COMPUTER SCIENCE TRIPOS Part II

Tuesday 6 June 2017 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) Into what three cases can a linear program in standard form be classified? [3 marks]

(b) Consider the (unweighted) vertex cover problem for the graph $G = (V, E)$ with $V = \{1, 2, 3\}$ and $E = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$.

(i) Write down the linear program relaxation for the vertex cover problem and solve the linear program. [6 marks]

(ii) Based on the solution of the linear program in (b)(i), derive an integer solution using the rounding approach described in the lecture. [2 marks]

(c) Consider the following randomised algorithm for the unweighted vertex cover problem:

```
Initialize S to be the empty set
For all edges e=(u,v) do
   If neither u nor v belongs to S
      Randomly choose u or v with probability 1/2
      and add the vertex to S
   End If
End For
Return S
```

Derive an upper bound, as tight as possible, on the approximation ratio of the algorithm.

Hint: Try to find an invariant that bounds from below the size of the intersection of the current solution $S = S(i)$ with the optimum solution, where $S(i)$ denotes the set S after the i-th iteration of the FOR loop. [9 marks]

2 Advanced Graphics

(*a*) Briefly explain the global illumination methods *radiosity* and *photon mapping*. Highlighting the strengths and weaknesses of each method, compare and contrast the two.
You will be marked for correctness, clarity and brevity. [8 marks]

Recall that the *signed distance field* (SDF) expression of a surface returns the signed nearest distance from a sample point to the surface. This is well-suited to ray-marching on a GPU. As an example, the SDF method describing a unit cube centred at the origin may be written in *openGL shading language* (GLSL) as:

```
float cube(vec3 pt) {
  return max(abs(pt.x), max(abs(pt.y), abs(pt.z))) - 1;
}
```

(*b*) Give an SDF method **cylY(pt, len, radius)** for a finite cylinder of specified length and radius, centred at the origin, parallel to the Y axis. [4 marks]

(*c*) Give an SDF method **hollowedSphere(pt)** which specifies the model shown in Figure 1: a unit sphere hollowed along each axis by a cylindrical hole of radius 0.5. [6 marks]

(*d*) How would you repeat your hollowed sphere at two-unit intervals infinitely across the XZ plane as illustrated in Figure 2? [2 marks]

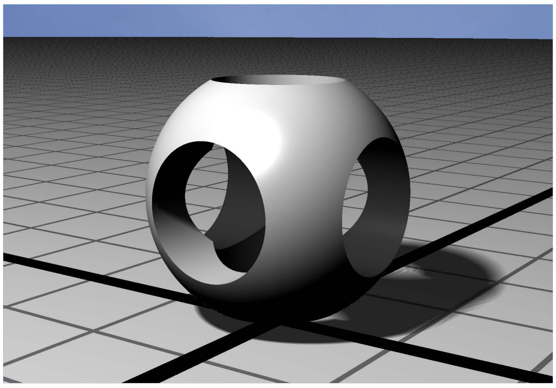


Figure 1: A unit sphere hollowed along each axis by a hole of radius 0.5

Ground plane is for illustration only

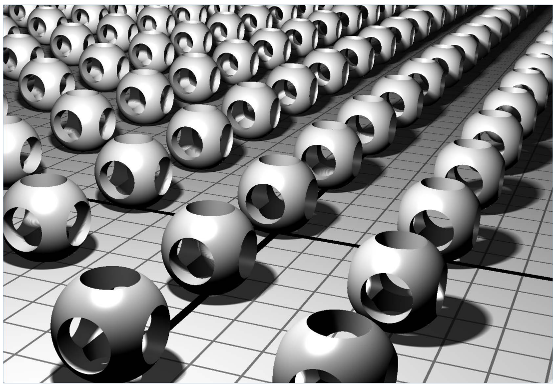


Figure 2: The hollowed unit sphere infinitely repeated in XZ at intervals of 2

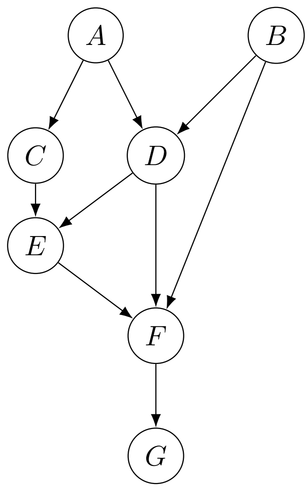
3 Machine Learning and Bayesian Inference

