CST2
COMPUTER SCIENCE TRIPOS Part II

Monday 7 June 2021 11:30 to 14:30 BST

COMPUTER SCIENCE Paper 9

*Answer **five** questions.*

*Submit each question answer in a **separate** PDF. As the file name, use your candidate number, paper and question number (e.g., `1234A-p9-q6.pdf`). Also write your candidate number, paper and question number at the start of each PDF.*

You must follow the official form and conduct instructions for this online examination

1 Advanced Algorithms

(a) Assume you have a randomised approximation algorithm for a maximisation problem, and your algorithm achieves an approximation ratio of 2. What can you deduce for

$$\mathbf{E}[C^*/C],$$

where C^* is the cost of the optimal solution, C is the cost of the solution of the approximation algorithm, and $\mathbf{E}[.]$ denotes the expectation? [4 marks]

(b) Consider the following optimisation problem on graphs: Given an undirected, edge-weighted graph $G = (V, E, w)$ with $w : E \to \mathbb{R}^+$, we want to find a subset $S \subseteq V$ such that $w(S, V \setminus S) = \sum_{e \in E(S, V \setminus S)} w(e)$ (the total sum of weights over all edges between S and $V \setminus S$) is maximised.

(i) Design a polynomial-time approximation algorithm for this problem. Also analyse its running time and prove an upper bound on the approximation ratio. [8 marks]

(ii) Find a graph which matches your upper bound on the approximation ratio from Part (b)(i) as closely as possible. [4 marks]

(iii) Consider now the following generalisation of the problem. Given an integer $k \geq 2$, we want to partition V into disjoint subsets $S_1, S_2, \ldots, S_k$ so that we maximise

$$\sum_{i=1}^{k} w(S_i, V \setminus S_i).$$

Describe an extension of your algorithm in Part (b)(i). What approximation ratio can you prove for this algorithm? [4 marks]

2 Bioinformatics

(a) Compute the local alignment between the following sequences: GATTACA, TATACG with the following rules: match score = +5, mismatch = −3, gap penalty = −4 and discuss how the alignment depends on the choices of match scores, mismatch and gap penalty. [5 marks]

(b) Discuss how a local alignment algorithm allows identification of internal sequence duplications. [3 marks]

(c) Define the UPGMA algorithm and state and justify its complexity. What is the output of the algorithm given the distance matrix of the species X_1, X_2, X_3, X_4 below?

$$\begin{pmatrix} species & X_1 & X_2 & X_3 \\ X_2 & 2 & & \\ X_3 & 4 & 4 & \\ X_4 & 6 & 6 & 6 \end{pmatrix}$$

[4 marks]

(d) Discuss a method to perform random access in DNA-based storage memory. [4 marks]

(e) Discuss with one example the complexity of the Gillespie algorithm and comment on the main differences with respect to a deterministic approach. [4 marks]

3 Business Studies

You create a startup that helps companies manage office space use to boost employee wellbeing while minimising the potential for workplace virus transmission.

After an intense 9 months of building a prototype and trialling it with early customers you demonstrate enough market traction to raise funding and hire a team of 15 people to support your growing customer base.

You've just completed your first month of work with the new team in your new and larger office space when the government announces a new 2-month lockdown where your staff must work from home.

As workers are no longer going to be in offices at the same rate, sales slow down. Confident that this is just temporary you decide to put 5 employees on furlough, with government support, until the lockdown is lifted.

(*a*) Discuss what needs to be done to support your employees as you transition the business to everyone working from home. [8 marks]

(*b*) Unfortunately, at the end of the initial lockdown period the government decides to extend it by another month. Looking at your financial forecasts you do not have enough funds to continue the business in its current form. Discuss what options are available to you. [12 marks]

4 Comparative Architectures

(*a*) A superscalar processor may speculatively execute loads even when one or more earlier stores have not yet computed their memory addresses. In practice, we would need to restart execution from the speculative load if a memory-carried dependency is subsequently detected.

