COMPUTER SCIENCE TRIPOS Part II

Tuesday 5 June 2018 1.30 to 4.30

COMPUTER SCIENCE Paper 7

*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
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

1 Advanced Algorithms

(a) What are the three possible cases for the solution of a linear program? For each of them, give an example of a linear program in standard form exhibiting this case. [6 marks]

(b) What is the set of optimal solutions for the following linear program?

$$\begin{aligned} \text{Minimize} \quad & -x_1 - x_2 \\ & -x_2 \geq -3 \\ & 2x_1 + x_2 \leq 8 \\ & x_1, x_2 \geq 0 \end{aligned}$$

[6 marks]

(c) For a given linear program $\mathbf{LP}_1$

$$\begin{aligned} \text{Maximize} \quad & \sum_{j=1}^{n} c_j x_j \\ & \sum_{j=1}^{n} a_{ij} x_j \leq b_i \qquad (1 \leq i \leq m) \\ & x_j \geq 0 \qquad (1 \leq j \leq n), \end{aligned}$$

consider a new linear program $\mathbf{LP}_2$:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{m} b_i y_i \\ & \sum_{i=1}^{m} a_{ij} y_i \geq c_j \qquad (1 \leq j \leq n) \\ & y_i \geq 0 \qquad (1 \leq i \leq m). \end{aligned}$$

(i) Prove that if x is a feasible solution for $\mathbf{LP}_1$ and y is a feasible solution for $\mathbf{LP}_2$, then $c^T x \leq b^T y$. [6 marks]

(ii) Using your answer in Part (c)(i), what can we conclude about $\mathbf{LP}_2$ if we know that $\mathbf{LP}_1$ is unbounded? [2 marks]

2 Advanced Graphics

(*a*) Given two signed distance field functions f and g, give the formula for their ...

(*i*) Union ($f \cup g$)

(*ii*) Intersection ($f \cap g$)

(*iii*) Difference ($f - g$)

[3 marks]

(*b*) Give clear definitions for the Virtual Reality industry's principles of *immersion* and *presence*. Compare the two concepts and explain the difference between them with examples demonstrating each. [5 marks]

(*c*) The *Doo-Sabin* subdivision scheme has kernel $(1/4)[\ldots, 0, 0, 1, 3, 3, 1, 0, 0, \ldots]$, defining a scheme in which each face is replaced by four new vertices.

(*i*) Give an expression for computing the position of a new vertex given the positions of the four old vertices of a face. [2 marks]

(*ii*) If the face does not have 4 vertices then you must weight each parent vertex differently to find the position of the child. Suggest possible weights for the vertices of faces with 3, 5, and n vertices, and justify your answer. [3 marks]

(*d*) There are several ray-tracing-friendly acceleration structures.

(*i*) Explain the *BSP tree* data structure. Explain how it is constructed and traversed. [3 marks]

(*ii*) Explain the *kd-tree* data structure. Explain how it is constructed and traversed. [3 marks]

(*iii*) Which of the two data structures is best-suited to ray-tracing a game of chess in real time? [1 mark]

3 Bioinformatics

(*a*) Explain with one example the difference between local and global alignment. [3 marks]

(*b*) Give one example why the multiple alignment, as implemented in the software Clustal, described in the course, needs a guide tree. [4 marks]

(*c*) What is the scope of phylogeny? [2 marks]

(*d*) Describe the UPGMA algorithm. [4 marks]

(*e*) Explain with one example what the ultrametric property of a phylogenetic tree tells us about the evolutionary process. [3 marks]

(*f*) Explain the steps and the complexity of the divide and conquer approach to sequence analysis. [4 marks]

4 Business Studies

(*a*) Describe the characterists of debt and equity financing, highlighting the differences between them. [4 marks]

(*b*) What is the difference between a loan and an overdraft? [1 mark]

Your software company is contracted to create a new control system for chocolate bar delivery in Cambridge. The contract is for a 6 month period, with payment of £400k against milestones in months 1, 3 and 6.