(*a*) For random variables (RVs) A_1, A_2 and B, define what it means for A_1 to be *conditionally independent of* A_2 *given* B, written $A_1 \perp A_2|B$. [1 mark]

(*b*) Given mutually disjoint sets X_1, X_2 and Y of random variables from some Bayesian network, define what it means to say that a path from $x_1 \in X_1$ to $x_2 \in X_2$ is *blocked* by Y. [5 marks]

(*c*) Given mutually disjoint sets X_1, X_2 and Y of random variables from some Bayesian network, define what it means for X_1 and X_2 to be *d-separated* by Y. What does this tell you about the probability distribution represented by the Bayesian network? [3 marks]

(*d*) Consider the following Bayesian network.



In each of the following cases, establish whether or not $X_1 \perp X_2|Y$:

(*i*) $X_1 = \{A, E\}$, $X_2 = \{G\}$ and $Y = \{F\}$. [2 marks]

(*ii*) $X_1 = \{A\}$, $X_2 = \{D\}$ and $Y = \{G\}$. [3 marks]

In each case explain your answer.

(*e*) Define how Bayesian networks and Markov random fields are used to represent probability distributions, and briefly describe the trade-offs involved in choosing one versus the other. [6 marks]

4 Bioinformatics

(*a*) Discuss the time and space complexity of dynamic programming algorithms in the context of sequence alignment problems. [6 marks]

(*b*) Describe the differences between Soft and Hard Clustering. [6 marks]

(*c*) Construct the de Bruijn graph of `TACCTTCAGCGCCTTC` by splitting it into k-mers for a suitable value of k. Comment on how the choice of k affects the construction. Discuss the use of de Bruijn graphs in the context of genome sequencing. [8 marks]

5 Business Studies

(*a*) Describe five types of Intellectual Property. [5 marks]

(*b*) Describe five topics that might be in a Market Requirements Document. [5 marks]

(*c*) Giles Murchiston, a PhD student, has invented a new algorithm to control a dog walking robot. He realises that his algorithm may have wider applications including driverless cars, factory and farm robots, drones, and uses such as domestic robot lawnmowers or robot vacuum cleaners. Advise Giles on the exploitation of his idea. [10 marks]

6 Comparative Architectures

(*a*) What features of a processor's instruction set are desirable if a pipelined implementation is planned? [5 marks]

(*b*) The performance of a processor typically improves when a modest number of pipeline stages are created. Why does it become difficult to maintain near linear performance gains with deeper pipelines? [5 marks]

(*c*) Clustered superscalar processors partition functional units into clusters. Data forwarding within a cluster operates as normal allowing dependent instructions to execute on consecutive clock cycles. Communication between clusters normally incurs an additional delay of 1 or 2 clock cycles. The clustering idea may also be extended to include the issue buffer (also known as the issue window).

(*i*) What problem does clustering attempt to solve? [5 marks]

(*ii*) Assume a processor has two symmetric clusters that contain both functional units and an issue buffer. In this processor, instructions must be steered to a particular cluster before they are inserted in an issue buffer. What should the two basic goals of a good steering policy be? [5 marks]

7 Denotational Semantics

(a) (i) Define the notion of least pre-fixed point $\mathsf{fix}(f)$ of a continuous endofunction f on a domain and state Tarski's fixed point theorem for it. [2 marks]

(ii) State Scott's fixed point induction principle. [2 marks]

(b) Let $f : D \to D$, $g : E \to E$, and $h : D \to E$ be continuous functions between domains such that $h \circ f = g \circ h : D \to E$. Show that if h is strict (that is, $h(\bot_D) = \bot_E$) then $\mathsf{fix}(g) = h\big(\mathsf{fix}(f)\big)$. [4 marks]

(c) Let Σ^* denote the set of strings over an alphabet Σ, and let $\mathcal{P}(\Sigma^*)$ be the domain of all subsets of Σ^* ordered by inclusion.

