COMPUTER SCIENCE TRIPOS Part II

Wednesday 3 June 2020 1.30 to 4.30

COMPUTER SCIENCE Paper 9

Answer ***five*** *questions.*

Submit the answers in five ***separate*** *bundles, each with its own cover sheet. On each cover sheet, write the numbers of* ***all*** *attempted questions, and circle the number of the question attached.*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Script paper	*Approved calculator permitted*
Blue cover sheets	
Tags	

1 Advanced Algorithms

(*a*) (*i*) What is the approximation ratio of an approximation algorithm? [2 marks]

(*ii*) State the definitions of PTAS and FPTAS. [4 marks]

(*b*) Consider the two approximation algorithms for VERTEX-COVER from the lectures (one greedy algorithm and one based on rounding a linear program).

(*i*) What are the approximation ratios of these two algorithms? [2 marks]

(*ii*) Construct an input graph that demonstrates the tightness of the approximation ratio of the greedy algorithm (for full marks, your construction should work for any even number of vertices n). [3 marks]

(*c*) Consider the following randomised algorithm to compute a solution of the VERTEX-COVER problem for an unweighted graph $G = (V, E)$:

```
Let C be the empty set
While E not empty do
       Pick any edge e={u,v} from E
       Choose x from {u,v} uniformly at random
       Add x to C
       Remove all edges incident to x from E
End While
Return C
```

(*i*) Explain briefly why the set C returned is a valid vertex cover. [2 marks]

(*ii*) Find a lower bound on the probability that the algorithms returns an optimal solution.
Hint: For each edge $e = \{u, v\}$ picked by the algorithm consider the event that the chosen vertex $x \in \{u, v\}$ added to C is also part of an optimal cover. [4 marks]

(*iii*) Given a lower bound $p \in (0, 1)$ on the probability that this algorithm returns an optimal solution, describe a new algorithm that returns an optimal solution with probability at least δ, for any given $\delta \in [p, 1)$. [3 marks]

2 Bioinformatics

(*a*) You are given a table of gene expression data. Each row corresponds to a gene and in the columns there is the gene expresson at different time steps (or different experimental conditions). Discuss at least one method, with one example, to identify genes with similar behaviour in time and the method's complexity and limitations. [4 marks]

(*b*) Discuss with one example how the stiffness parameter affects soft k-means clustering. [4 marks]

(*c*) Describe the advantage of using suffix arrays to find matches in genome sequencing. [4 marks]

(*d*) Describe solutions to the problem of 'bubbles' in De Bruijn graphs of genomes. [4 marks]

(*e*) Describe opportunities and challenges presented by DNA storage of data including a technique for indexed retrieval. [4 marks]

3 Business Studies

Giles Murchiston, a former PhD student now graduated, approaches you with a potential Part II project. He is interested in developing a new autonomous transport system for Cambridge and would like to you to conduct a feasibility study given some of the recent research results from the Department.

(*a*) Describe five types of intellectual property that might have arisen from research at the Department that Giles might need to consider licensing. [5 marks]

(*b*) Given that a new transport system is likely to need investment, describe five characteristics of the transport system you should investigate that investors would be interested in. [10 marks]

(*c*) When marketing the transport system describe five criteria Giles should consider in relation to customer adoption. [5 marks]

4 Comparative Architectures

(*a*) The performance of single-chip computers has improved rapidly over the past 40 years. Describe the major turning points in computer architecture as technology, applications and target markets have driven change. [10 marks]

(*b*) Your processor supports simultaneous multithreading (SMT) and has hardware support for two threads. What characteristics would you consider when deciding which two threads would be best to schedule together? [4 marks]

(*c*) An analysis of a multicore processor used within a datacentre suggests that its performance could be improved by increasing the number of threads supported in hardware. This could be done by adding more thread contexts to a core with support for simultaneous multithreading and/or adding more cores. Contrast these two different approaches and describe their limitations if we accept our area budget (i.e. die size) is fixed. [6 marks]

5 Computer Vision

(*a*) (*i*) For an image $I(x, y)$, define its gradient vector field $\vec{\nabla} I(x, y)$. [1 mark]