(*c*) Create an outline cashflow for the project assuming staff costs of £75k per month and overheads of £60k per month. [5 marks]

(*d*) What is your working capital requirement for the project allowing a contingency of a two month delay to one of either the second or third delivery milestones? [5 marks]

(*e*) How would you suggest to finance this working capital requirement, justifying your answer? [5 marks]

5 Comparative Architectures

(*a*) Briefly describe three microarchitectural techniques or elements that can be used to improve the performance of a scalar pipelined processor. You are unable to fetch more than a single instruction per clock cycle or make any changes to the Instruction Set Architecture (ISA). [4 marks]

(*b*) Imagine two processor implementations with equal performance. One is a superscalar design with support for out-of-order execution. The other is an in-order scalar processor. In what circumstances might the superscalar design be more power efficient? [6 marks]

(*c*) Loads and stores are often reordered in a superscalar processor. Describe how some loads can be issued speculatively before the addresses of older stores are known and how mispredictions are detected and handled. [6 marks]

(*d*) How can memory reference speculation be supported in a VLIW processor? [4 marks]

6 Denotational Semantics

Consider the concocted language PCF_* obtained from the language PCF by extending it with:

- Types
  $\tau ::= \cdots \mid \tau * \tau$
- Expressions
  $M ::= \cdots \mid \mathbf{pair}(M, M) \mid \mathbf{left}(M) \mid \mathbf{right}(M)$
- Typing rules

$$\frac{\Gamma \vdash M_1 : \tau_1 \quad \Gamma \vdash M_2 : \tau_2}{\Gamma \vdash \mathbf{pair}(M_1, M_2) : \tau_1 * \tau_2} \qquad \frac{\Gamma \vdash M : \tau_1 * \tau_2}{\Gamma \vdash \mathbf{left}(M) : \tau_1} \qquad \frac{\Gamma \vdash M : \tau_1 * \tau_2}{\Gamma \vdash \mathbf{right}(M) : \tau_2}$$

- Values
  $V ::= \cdots \mid \mathbf{pair}(V, V)$
- Operational semantics

$$\frac{M_1 \Downarrow_{\tau_1} V_1 \qquad M_2 \Downarrow_{\tau_2} V_2}{\mathbf{pair}(M_1, M_2) \Downarrow_{\tau_1 * \tau_2} \mathbf{pair}(V_1, V_2)}$$

$$\frac{M \Downarrow_{\tau_1 * \tau_2} \mathbf{pair}(V_1, V_2)}{\mathbf{left}(M) \Downarrow_{\tau_1} V_1} \qquad \frac{M \Downarrow_{\tau_1 * \tau_2} \mathbf{pair}(V_1, V_2)}{\mathbf{right}(M) \Downarrow_{\tau_2} V_2}$$

(*a*) Give a denotational semantics for the above extension of PCF. [3 marks]

(*b*) Show that the denotation of types are domains and that the denotation of terms are continuous functions. You may use any standard results provided that you state them clearly. [5 marks]

(*c*) State the soundness property for a denotational semantics of PCF_*. [2 marks]

(*d*) Show that your denotational semantics of PCF_* is sound. You may use any standard results provided that you state them clearly. [4 marks]

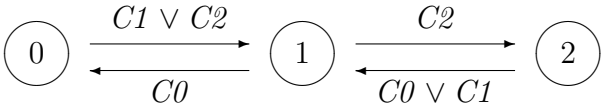
(*e*) State the adequacy property for a denotational semantics of PCF_*. [2 marks]

(*f*) Establish whether or not your denotational semantics of PCF_* is adequate. You may use any standard results provided that you state them clearly. [4 marks]

7 Hoare Logic and Model Checking

(a) Give a formal definition of a Kripke structure, as a 3- or 4-tuple, briefly explaining the roles of its components. What might a Kripke structure model? [4 marks]