Consider the continuous functions

$$\begin{array}{lcl} \text{(union)} & + & : \ \mathcal{P}(\Sigma^*) \times \mathcal{P}(\Sigma^*) \to \mathcal{P}(\Sigma^*) \\ \text{(concatenation)} & \cdot & : \ \mathcal{P}(\Sigma^*) \times \mathcal{P}(\Sigma^*) \to \mathcal{P}(\Sigma^*) \end{array}$$

given, for all $X, Y \subseteq \Sigma^*$, by

$$\begin{array}{rcl} X + Y & = & \{\, w \in \Sigma^* \mid w \in X \text{ or } w \in Y \,\} \\ X \cdot Y & = & \{\, uv \in \Sigma^* \mid u \in X \text{ and } v \in Y \,\} \end{array}$$

(i) Using part (b), or otherwise, show that, for all $A, B, C \subseteq \Sigma^*$,

(1) $\mathsf{fix}\big(\lambda X.\, C \cdot B + A \cdot X\big) \ = \ \mathsf{fix}\big(\lambda X.\, C + A \cdot X\big) \cdot B$ [4 marks]

(2) $\mathsf{fix}\big(\lambda X.\, A \cdot C + A \cdot B \cdot X\big) \ = \ A \cdot \mathsf{fix}\big(\lambda X.\, C + B \cdot A \cdot X\big)$ [4 marks]

(ii) For $P \subseteq \Sigma^*$, let $P^\star \subseteq \Sigma^*$ be defined as

$$P^\star \ = \ \mathsf{fix}\big(\lambda X.\, E + P \cdot X\big)$$

where E denotes the singleton set $\{\varepsilon\}$ consisting of the empty string ε.

Using part (c)(i), or otherwise, show that, for all $S, T \subseteq \Sigma^*$,

$$(S \cdot T)^\star \cdot S \ = \ S \cdot (T \cdot S)^\star$$ [4 marks]

8 Hoare Logic and Model Checking

Consider a programming language that consists of commands `C` composed from assignments `V := E` (where `E` is an expression) using sequences `C1;C2`, conditionals `IF S THEN C1 ELSE C2` (where `S` is a statement) and while-loops `WHILE S DO C`.

(*a*) Carefully explain the meaning of total correctness Hoare triples. [2 marks]

(*b*) Suggest a command C such that the following partial correctness triple holds.

$$\{X = x\}\ C\ \{1 = 2\}$$

Explain why the triple holds. [4 marks]

(*c*) Consider Hoare triples of the form $\{P\}$ `X := E` $\{P[\mathtt{E}/\mathtt{X}]\}$ where P, `X` and `E` range over formulas, variables and expressions, respectively. Recall that $P[\mathtt{E}/\mathtt{X}]$ denotes P with `E` substituted for every occurrence of `X` in P.

Write down an instance of such a triple that cannot be proved using Hoare logic and explain why it cannot be proved. [4 marks]

(*d*) Write down a partial correctness specification for a command that adds the initial values stored in variables `X` and `Y`. The command should store the result in a variable `Z`. [4 marks]

(*e*) Propose a loop invariant for proving the following partial correctness triple.

$$\{\mathtt{X} = n \land \mathtt{Y} = 0 \land n \geq 0\}$$

```
WHILE X > 0 DO (Y := Y + X; X := X - 1)
```

$$\{\mathtt{Y} = \sum_{i=1}^{n} i\}$$

[6 marks]

9 Human–Computer Interaction

This question relates to the design of interactive augmented reality applications. Consider a future application in which a (static) public billboard is augmented so that people passing in the street can see and modify an enhanced view, for example adding and editing caption text or speech bubbles over their personalised view of the billboard. This augmented functionality will be available either using hand-held mobile devices such as a touch-screen phone, or head-mounted displays such as Microsoft Hololens. Whichever type of device is used, users should be able to control the application either using hand gestures or gaze control.