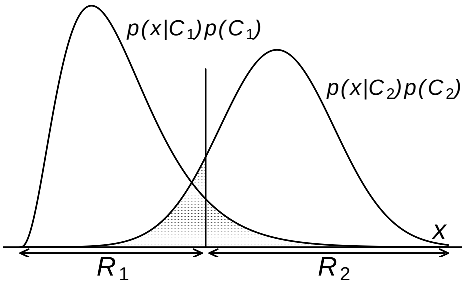
(*ii*) Why is this vector field a useful thing to compute? [1 mark]

(*iii*) Define the gradient magnitude over the image plane (x, y). [1 mark]

(*iv*) Define the gradient direction over the image plane (x, y). [1 mark]

(*v*) Explain how the gradient vector field is used in the Canny edge detector, the main steps in its use, and its advantages. [3 marks]

(*b*) A Bayesian classifier uses observations x to assign visual objects to either one of two classes, C_1 or C_2. Their baseline prior probabilities are $p(C_1)$ and $p(C_2)$, with sum $p(C_1) + p(C_2) = 1$. Observations x have unconditional probability $p(x)$, and the class-conditional probabilities of a given observation x are $p(x|C_1)$ and $p(x|C_2)$.



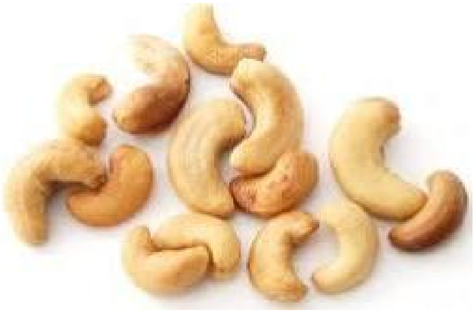
(*i*) Using the above quantities provide an expression for $p(C_k|x)$, the likelihood of class C_k given an observation x. (Here $k \in \{1, 2\}$.) [2 marks]

(*ii*) Provide a decision rule using $p(C_j|x)$ and $p(C_k|x)$ for assigning classes $\{C_j, C_k\}$ based on observations x, that will minimise errors. [2 marks]

(*iii*) Now express your decision rule instead using only the quantities $p(C_j)$, $p(C_k)$, $p(x|C_j)$, $p(x|C_k)$, and relate it to the diagram above. [1 mark]

(*iv*) If the classifier decision rule assigns class C_1 if $x \in R_1$, and C_2 if $x \in R_2$ as shown in the figure, what is the total probability of error? [2 marks]

(*c*) Propose an algorithm for shape classification that could correctly classify all of the objects shown here as cashew nuts, despite their variations in size (or hence distance), pose angles, colours, and intrinsic shapes. How can shape grammars, active contours, boundary descriptors, zeroes of curvature, and codon constraints enable a classifier to achieve those invariances?



[6 marks]

6 Cryptography

(*a*) Consider a message-authentication code Mac expected to provide *existential unforgeability under adaptive chosen-message attack.*

(*i*) What requirement does existential unforgeability impose on any padding function applied to the message by Mac and why? [4 marks]

(*ii*) What is an example of a padding function that satisfies that requirement? [2 marks]

(*b*) While reviewing the *MacGyver* burglar alarm system, you notice that a sensor S uses the following stream authentication protocol to report its status to the controller C once every second over a data wire:

$$\begin{array}{lclll}
C & \rightarrow & S: & R & \text{with } R \in_{\mathsf{R}} \{0,1\}^{128} \\
S & \rightarrow & C: & (M_1, T_1) & \text{with } T_1 = \mathsf{trunc}_{32}(\mathsf{Mac}_K(M_1, R)) \\
S & \rightarrow & C: & (M_2, T_2) & \text{with } T_2 = \mathsf{trunc}_{32}(\mathsf{Mac}_K(M_2, T_1)) \\
 & \vdots & & & \\
S & \rightarrow & C: & (M_i, T_i) & \text{with } T_i = \mathsf{trunc}_{32}(\mathsf{Mac}_K(M_i, T_{i-1})) \\
 & \vdots & & &
\end{array}$$

The controller C picks a new 128-bit random value R when the system is powered up. Each message (M_i, T_i) is sent i seconds after that. The messages M_i are normally all identical, of the form $M = 0$ meaning "no burglary has happened in the last second". Mac is a 128-bit message-authentication code function, using a private key K shared between S and C. Because of the very limited data rate available on the alarm-wire interface, the output of Mac is truncated to the first 32 bits.