(*i*) With the help of some additional hardware it is possible to record which loads cause such ordering violations. Briefly outline how this could be done and how such a record could be used to help improve performance. [3 marks]

(*ii*) Describe why such a scheme may unnecessarily delay the issuing of a load even when the mechanism correctly recalls that the load has led to an order violation between a store and load in the past? [4 marks]

(*b*) Why might it also be advantageous for a superscalar processor to predict whether a particular load will hit or miss in the processor's L1 data cache? [3 marks]

(*c*) You are asked to design hardware to run artificial neural network applications in a high-performance and energy-efficient manner. Such workloads can typically make good use of many multiply-accumulate (MAC) units operating in parallel and narrow datatypes. Your system is required to support a range of different neural networks that vary considerably in the type of computations they perform. You consider three approaches: (1) to use a multicore processor; (2) to design a single domain-specific accelerator; (3) to compose your design from two or more domain-specific accelerators where each is specialised for different types of neural network.

(*i*) What are the advantages and disadvantages of each approach? [6 marks]

(*ii*) Describe one possible way of organising the multicore processor and a possible choice for the architecture(s) of its individual cores. Briefly justify your design decisions. [4 marks]

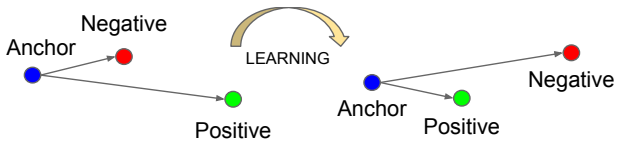
5 Computer Vision

(*a*) Inferring a 3D object shape from shading variation across a surface depends on assumptions about how Lambertian or how specular each area is. For a surface reflectance map $\phi(i, e, g)$ having a mixed form,

$$\phi(i, e, g) = \frac{s(n+1)(2\cos(i)\cos(e) - \cos(g))^n}{2} + (1-s)\cos(i)$$

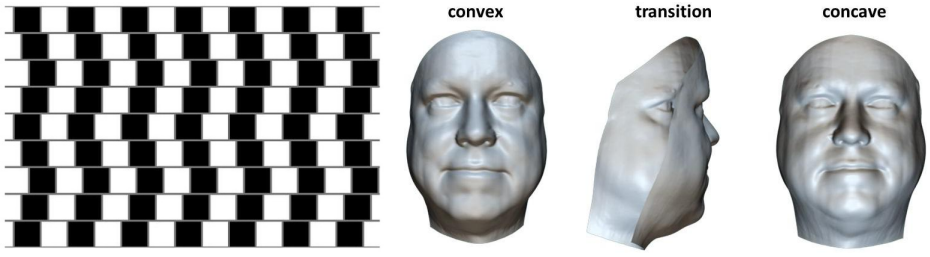
give a range of values for s and n that would arise for: (*i*) a matte surface, and (*ii*) a glossy surface. What form of reflectance map $\phi(i, e, g)$ describes (*iii*) a mirror, and what form describes (*iv*) the lunar surface? (*v*) Why is $\phi(i, e, g)$ as specified above sometimes called the "Face Powder Equation"? (*vi*) How does the lunar form of $\phi(i, e, g)$ explain why the full moon looks like a flat 2D penny in the sky, rather than a 3D sphere like a ping-pong ball? [8 marks]

(*b*) A breakthrough in face recognition accuracy arose when machine learning on big datasets minimised a loss function involving terms like $\| f(x_i^a) - f(x_i^p) \|^2$ and $\| f(x_i^a) - f(x_i^n) \|^2$ on triples of embeddings for $f(x_i^a)$ (anchor faces), $f(x_i^p)$ (positive examples: same face), and $f(x_i^n)$ (negative examples: different faces).



This approach treats false matches and failures-to-match as equally bad errors. But their costs are vastly different for a 1-to-1 face verification system (that just makes a 'yes/no' decision), versus a face identification system that may need to search a database the size of an entire nation, returning an actual identity. Propose a parameterised loss function for an algorithm that can be tuned for the different costs of the two error types, false matches and failures-to-match. Explain how its parameter(s) should reflect the numbers of potential false match collisions that must be avoided in a large-scale search. [6 marks]

(*c*) A surprising aspect of human vision is the prevalence of quite striking illusions, which cannot be defeated even by being aware of them. Are visual illusions "bugs", or "features" that should be built into computer vision algorithms? Consider in your answer both the tiling illusion (in which all horizontal lines really are parallel), and the hollow mask illusion below (in which the face always appears convex even when the mask is concave in presentation).