(b) A lift controller manages a lift moving between floors 0, 1 and 2. There are three 'call' input buttons (*C0*, *C1*, *C2*) within the lift requesting the lift to move to the corresponding floor. These are duplicated at each floor to avoid the need for a separate 'call' button. They are internally latched as usual—a call button-press stays active until reaching the associated floor, but can be immediately reactivated (e.g. useful when one realises the lift is setting off in the wrong direction!). The controller is (rather informally) specified by a hardware-style state transition diagram with three inputs and three states as in the diagram:



Give a Kripke-structure model for the controller, explaining any necessary changes or clarifications you make. You need not model the internal structure of the call buttons, it suffices to treat them as (a) non-deterministically becoming active and (b) deactivated on arrival at the corresponding floor. [*Hint:* Two possible answers have 12 and 24 states in the Kripke structure.] [6 marks]

(c) Give formulae (in a temporal logic of your choice, but which you should name) corresponding to

(i) If I press button *C0* the lift will eventually arrive at floor 0

(ii) If I press button *C1* the lift will eventually arrive at floor 1

(iii) If I press button *C2* the lift will eventually arrive at floor 2

[*Hint:* You might wish to check your Kripke model above defines any predicates you use in your answers.] [4 marks]

(d) Which of your formulae in Part (c) are valid in your Kripke model? [2 marks]

(e) Improve the state transition diagram in Part (b) to fix any problems you discover in Part (d). It is not necessary to give a Kripke model. [4 marks]

8 Human–Computer Interaction

This question relates to the design of control software for home automation. There is a market opportunity arising from increasing deployment of software-controlled light bulbs and fittings. As householders acquire larger numbers of such products from different suppliers, they will wish to create lighting schemes, coordinating lights located in particular areas of the house or in individual rooms, where contrasting schemes might be designed for various times of day or activities.

You are asked to consider ways in which HCI research can be conducted to compare the relative advantages of two technical approaches to this opportunity. The first approach is to use machine learning algorithms to infer the schemes from user actions. The second approach is to provide a programming language with which users can define the schemes.

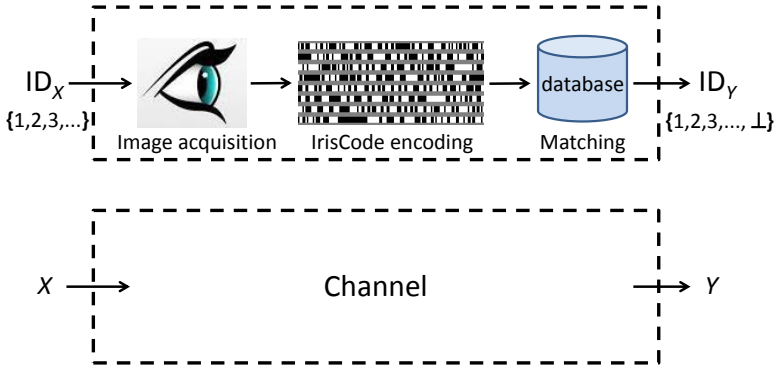
(*a*) For *each* of the two technical approaches, describe a usability problem that you might expect to arise that is specific to the lighting scheme application, and a design strategy that could be taken to reduce the impact of this problem, noting any trade-offs that may result. [8 marks]

(*b*) Propose a formative empirical research method that could be used to compare the relative desirability of the two technical approaches, from the perspective of the home lives of potential customers. Your proposal should include a description of the practical steps involved in carrying out the research, the data that will be collected, the way that it will be analysed, and the kind of recommendations that you expect to be able to make. [6 marks]

(*c*) Your company has deployed alpha-release versions of both systems. Early feedback from users suggests that there is a problem when a new lighting scheme must be defined quickly, for example when guests arrive for a party. Propose a summative empirical research method that could be used to compare the speed of the two technical approaches, in order to find which is faster. Your proposal should include a description of the practical steps involved in carrying out the research, the data that will be collected, the way that it will be analysed, and the kind of recommendations that you expect to be able to make. [6 marks]

9 Information Theory

A classifier or pattern recognition system such as illustrated in the diagram below can be treated abstractly as a channel, with some object as input represented by discrete random variable $X = \{x_1, \ldots, x_J\}$, and the decision output represented by discrete random variable $Y = \{y_1, \ldots, y_K\}$. Note that J and K need not be the same. Errors may be made, and the output $\perp$ may signify that no decision was possible.