(*i*) How can an attacker, who has been observing this communication since power up, eventually predict future tags T_i for the constant message $M_i = M$? [4 marks]

(*ii*) How long will it take, on average, after powerup until the attacker can start sending simulated sensor messages? [4 marks]

(*iii*) What security implication does the predictability of message-authentication codes from a sensor have for a burglar alarm system? [2 marks]

(*iv*) How can you improve the protocol to practically eliminate the risk of that attack, without increasing the number of bits transmitted over the wire? [4 marks]

7 Denotational Semantics

(*a*) (*i*) Define the notion of admissible subset of a domain and state Scott's fixed point induction principle. [4 marks]

(*ii*) Let $(D, \sqsubseteq_D)$ and $(E, \sqsubseteq_E)$ be domains and let $f : D \to E$ and $g : E \to D$ be continuous functions.

Using Scott's fixed point induction principle prove

(A) $\mathit{fix}(f \circ g) \sqsubseteq_E f\big(\mathit{fix}(g \circ f)\big)$

(B) $f\big(\mathit{fix}(g \circ f)\big) \sqsubseteq_E \mathit{fix}(f \circ g)$

[8 marks]

(*b*) (*i*) Define the contextual-equivalence relation $P_1 \cong_{\mathrm{ctx}} P_2 : \tau$ for pairs of closed PCF expressions P_1, P_2 and a PCF type τ. [2 marks]

(*ii*) Prove or disprove the following statement.

For every pair of PCF types σ, τ and every pair of closed PCF expressions M of type $\sigma \to \tau$ and N of type $\tau \to \sigma$,

$$\mathbf{fix}\big(\mathbf{fn}\ y : \tau.\ M(N(y))\big) \cong_{\mathrm{ctx}} M\big(\mathbf{fix}\big(\mathbf{fn}\ x : \sigma.\ N(M(x))\big)\big)\ : \tau$$

[6 marks]

8 Hoare Logic and Model Checking

We consider the LTL temporal logic over atomic propositions $p \in \mathsf{AP}$:
$\phi \in \mathsf{PathProp} ::= p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \phi_1 \rightarrow \phi_2 \mid \mathsf{X}\ \phi \mid \mathsf{F}\ \phi \mid \mathsf{G}\ \phi \mid \phi_1\ \mathsf{U}\ \phi_2$.

(*a*) Precisely state the semantics of the until operator $\phi_1\ \mathsf{U}\ \phi_2$. [2 marks]

(*b*) Express $\mathsf{F}\ \phi$ in terms of the until operator $-\ \mathsf{U}\ =$. [2 marks]

(*c*) Give models $\mathcal{M}_1, \mathcal{M}_2$ such that $\mathcal{M}_1 \models \mathsf{G}\ (p \vee q)$ and $\mathcal{M}_2 \models (\mathsf{G}\ p) \vee (\mathsf{G}\ q)$, but either $\mathcal{M}_1 \not\models (\mathsf{G}\ p) \vee (\mathsf{G}\ q)$ or $\mathcal{M}_2 \not\models \mathsf{G}\ (p \vee q)$ (indicate which). [3 marks]

(*d*) Starting from any strictly positive integer n, the transition system induced by going to $\mathsf{n}/2$ if n is even, and to $3 \times n + 1$ if n is odd, is conjectured to always pass through 1.

(*i*) Precisely describe this conjecture in the form of a model and an LTL formula. [4 marks]

(*ii*) Describe what shape a counterexample to this conjecture would have. [2 marks]

(*e*) Alice (a) and Bob (b) share a bicycle. To ensure they do not have problems, they have a protocol: they can express an interest for it (e), use it (u), or not need it (n), yielding atomic propositions $AP = Pers \times Act$, where $Pers ::= \mathsf{a} \mid \mathsf{b}$ and $Act ::= \mathsf{e} \mid \mathsf{u} \mid \mathsf{n}$, so that for example a state labelled with $\{\mathsf{ae}, \mathsf{bu}\}$ is one where Alice has expressed interest in using the bike, and Bob has taken it.