(*a*) Discuss the requirements for sensor configuration and computer vision processing, to achieve the necessary detection and registration of the augmented interaction in *each of the four* possible combinations of the display and control methods described above. [8 marks]

(*b*) This part relates only to the head-mounted display version of the proposed product.

(*i*) Describe a technical approach that applies Bayes' theorem to improve the performance of command gesture recognition. [6 marks]

(*ii*) Describe a technical approach that applies Bayes' theorem to improve the performance of gaze-controlled text entry. [6 marks]

10 Information Theory

(*a*) An error-correcting (7/4) Hamming code combines four data bits b_3, b_5, b_6, b_7 with three error-correcting bits: $b_1 = b_3 \oplus b_5 \oplus b_7$, $b_2 = b_3 \oplus b_6 \oplus b_7$, and $b_4 = b_5 \oplus b_6 \oplus b_7$. The 7-bit block is then sent through a noisy channel, which corrupts one of the seven bits. The following incorrect bit pattern is received:

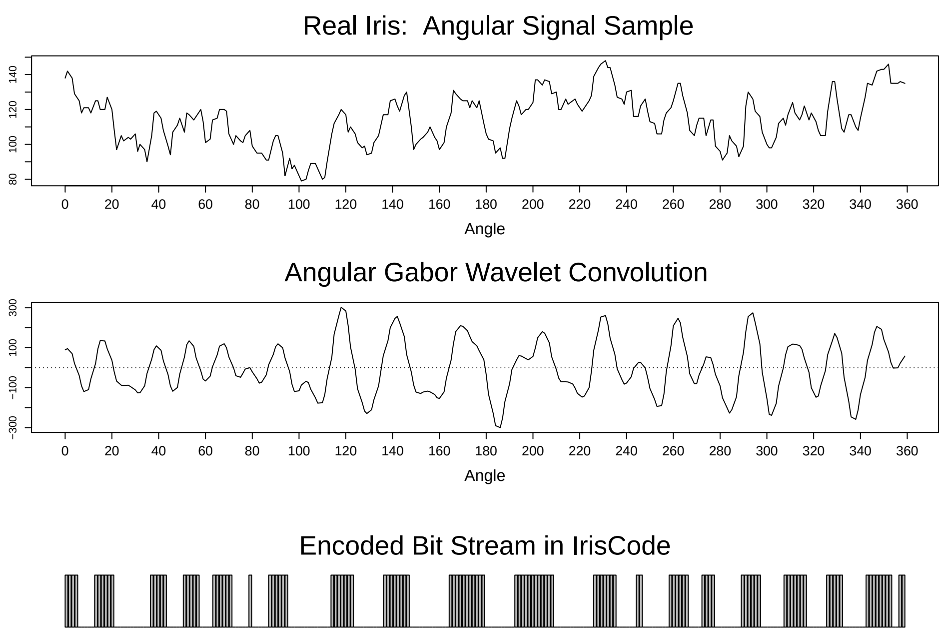
b_1	b_2	b_3	b_4	b_5	b_6	b_7
1	1	0	1	0	0	0

Evaluate three syndromes that are derived upon reception of the 7-bit block: $s_1 = b_1 \oplus b_3 \oplus b_5 \oplus b_7$, $s_2 = b_2 \oplus b_3 \oplus b_6 \oplus b_7$, $s_4 = b_4 \oplus b_5 \oplus b_6 \oplus b_7$, and provide the corrected 7-bit block that was the input to this noisy channel. [5 marks]

(*b*) Consider a noisy channel that suffers from additive noise which is not uniform in power across frequencies, but where instead the noise power attenuates exponentially with frequency ω. This channel has a signal-to-noise ratio $\text{SNR}(\omega)$ that therefore improves at higher frequencies, as: $\text{SNR}(\omega) = 2^{\omega} - 1$ over the possible range. Within some interval $[\omega_1, \omega_2]$ of the spectrum, with $\omega_2 > \omega_1 > 0$, what is the information capacity C of this noisy channel in bits/sec? [5 marks]