[6 marks]

6 Cryptography

(*a*) *CrashHash* is a cryptographic hash function invented by your colleague this morning. It zero-pads input X, splits it into n 256-bit blocks $x_1 \| x_2 \| \ldots \| x_n = X \| 0^{(-|X|) \bmod 256}$ and then appends a length-indicator block $x_{n+1} = \langle |X| \rangle$, as in the Merkle–Damgård construction. It then iterates a 512-bit to 256-bit compression function of the form $C(K, M) = E_K(M)$, where $E_K(M)$ is a blockcipher E applied with 256-bit key K to 256-bit message block M, as

$$\begin{aligned} z_1 &= C(\langle 0 \rangle, x_1) \\ z_i &= C(z_{i-1}, x_i) \qquad (1 < i \leq n+1) \end{aligned}$$

The value $H(X) = z_{n+1}$ is the hash value returned. Show that *CrashHash* is not collision resistant, even if E is replaced with an *ideal cipher*. [6 marks]

(*b*) (*i*) How can one modify an implementation of the DES encryption function to obtain the decryption function? [4 marks]

(*ii*) Name two other features of DES that made it well suited for hardware implementation. [2 marks]

(*c*) Your colleague has generated a set of $m = 200\,000$ RSA key pairs that include a modulus $n_i = p_i q_i$ where p_i and q_i are 1536-bit prime numbers (for $1 \leq i \leq m$). The corresponding p_i and q_i values were discarded immediately after key generation and are no longer available.

Due to a bug in your colleague's key-generation software, two types of fault have appeared in a random subset of the issued key pairs:

(*i*) For some key pairs i we have $p_i = q_i$.

(*ii*) For some key pairs i there exists another key pair j in that set with $p_i = p_j$ and $i \neq j$.

Suggest practical tests that can identify all public keys affected by either of these problems and state how often the algorithms involved have to be executed for this task. [4 marks]

(*d*) Calculate $7^{2000} \bmod 100$ by hand. [4 marks]

7 Denotational Semantics

Say whether the following statements are true or false with justification. You may use standard results provided that you state them clearly.

(*a*) For all PCF types τ and terms $M \in \mathrm{PCF}_\tau$, if $[\![M]\!] = \bot_{[\![\tau]\!]}$ then $M \cong_{\mathrm{ctx}} \Omega_\tau : \tau$.
[4 marks]

(*b*) For all PCF types τ and terms $M \in \mathrm{PCF}_\tau$, if $[\![M]\!] = \bot_{[\![\tau]\!]}$ then $M \not\Downarrow_\tau$.
[4 marks]

(*c*) For all PCF types τ and terms $M \in \mathrm{PCF}_\tau$, if $M \cong_{\mathrm{ctx}} \Omega_\tau : \tau$ then $M \not\Downarrow_\tau$.
[4 marks]

(*d*) For all PCF types τ and terms $M \in \mathrm{PCF}_\tau$, if $M \not\Downarrow_\tau$ then $M \cong_{\mathrm{ctx}} \Omega_\tau : \tau$.
[*Hint:* Recall the extensionality properties of contextual equivalence.]
[4 marks]

(*e*) For all PCF types τ and terms $M \in \mathrm{PCF}_\tau$, if $M \cong_{\mathrm{ctx}} \Omega_\tau : \tau$ then $[\![M]\!] = \bot_{[\![\tau]\!]}$.
[*Hint:* Recall the parallel-or test functions.] [4 marks]

8 Hoare Logic and Model Checking

We consider the CTL temporal logic over atomic propositions $p \in \mathsf{AP}$:
$\psi \in \mathsf{StateProp} ::= \bot \mid \top \mid \neg\psi \mid \psi_1 \wedge \psi_2 \mid \psi_1 \vee \psi_2 \mid \psi_1 \to \psi_2 \mid p \mid \mathsf{A}\ \phi \mid \mathsf{E}\ \phi$,
$\phi \in \mathsf{PathProp} ::= \mathsf{X}\ \psi \mid \mathsf{F}\ \psi \mid \mathsf{G}\ \psi \mid \psi_1\ \mathsf{U}\ \psi_2$.