(*a*) Define a channel matrix and provide it for such a system. [4 marks]

(*b*) Assuming there is a known probability distribution $p(x_j)$ over all possible inputs x_j, and that correct decisions about them are y_k with $k = j$, give an expression for the average probability of error, P_e. [4 marks]

(*c*) Give an expression for the mutual information $I(X;Y)$ of this system, in terms of the entropy $H(X)$ of the set of possible inputs and the conditional entropy $H(X|Y)$ of the inputs given the output decisions Y. [2 marks]

(*d*) Now give an expression for the mutual information $I(Y;X)$ of this system in terms of the entropy $H(Y)$ of the output decisions and the conditional entropy $H(Y|X)$ of the decisions given the inputs X. [2 marks]

(*e*) Suppose now that there are $J = 2^N$ (for integer N) possible inputs and that they are all equiprobable: $\forall j,\ p(x_j) = J^{-1} = 2^{-N}$. In terms of N, what is the entropy $H(X)$ of this set of possible inputs? [3 marks]

(*f*) In India, $J \approx 1$ billion $\approx 2^{30}$ citizens have been enrolled in the identification system illustrated in the diagram. All citizens present themselves equiprobably. Suppose that the average remaining uncertainty about input persons X given inferred identities Y is $H(X|Y) = 1$ bit: the system computes that the IrisCode may equally well arise from either one of two persons among the billion. What then is the mutual information of the system, in bits per identity? [3 marks]

(*g*) What then is the channel capacity C of this system, in bits per identity, given the assumption of equiprobable presenting identities? [2 marks]

10 Machine Learning and Bayesian Inference

A *linear maximum-margin classifier* computes a function

$$f_{\mathbf{w},w_0}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

and assigns a class as $\text{sgn}(f_{\mathbf{w},w_0}(\mathbf{x}))$ where $\text{sgn}(x) = 1$ if $x \geq 0$ and $\text{sgn}(x) = -1$ otherwise. It is trained using a training sequence $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ and the aim in training is to solve the problem

$$\text{argmax}_{\mathbf{w},w_0} \left(\min_i \frac{y_i f_{\mathbf{w},w_0}(\mathbf{x}_i)}{||\mathbf{w}||} \right). \tag{1}$$

(*a*) Give a brief explanation of how Equation 1 is derived. You may assume that the distance from $\mathbf{x}'$ to the line $f_{\mathbf{w},w_0}(\mathbf{x}) = 0$ is $|f_{\mathbf{w},w_0}(\mathbf{x}')|/||\mathbf{w}||$. [2 marks]

(*b*) Why is Equation 1 not used in practice? Explain how an alternative optimization problem is derived that can form the basis of a practical learning algorithm. You need only derive a statement of the primal optimization problem. [3 marks]

(*c*) Explain how the linear maximum-margin classifier can be modified to be nonlinear and to allow misclassification of the training examples. Give a derivation of the modified optimization problem needed for training. You need only derive a statement of the primal optimization problem. [4 marks]

(*d*) As part of the derivation of the full learning algorithm we find that the function f might be expressible in terms of m new parameters α_i as

$$f_{\alpha_1,\ldots,\alpha_m,w_0}(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + w_0. \tag{2}$$

Explain the purpose of K in Equation 2 and explain why its use might be beneficial. [3 marks]