(*c*) Explain the Lempel-Ziv lossless compression algorithm, including its principle, its steps, its payload requirements, and its adaptive variants. [5 marks]

(*d*) The following traces show how an IrisCode bit sequence (bottom) is computed from a sample of human iris texture (top), for compact pattern encoding whose high entropy makes identification very reliable. Explain the rôles of Gabor wavelets and of Logan's Theorem in reference to these traces. [5 marks]



11 Natural Language Processing

Compositional semantics is concerned with modelling the meaning of phrases and sentences. It relies on the principle of compositionality, which states that the meaning of a phrase or a sentence can be derived from the meanings of its parts, guided by syntactic structure.

(*a*) What do the following examples demonstrate about the difficulties that compositional semantic models face?

(*i*) *it snows*

(*ii*) *rock the boat*

(*iii*) *enjoy a cigarette*
enjoy a book

(*iv*) *inject life*
inject hydrogen

[2 marks each]

(*b*) Distributional representations of phrases and sentences can be derived by combining, or *composing*, the distributional representations of individual words in them. Briefly describe:

(*i*) three limitations of vector mixture models;

(*ii*) how lexical function models address these limitations.

[3 marks each]

(*c*) Describe how adjectives can be represented as lexical functions and how such lexical functions can be learned from a text corpus. [6 marks]

12 Optimising Compilers

The following three-address pseudo-code shows a compiler's intermediate representation of a basic block, with destination registers as the first operand of each non-store instruction. The compiler has already determined that there are no aliases between instructions accessing memory (ie: instructions 1, 2, 6, 8, 9 all access different memory locations).

```
1:  load  $1, 0($5)
2:  load  $2, 8($5)
3:  sub   $3, $1, $2
4:  add   $4, $1, $2
5:  add   $4, $4, #8
6:  store $4, 0($3)
7:  and   $2, $4, #4095
8:  load  $1, 16($5)
9:  store $2, 0($1)
```

(*a*) What are the two goals of instruction scheduling? [2 marks]

(*b*) With reference to the above code, describe the constraints on reordering instructions. [4 marks]

(*c*) Briefly describe an instruction scheduling algorithm, then apply it to this code when targeting a simple 5-stage pipelined processor (IF, RF, EX, MEM, WB), showing the time saving against the original schedule. You may assume each instruction takes one cycle to execute and that the hardware supports delayed loads with feed-forwarding and interlocks. [8 marks]

(*d*) How does register allocation interact with instruction scheduling? [3 marks]

(*e*) Assuming that all registers are dead at the end of this code, are more registers required to schedule this basic block without any stalls? Justify your answer. [3 marks]

13 Principles of Communications

(*a*) The Transmission Control Protocol (TCP) employs a congestion control algorithm to respond to feedback about network conditions. Describe three common circumstances in which this might not be triggered. [10 marks]

(*b*) Wikipedia occasionally suffers from so-called "edit wars", when different authors successively alter content, backwards and forwards, between two different viewpoints. How would you design a system to stabilize such behaviour by introducing negative feedback controls to slow down or even stop the ability of such authors to make edits to articles in this situation? A qualitative answer identifying the features of such a system will suffice. [10 marks]

14 Security II

(*a*) Single Sign-On systems (SSO) may be classified (after Pashalidis and Mitchell) as either pseudo-SSO or true-SSO, and as either local or proxy-based. These two dimensions partition the space of SSO systems into 4 quadrants.

(*i*) After describing what an SSO is and the purpose it serves, explain the terms "pseudo-SSO", "true-SSO", "local" and "proxy-based". [4 marks]

(*ii*) Describe an example SSO system for each quadrant (real or imaginary, but with a concrete use case), briefly sketching how it works. [8 marks]

(*iii*) What are the security, privacy, usability and deployability problems inherent to each of the four quadrants? [8 marks]

END OF PAPER