(a) Alice (a), Bob (b), and Carol (c) are bank tellers. They sit at their till (t), until they need to give money to their customers, which they can do in gold (g) or silver (s) coins, which are stored in two different vaults. Each vault needs two tellers to turn keys simultaneously, but if it determines that there are more than two tellers present, it locks itself and phones the police. Once they have retrieved the coins, they can return with coin (r). This yields atomic propositions $AP = Pers \times Loc$, where $Pers ::= \mathsf{a} \mid \mathsf{b} \mid \mathsf{c}$ and $Loc ::= \mathsf{t} \mid \mathsf{g} \mid \mathsf{s} \mid \mathsf{r}$, so that for example a state labelled with $\{\mathsf{at}, \mathsf{bs}, \mathsf{cr}\}$ is one where Alice is at her desk, Bob is waiting to open the silver vault, and Carol is returning with coin.
Give CTL formulas for

(i) Alice repeatedly serves clients with coins. [2 marks]

(ii) When Bob picks a vault, he stays there until it gets opened. [2 marks]

(iii) Carol is always able to serve clients with coins. [2 marks]

(iv) The vaults never lock. [2 marks]

(b) Explain why Carol's property cannot be expressed in LTL. [2 marks]

(c) A chemist is trying to determine what can be synthesised from the chemicals they have. They know all possible reactions: $2\ H_2 + O_2 \to 2\ H_2O$, $C + O_2 \to CO_2$, etc.

(i) Describe a model of reactions from given starting quantities. [3 marks]

(ii) Keeping track of the exact amount of chemicals is challenging, so the chemist looks only at whether each is present. Describe such an abstract model, and give a simulation relation between the two models. [4 marks]

(iii) State, for each of these two properties, which of the models (i) and (ii) is such that if the property holds for it, then it holds for the other:
1/ that dangerous chemicals like CO are never synthesised;
2/ that desirable chemicals like pure gold (Au) can be synthesised.
Explain why. [3 marks]

9 Information Theory

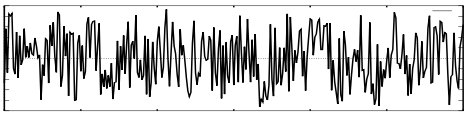
(*a*) A long-term and self-replicating data storage system based on DNA sequences is being developed. Advantages include huge information density ($\sim 10^{19}$ bits/cm^3) and extreme persistence: dinosaur DNA can still be extracted from fossils. The letters A,C,G,T each occur with equal probability, independently, without sequence constraints. Consider sequences consisting of 100 such letters.

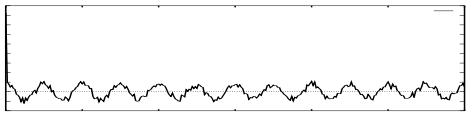
(*i*) How many sequences are possible, and with what probabilities? [2 marks]

(*ii*) Random variable X selects such a sequence. Calculate $H(X)$, the entropy of X, starting from Shannon's definition. [2 marks]

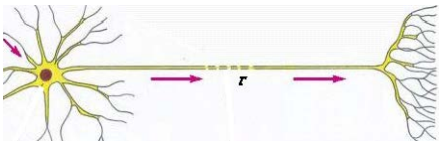
(*iii*) Sequence replication may be corrupted such that the last two letters are reproduced randomly in the post-replication sequences, denoted Y. What is the conditional entropy $H(X|Y)$, and what is the mutual information $I(X;Y)$ for this error-prone replication process? [4 marks]

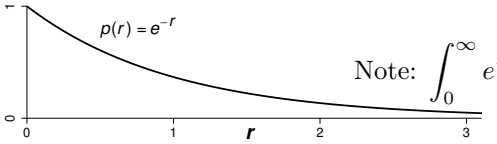
(*b*) Financial markets generate daily asset valuations like the time-series $f(t)$ in the left panel, reflecting the dynamics of greed and fear. But underlying such fluctuating indices there may exist meaningful trends, such as a business cycle (right panel). Write an auto-correlation integral that can extract the coherent quasi-periodic signal on the right from noisy valuations $f(t)$, and explain how computing the Fourier transform $F(\omega)$ of $f(t)$ makes it efficient. [5 marks]





(*c*) Brain tissue contains about 10^5 neurones per mm^3, and each neurone has a single output axon whose length r (in dimensionless units) before terminating at synapses to other neurones has probability density distribution $p(r) = e^{-r}$.