Give LTL formulas for

(*i*) Alice does not keep the bike forever. [2 marks]

(*ii*) Non-starvation: if Alice expresses an interest in having the bike for long enough, she eventually gets it. [2 marks]

(*iii*) Alice cannot take the bike twice in a row if Bob expresses interest throughout. [3 marks]

9 Information Theory

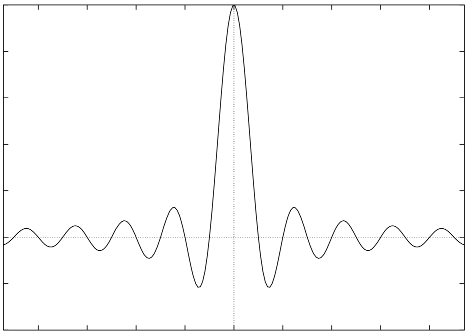
(*a*) A machine learning system was trained to learn about random variable X using a training set in which the probability distribution of its values $\{x_i\}$ was $p(x)$, having entropy $H(p)$; but then a test data set had a different distribution $q(x)$.

(*i*) Define the cross-entropy $H(p, q)$ between distributions $p(x)$ and $q(x)$. What condition minimises its value, and then what value is it? [3 marks]

(*ii*) Define the Kullback-Leibler distance (or the relative entropy) $D_{KL}(p\|q)$ between the distributions $p(x)$ and $q(x)$. What condition minimises its value, and then what value is it? [3 marks]

(*iii*) Show that: $H(p, q) = D_{KL}(p\|q) + H(p)$. [2 marks]

(*b*) The function $\text{sinc}(x) = \dfrac{\sin(\pi x)}{\pi x}$ for $x \neq 0$ as plotted below plays an important role in the Sampling Theorem. By considering its Fourier transform, show that this function is unchanged in form after convolution with itself, and show that it even remains unchanged in form after convolution with any higher frequency sinc function, $\text{sinc}(ax)$ for $a > 1$; but that if $0 < a < 1$, then the result of the convolution is instead that lower frequency sinc function $\text{sinc}(ax)$. [4 marks]



(*c*) Explain the use of run-length encoding in JPEG compression. How does it enable compession factors of typically at least 10:1 with no perceptible loss, and even compression factors of 30:1 or higher? What is the role of the quantisation table in achieving this, and what is the relevant fact about spatial frequency sensitivity in human visual perception? [3 marks]

(*d*) (*i*) Define the *genetic isopoint* of a human population. (*ii*) For most Europeans today, in what century did it occur? (*iii*) For a large well-mixed population of size m, approximately how many generations N ago was the genetic isopoint? (*iv*) Regarding genetic transmission as a lossy information channel, what sampling fact becomes critical for the effect of an ancestor once a family tree extends back at least $N = 15$ generations? (*v*) What does Information Theory imply is achieved by sexual (as opposed to asexual) reproduction? [5 marks]

10 Machine Learning and Bayesian Inference

The central limit theorem tells us that, if X_i are random variables, μ is the mean of X_i and σ^2 is the variance of X_i then, under suitable conditions

$$\frac{\hat{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

where $N(0,1)$ denotes the normal density with mean 0 and variance 1, and

$$\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(*a*) Let Y have density $N(0,1)$. We know that, for a parameter p, there is a constant z_p such that

$$\Pr(-z_p \leq Y \leq z_p) > p.$$

Show that with probability at least p, the quantity μ as defined above is in the interval described by $\hat{X}_n \pm z_p(\sigma/\sqrt{n})$. [3 marks]

(*b*) The quantity $\hat{X}_n$ can be regarded as an estimate of μ. If the random variables X_i take values in $\{0,1\}$, and are also independent and identically distributed, explain why it might make sense to estimate σ^2 as

$$\sigma^2 \simeq s = \hat{X}_n(1 - \hat{X}_n).$$

[3 marks]

