob[i]`. [**2** marks]

(c) Given methods `sample` and `randpoiss`, sketch a pseudocode method `mutate(A, L, mu,t)` that takes a sequence A of length `L` and mutates it according to the Jukes-Cantor model over time t with mutation parameter μ. Allow mutations from a base to itself. [**3** marks]

7. What does it mean for a sequence of random variables $X_0, X_1, X_2, \ldots$ to have the Markov property? Express your answer in plain English and in mathematical notation. [**1** mark]

8. Suppose we are modelling regions of DNA sequences as being either being rich in purines, rich in pyrimidines or having a balance of the two. We use an HMM with three states, U for purine rich, Y for pyrimidine rich and B for balanced. A U state is followed by a Y state 5% of the time, and a B state 10% of the time. Y is followed by U 10% of the time and B 10% of the time. A B state is followed by any state with equal probability. The process starts in any state with equal probability.

   Each state can emit any of the 4 bases A, C, G or T. Emission probabilities are given by

	A	C	G	T
U	0.4	0.1	0.4	0.1
Y	0.1	0.4	0.1	0.4
B	0.25	0.25	0.25	0.25

   (a) Sketch a diagram of the HMM, showing all states, possible transitions and transition probabilities. Include the begin state. [**2** marks]

   (b) Given a state sequence π, explain why the length of a run of Ys in π is geometrically distributed with parameter 0.2. [**2** marks]

(c) What is the joint probability $P(x, \pi)$ of the state sequence $\pi = UB$ and the symbol sequence $x = GT$? (You can leave your answer as a product or sum of numbers). [**1** mark]

(d) The forward algorithm calculates the quantity $P(x)$ by iteratively filling out the matrix f. For the symbol sequence $x = ACA$, what probability does $f_Y(3)$ represent and what other quantities do you need to calculate $P(x)$ assuming that the end state is not modeled? [**1** mark]

(e) Using the correct notation and terms obtained from the forward and backward algorithms, write an expression for the posterior probability of being in state Y at time 3 given the sequence $x = ACA$. [**1** mark]

(f) For the HMM discussed here and a sequence x of length L, what is the algorithmic complexity of calculating $P(x)$ via the forward algorithm? What is the complexity of the naive method of directly calculating the sum $P(x) = \sum_\pi P(x, \pi)$? [**1** mark]

(g) Suppose that the emission and transition probabilities of the HMM discussed here were unknown but that a state sequence, π, and a symbol sequence, x had been observed. Illustrate how the parameters could be estimated from $\pi = UUUYB$ and $x = AACGG$ and explain why pseudo-counts are used in this procedure. [**2** marks]

(h) What is the input and output of the Viterbi algorithm? What is the value of z in the Viterbi matrix (you can leave your answer as a sum or product of numbers)?

	0	A	C
0	1	0	0
U	0	0.1333	0.0113
Y	0	0.0333	0.0111
B	0	0.0833	z

[**2** marks]

(i) What are the inputs and outputs of the Baum-Welch algorithm and why may it be necessary to run it several times on the same problem? [**1** mark]

Section C: Alignment and phylogenetics (20 marks)

9. (a) Let the symmetric matrix

$$D = \begin{pmatrix} 0 & & & \\ 6 & 0 & & \\ 4 & 6 & 0 & \\ 2 & 6 & 4 & 0 \end{pmatrix}$$

specify the pairwise distances, D_{ij}, between the four sequences $x_1, \ldots, x_4$. Construct a UPGMA tree from D showing your working. [**2** marks]

(b) Consider a phylogeny representing the ancestral relationships between species that were all sampled at the same time. If the distance to the root of the tree is not the same for all leaves, what can we infer about the mutation process giving rise to these species? [**1** mark]

(c) What is an advantage of using UPGMA over neighbour-joining and what is an advantage of using neighbour-joining over UPGMA? [**1** mark]

10. (a) What is the space complexity and what is the time complexity of using a dynamic programming approach to find a multiple sequence alignment of k sequences of length L and why do we prefer to use a heuristic approach? [**1** mark]

(b) Explain how the neutral X character is used in the Feng-Doolittle method for progressive alignment and what effect it has. [**1** mark]

11. Consider the four aligned sequences, `W`,`X`,`Y`, and `Z`:

```
     1234
W:   CACT
X:   ACAT
Y:   AAAT
Z:   GCCT
```

(a) What does it mean for site to be parsimony informative? Specify which sites in this alignment are parsimony informative. [**1** mark]

(b) By enumerating the three possible unrooted trees on the four taxa and finding the parsimony score for each tree, identify the maximum parsimony tree for this alignment. [**2** marks]

12. (a) Briefly describe a heuristic algorithm to find a tree with a low parsimony score and describe why it is not guaranteed to find the maximum parsimony tree. [**2** marks]

(b) Give two reasons to prefer likelihood based tree reconstruction methods over parsimony methods. [**1** mark]

13. Suppose we have a multiple sequence alignment D, a tree g and a mutation process with parameters μ. What assumptions do we typically make in order to make the calculation of the likelihood $P(D|g, \mu)$ tractable? [**2** marks]

14. The partially completed F matrix for calculating the local alignment of the sequences AGC and GACCT is given below. The score matrix is given by $s(a, b) = -2$ when $a \neq b$ and $s(a, a) = 4$ while the gap penalty is $d = -3$.

		G	A	C	C	T
	0	0	0	0	0	0
A	0	0	4	1	0	0
G	0	4	1	2	0	0
C	0	1	2	w	x	y

(a) Complete the matrix F by finding values for w, x and y. [**1** mark]

(b) Give the score for the best local alignment of these two sequences and provide an alignment that has this score. [**1** mark]

(c) Describe the changes required to transform the local alignment algorithm into the overlap alignment algorithm which does not penalise overhanging ends. [**2** marks]

15. Define an affine gap penalty, state why it is preferable to a linear penalty from a biological point of view and discuss issues of computational complexity when using an affine gap penalty for global alignment. [**2** marks]

ID: _____

Blank page — will not be marked

ID: _____

Blank page — will not be marked