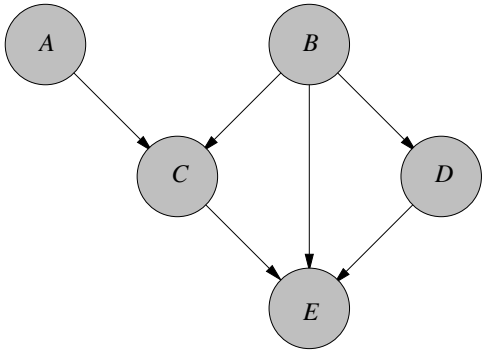
Note: $\int_0^\infty e^{-r}dr = 1$

(*i*) Define differential entropy h for continuous random variables in terms of general probability density distribution $p(x)$, and then calculate the value of h in bits for this axonal length distribution $p(r) = e^{-r}$. [5 marks]

(*ii*) If the axon's branching terminals make altogether about 1,000 synapses (connections) with different neurones within the axonal tree's 1 mm^3 volume, uniformly distributed, roughly how many bits of entropy describe the uncertainty of whether a neurone gets such a connection? [2 marks]

10 Machine Learning and Bayesian Inference

Consider the following Bayesian Network:



All random variables (RVs) are Boolean. For an RV R we denote $R = T$ by r and $R = F$ by $\overline{r}$. We have $\Pr(a) = 0.1$, $\Pr(b) = 0.2$, $\Pr(d|b) = 0.7$ and $\Pr(d|\overline{b}) = 0.4$. For the remaining RVs we have

A	B	$\Pr(c\|A,B)$
F	F	0.2
F	T	0.2
T	F	0.5
T	T	0.6

C	B	D	$\Pr(e\|C,B,D)$
F	F	F	0.3
F	F	T	0.5
F	T	F	0.6
F	T	T	0.3
T	F	F	0.1
T	F	T	0.2
T	T	F	0.1
T	T	T	0.9

In this question you must use the *Variable Elimination* algorithm to compute $\Pr(A|\overline{e})$. You should begin with the factorisation

$$\Pr(A|\overline{e}) = \Pr(A) \sum_B \Pr(B) \sum_C \Pr(C|A,B) \sum_D \Pr(D|B) \Pr(\overline{e}|B,C,D).$$

You should express factors as tables of integers, leaving any necessary normalisation until the final step in Part (d).

(a) Define *conditional independence* of two RVs X and Y with respect to a third RV Z. [2 marks]

(b) Deduce the factor $F_{E,\overline{D}}(B,C)$ corresponding to the summation over D. [8 marks]

(c) Deduce the factor $F_{E,\overline{D},\overline{C}}(A,B)$ corresponding to the summation over C. [6 marks]

(d) Complete the computation to find the distribution $\Pr(A|\overline{e})$. [4 marks]

11 Mobile and Sensor Systems

(*a*) Contrast the measurement of biosignals from conventional medical devices with those from smartphone and wearable device sensors. [5 marks]

(*b*) A researcher wishes to screen users for the hand tremor symptom via a smartphone app. They recruit a cohort of people with diagnosed tremor and a control cohort about which nothing is known. The researcher has a smartphone that collects data from its sensors. They give it to each participant and ask them to hold it as still as they can while sitting. The researcher then uses the data captured to develop a tremor classifier that has sensitivity 0.97 and specificity 0.99.

(*i*) Why is getting a high specificity a particular priority for a smartphone-based screening app? [4 marks]

(*ii*) Explain why the prevalence of the disease must also be taken into account in deciding whether to deploy this app, illustrating your answer by considering tremor prevalences of 0.1% and 5%. [3 marks]

(*iii*) The smartphone app is deployed. It asks users to test their tremor monthly at home, when sitting and trying to hold their phone still. Why might the screening be less effective than expected from the collected data? [3 marks]

(*iv*) The researcher changes the app to do background tremor screening. The app now constantly monitors for the user holding the phone appropriately. When it observes this it captures sensor data and applies the tremor classifier. Discuss the advantages and disadvantages of this approach compared to the previous approach. [5 marks]

12 Optimising Compilers

The following excerpt from a program in C-style code is optimised with a compiler using code-motion transformations. The function `read()` returns a signed integer from the user.

```
 l0:  a = read();
 l1:  b = read();
 l2:  p = &a;
 l3:  q = &b;
 l4:  r = &p;
 l5:  if (read() > 0) {
 l6:    a = b + 5;
 l7:  } else {
 l8:    i = 0;
 l9:    while (i < 10) {
l10:      c = b + 5;
l11:      **r += *q;
l12:      i += 1;
l13:    }
l14:    a += c;
l15:  }
l16:  print(a);
```