(*c*) Define what it means for an estimate such as that suggested in Part (*b*) to be *unbiased*. Is the estimate suggested in Part (*b*) unbiased? Provide a proof of your answer. [7 marks]

(*d*) We have two binary classifiers h_1 and h_2, and a test set $\mathbf{s}$ containing 1000 examples. During testing, h_1 makes 105 errors and h_2 makes 120 errors. Explain how we can estimate a confidence interval of the kind defined in Part (*a*) for the difference $(\text{er}(h_1) - \text{er}(h_2))$ between the true error probabilities $\text{er}(h_1)$ and $\text{er}(h_2)$ of the classifiers. Your answer should be careful to state any assumptions or approximations being made. [7 marks]

11 Mobile and Sensor Systems

A company manufactures small location tags containing a speaker and a Bluetooth radio designed to be attached to personal effects such as keys. A Bluetooth-enabled smartphone can instruct the device to make an audible noise to aid in finding it when lost. The company wishes to support *background* location estimation of the tags relative to the phone.

(*a*) They propose to use Bluetooth 5.1, which supports an antenna array for direction finding from Bluetooth packet exchanges.

(*i*) Explain how an antenna array enables direction finding. Discuss any practical challenges in direction finding indoors. [4 marks]

(*ii*) State and justify the upper bound on the spacing between antenna array elements and estimate this value in metres for Bluetooth. Comment on the practicality of your estimate. Is there any disadvantage to having closer spacing? [4 marks]

(*iii*) The antenna array could be on the phone or on the location tag or both. Discuss the commercial trade-offs of these options. [3 marks]

(*b*) In addition to direction, the system must estimate the distance to the tag from the phone.

(*i*) Discuss how feasible it is to use Received Signal Strength (RSS) measurements to estimate distance in this context. [1 mark]

(*ii*) Explain how WiFi 802.11mc Fine Timing Meansurement estimates distance. Discuss whether an analogous scheme could be implemented using standard Bluetooth 5.1 hardware for this application. [5 marks]

(*iii*) Given that the speakers and microphones on smartphones and on the tags can produce and receive ultrasonic audio, describe an approach to estimating distance. How would you optimise your system for the power-sensitive location tags? [3 marks]

12 Optimising Compilers

The following code for a function is given to a compiler for optimisation, where arguments are passed in through variables arg0, arg1 and arg2 and the result is returned through variable res0. Describe five optimisations that the compiler can carry out, explaining for each the analysis required, the actual transformation, how it impacts the code and showing the resulting code after all optimisations have been performed.

```
entry foo
      mov t0, #0x1000      // t0 = 0x1000
      mov t1, #0           // t1 = 0
      mov t2, #0           // t2 = 0
      mul t3, arg0, #2     // t3 = arg0 * 2
      mul t4, arg1, #4     // t4 = arg1 * 4
      mov t5, #&lab4       // t5 = address of lab4
      sti t0, t5           // *t0 = t5
      b lab7               // branch to lab7
lab1: mov t6, #1           // t6 = 1
      cmpeq t1, arg2, lab3 // branch to lab3 if t1 == arg2
      mul t7, t4, t4       // t7 = t4 * t4
      cmpeq t1, t7, lab5   // branch to lab5 if t1 == t7
      b lab6               // branch to lab6
lab2: mov t8, #&lab5       // t8 = address of lab5
      sti t0, t8           // *t0 = t8
      b lab5               // branch to lab5
lab3: mul t9, t4, t4       // t9 = t4 * t4
      add t2, t2, t9       // t2 = t2 + t9
      add t6, t6, #1       // t6 = t6 + 1
      ldi t10, t0          // t10 = *t0
      bi t10               // branch to address in t10
lab4: add t2, t2, t1       // t2 = t2 + t1
      b lab6               // branch to lab6
lab5: add t2, t2, t6       // t2 = t2 + t6
lab6: add t1, t1, #1       // t1 = t1 + 1
lab7: cmpne t1, t3, lab1   // branch to lab1 if t1 != t3
      mul res0, t2, #4     // res0 = t2 * 4
      mov t11, #&lab1      // t11 = address of lab1
      sti t0, t11          // *t0 = t11
      exit
```