(*e*) Your boss can not afford to provide you with a solver capable of training your system using the algorithm in Parts (*c*) and (*d*). Your boss does however provide you with a solver for *linear programs*. For a matrix $\mathbf{A}$ and vectors $\mathbf{b}$ and $\mathbf{c}$, this solves problems of the form

$$\text{Find } \mathbf{x} \text{ minimizing } \mathbf{b}^T\mathbf{x} \text{ with constraints } \mathbf{A}\mathbf{x} \geq \mathbf{c} \text{ and } \mathbf{x} \geq \mathbf{0}.$$

Suggest a way in which you could use this optimizer to (approximately) train your system. [8 marks]

11 Natural Language Processing

(*a*) The following text is from a children's story by Beatrix Potter (with slight modifications):

> Mr Jeremy put on a macintosh, and a pair of shiny shoes; he took his fishing rod and basket, and set off with enormous hops to the place where he kept his boat. The boat was round and green, and very like the other lily-leaves. It was tied to a water-plant in the middle of the pond.

Describe six features that are used in pronoun resolution algorithms using classifiers. For each feature, explain the range of values it can take, giving illustrative examples from the text above, and list any resources or systems necessary to derive the feature. [8 marks]

(*b*) Describe the data you would need to train and test a classifier for pronoun resolution that made use of these features. Illustrate your answer with examples from the text in (*a*). [3 marks]

(*c*) Given a trained classifier, how could you set up an experiment to evaluate its performance on some new text? [3 marks]

(*d*) On conducting such an experiment, you obtain worse performance than the results previously reported for a standard dataset. How would you investigate possible causes of the difference? [6 marks]

12 Optimising Compilers

(*a*) Describe a procedure's flowgraph and the concept of basic blocks. What language construct makes flowgraph construction imprecise? [3 marks]

(*b*) Describe the principle of dominance within a flowgraph, including strict dominance, dominance trees and the dominance frontier. [4 marks]

(*c*) Define data-flow equations to find each node's dominators and an algorithm to compute this. [5 marks]

(*d*) Describe how to use dominance information to find loops in a flowgraph. [2 marks]

(*e*) What is static single assignment (SSA) form? [1 mark]

(*f*) Use the dominance frontier to place ϕ nodes into the following code and thereby convert it to SSA form.

```
b1:   LDR  x, #0xFF00
      LDR  y, #0xFF08
      BEQ  x, #0, b9
b2:   BEQ  x, #1, b4
b3:   ADD  x, x, #2
      BR   b5
b4:   ADD  x, x, y
b5:   BEQ  x, #8, b7
b6:   STR  x, #0xFFA0
      BR   b8
b7:   STR  y, #0xFFA0
b8:   ADD  x, x, y
      BR   b10
b9:   ADD  y, y, #1
b10:  ADD  x, x, #1
      STR  x, #0xFF00
```

[5 marks]

13 Principles of Communications

(*a*) Routing across the Internet is held together by the Border Gateway Protocol (BGP). Describe how BGP differs from the Interior Gateway Protocols (for example, a link-state scheme), and how this leads to a number of scalability challenges. [10 marks]

(*b*) Multicast routing was introduced to meet the demand for live-streamed content to multiple simultaneous receivers. It has seen limited deployment due to a number of operational challenges, including how it might interact with inter-domain routing, accounting, and security. Describe these challenges. [10 marks]

14 Security II

(*a*) Compare and contrast the usability and security of messaging using encrypted email (either GPG or PGP) and end-to-end encrypted messaging (WhatsApp, Signal, or iMessage). [5 marks]

(*b*) Let's Encrypt provides free TLS certificates and has a secure protocol for issuing them automatically. This means that software can automatically provision itself with a valid TLS certificate when it is installed without manual intervention. How does this change the security ecosystem for TLS provision? Justify your answer with reference to security economics, network security, and usability. [5 marks]

(*c*) You are responsible for designing, implementing and maintaining a user-authentication service for a new e-commerce website. The website will include a marketplace for third party vendors, multiple methods of processing payments and is optimised for mobile and desktop computers. Describe and justify your design in detail. [10 marks]

END OF PAPER