(a) Describe loop-invariant code motion (LICM) and which expresion(s) in the loop above it should move. [2 marks]

(b) Describe a simple data-flow analysis and a way of using it to identify loop-invariant expressions. Use this to analyse the code above. [5 marks]

(c) Explain whether all expressions described in Part (a) are found through the analysis in Part (b). [2 marks]

(d) Describe an analysis that can aid in making LICM more precise in this example. [3 marks]

(e) Apply the analysis from Part (d) to the code above and redo the analysis from Part (b) to show which expressions described in Part (a) are now found. [4 marks]

(f) Describe another *code motion* transformation that could be applied to the code after LICM and show the final code after its application. [4 marks]

13 Principles of Communications

(*a*) Multicast routing provides IP packet delivery from a source to a set of receivers. One clear use case for this is for large scale content distribution (e.g. software updates). In such a case, we would expect the end-to-end protocol to provide flow and congestion control. How might such a reliable multicast transport protocol be designed? [10 marks]

(*b*) Mobile systems move. In cellular networks, this is can be handled by the radio access network, and measuring signal strength to determine to which cell a handset is best assigned. In the Internet, the IP layer typically hides this information. How might we combine information across layers to make mobile IP routing more efficient? In your discussion, pay attention to problems of software layering, and also of managing the dynamics (e.g. route flapping) in such systems. [10 marks]

14 Quantum Computing

(*a*) Find the eigenvectors, eigenvalues and spectral decomposition of the observable

$$\mathrm{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and give the outcome of measuring the expectation of the observable on the states:

(*i*) $|0\rangle$

(*ii*) $\frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$

(*iii*) $\frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle$

[8 marks]

(*b*) A quantum mechanical system has Hamiltonian

$$\mathrm{H} = \mathrm{H}_1 + 2\mathrm{H}_2$$

It is desired to use a quantum computer to approximately simulate the operator $e^{-i\mathrm{H}t}$ for some t. It is possible to build quantum circuits U_1 and U_2 to perform the operations

$$U_1 = e^{-i\mathrm{H}_1 t}$$
$$U_2 = e^{-i\mathrm{H}_2 t}$$

Give a circuit, U, consisting of one of more instances of U_1 and U_2 that approximates $e^{-i\mathrm{H}t}$ such that $e^{-i\mathrm{H}t} - U = \mathcal{O}(t^3)$. Show your calculations to verify that the circuit does indeed achieve this. [8 marks]

(*c*) Quantum Phase Estimation can be used to estimate the ground state energy of quantum mechanical systems. The Inverse Quantum Fourier Transform is a key component of Quantum Phase Estimation. Give the circuit for the 2-qubit Inverse Quantum Fourier Transform using only gates from the set $\{H, C\!T, \mathrm{CNOT}\}$, where C$T$ is a controlled T gate. [4 marks]

15 Types

(*a*) In a simply-typed lambda calculus augmented with first-class continuations, booleans, a list type and its iterator (i.e., fold, but not full recursion), write a function

$$\texttt{every} : (X \to \mathsf{Bool}) \to \mathsf{List}\, X \to \mathsf{Bool}$$

such that $\texttt{every}\, p\, xs$ returns true if every element of xs satisfies p, and false otherwise. This function should also stop iterating over the list as soon as it finds a false element. You may use SML- or OCaml-style notation if desired, but explain any notation used beyond the basic lambda calculus.

[4 marks]

(*b*) In the monadic lambda calculus with state, suppose we change the typing rule for reading locations to not cause a monadic effect: If we suggest changing the monadic lambda calculus to permit treating reads as pure:

$$\frac{l : X \in \Sigma}{\Sigma; \Gamma \vdash\, !l : X}$$

(*i*) Is this rule still typesafe? Informally but carefully justify your answer. [2 marks]

(*ii*) Is the following *common subexpression elimination* transformation sound? Either give an argument why it is, or supply a counterexample and explain why it shows it is not. [6 marks]

```
let x = return e1;                 let x = return e1;
let y = e2;             =====>     let y = e2;
let z = return e1;                 [z/x]e3
e3
```

(*c*) In System F augmented with existential types, give an existential type for the interface of the natural numbers, and give an implementation for it. [8 marks]

END OF PAPER