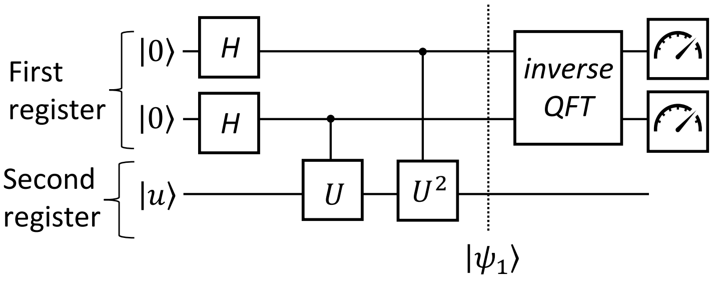
[4 marks each]

13 Principles of Communications

(*a*) In general, weighted fair queueing can be used to provide resource guarantees to packet flows in networks. Explain the problem with using a simple priority queue to do the same. Define weighted fair queueing. [5 marks]

(*b*) The Data Centre environment often allows us to make some simplifying assumptions about traffic that mean we can employ a priority queueing system in switches, combined with some traffic regulation in end systems. *QJump* is one approach to this - describe the assumptions that allow this, and the mechanisms QJump uses to offer delay bounding for high priority traffic without resource starvation for bulk traffic flows. [15 marks]

14 Quantum Computing



(*a*) Show that $U = \frac{1}{\sqrt{5}}\begin{bmatrix}1 & 2\\ 2 & -1\end{bmatrix}$ is unitary. [1 mark]

(*b*) Show that $\begin{bmatrix}1+\sqrt{5}\\ 2\end{bmatrix}$ and $\begin{bmatrix}1-\sqrt{5}\\ 2\end{bmatrix}$ are (un-normalised) eigenvectors of U, and give the eigenvalues. [2 marks]

(*c*) The figure shows the quantum circuit for quantum phase estimation to two bits of precision for a single-qubit unitary. Quantum phase estimation is performed to two bits of precision with $U = \frac{1}{\sqrt{5}}\begin{bmatrix}1 & 2\\ 2 & -1\end{bmatrix}$ and $|u\rangle = \begin{bmatrix}1\\ 0\end{bmatrix}$, find the quantum state, $|\psi_1\rangle$, before the inverse quantum Fourier transform is executed. [7 marks]

(*d*) What possible measurement outcomes can occur (i.e., after the inverse quantum Fourier transform, with measurement in the computational basis)? Give probabilities for each possible outcome. [6 marks]

(*e*) Give two applications of quantum phase estimation, and for each give the state, $|u\rangle$, in which the second register should be prepared, and briefly outline how these can be prepared in practise. [4 marks]

15 Types

(*a*) Consider System F extended with existential types, products, and a natural number type.

(*i*) Give an existential type corresponding to an abstract type of booleans with constructors for true and false, as well as a conditional test (if-then-else) operation. [3 marks]

(*ii*) Give an implementation of this type, using the natural numbers as the representation of booleans. [4 marks]

(*b*) Suppose we extend the simply-typed lambda calculus with the ability to raise exceptions with the fail construct, and the ability to catch exceptions with the $\mathsf{try}\, e_0\, \mathsf{except}\, e_1$ construct. Suppose also that we track exceptions monadically, with the type $\mathsf{Exn}\, A$ representing possibly-failing computations of A.

(*i*) Give a typing rule for signalling an error with fail. [2 marks]

(*ii*) Give a typing rule for trapping an error with $\mathsf{try}\, e_0\, \mathsf{except}\, e_1$. Does your type for this term have an effect? Justify your design. [5 marks]

(*c*) Consider the simply-typed lambda calculus extended with natural numbers and reference types, but without monadic effect tracking.

(*i*) Give an expression of type $1 \to \mathbb{N}$, which evaluates to a function which counts the number of times it has been called. [2 marks]

(*ii*) Without using explicit recursion, give an expression and its type in the simply-typed lambda calculus with references whose execution never halts. [4 marks]

END OF PAPER