COMPUTER SCIENCE TRIPOS Part II

Thursday 8 June 2017 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
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

1 Advanced Algorithms

(*a*) Give two examples of greedy algorithms and state their approximation ratios. [4 marks]

(*b*) Consider the Centre Selection Problem, defined as follows. The input consists of n points $p_1, p_2, \ldots, p_n$ in a metric space, and an integer $k > 0$. The goal is to find k centres $C = \{c_1, c_2, \ldots, c_k\}$ (not necessarily from among the n points) so that $r(C) = \max_{1 \leq i \leq n} \operatorname{dist}(p_i, C)$, where $\operatorname{dist}(p_i, C) = \min_{1 \leq j \leq k} \operatorname{dist}(p_i, c_j)$, is minimised.

(*i*) Consider the standard greedy approach: solve the problem optimally for $k = 1$ and then extend the solution to larger values of k by adding the optimal point to the current solution. Why is this likely to give a poor result? [4 marks]

(*ii*) Consider the following algorithm to solve the Centre Selection Problem:

```
Let C be the empty set
Repeat k times
   Select a point p_i with maximum distance dist(p_i,C)
   Add point p_i to the set C
Return C
```

Derive a lower bound for this algorithm on the minimum pairwise distance among the chosen centres C. [4 marks]

(*iii*) Give an upper bound, as tight as possible, on the approximation ratio of the algorithm in part (*b*)(*ii*). [2 marks]

(*iv*) Give a detailed analysis in order to justify your answer for part (*b*)(*iii*). *Hint*: Exploit the lower bound derived in part (*b*)(*ii*) in order to construct disjoint balls around the centre points. [6 marks]

2 Bioinformatics

(*a*) For problems involving hidden Markov models (HMM), when would you use the Baum-Welsh algorithm and when the Viterbi algorithm and why? [6 marks]

(*b*) Discuss how a sequence alignment might be evaluated statistically, illustrating your answer with an example. [6 marks]

(*c*) What is the condition for fitting a phylogenetic tree to a matrix? [2 marks]

(*d*) Discuss how to find matches in a genome sequence efficiently. [6 marks]

3 Computer Systems Modelling

Consider the simulation of a simple server with an unspecified arrival process and and unspecified service distribution. Observations of when customers arrive into the queue, and when they complete service, are readily available. Service is strictly first-come-first-served.

(*a*) As an initial step it is assumed that the arrival process is Poisson with arrival rate λ.

(*i*) Describe the inverse transform method for generating continuous random variables with specified distributions, assuming a source of random variables from $U(0, 1)$, that is, uniformly distributed on the interval $[0 : 1]$. [5 marks]

(*ii*) Apply this method to generate random variables representing the time intervals between arrivals. [3 marks]

(*iii*) A student generates arrival events by dividing time into small intervals Δt such that $\lambda\Delta t \ll 1$ and takes a sample u from $U[0, 1)$. If $u < \lambda\Delta t$ then an arrival is generated, otherwise not. Is this a good way of generating Poisson arrivals with rate λ? Explain your answer. [2 marks]

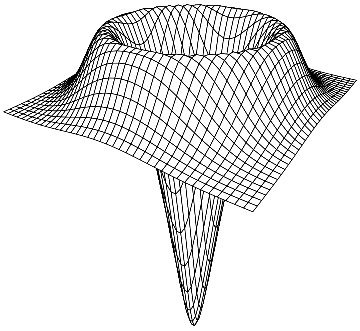
(*b*) In order to investigate the arrival process, observations of arrivals into the system are gathered and stored as a sequence $Y_1, Y_2, \ldots, Y_n$ of time intervals between arrivals.

(*i*) Given an assumption that the arrival process is Poisson, how could λ be estimated? [2 marks]

(*ii*) How might one use the Kolmogorov-Smirnov test to test the hypothesis that the arrival process is Poisson? [8 marks]

4 Computer Vision

(*a*) Define the gradient vector field $\vec{\nabla} f(x,y)$ over an image $f(x,y)$, and explain what makes it useful. Contrast its features and capabilities with the $\nabla^2 G_\sigma(x,y)$ (Laplacian of a Gaussian) operator shown below. Identify their respective orders as differential operators, explain how they can be implemented, and discuss any neurobiological analogues for both. [8 marks]



(*b*) Define a "hypercolumn" of neurones in the brain's primary visual cortex. Explain what are the main coding variables being spanned by a hypercolumn, roughly how many neurones it encompasses, how much of visual space it processes, and make a drawing of its architectural organisation. [7 marks]

(*c*) Explain the Retinex Algorithm, starting with the problem it seeks to solve and why the problem arises. [5 marks]

5 Denotational Semantics

(*a*) (*i*) Give the grammar defining the PCF expressions that are values. [2 marks]

(*ii*) Prove or disprove that, for every PCF type τ, there is a closed PCF expression that is not a value of type τ. [2 marks]

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

(*ii*) Prove or disprove that

$$(\mathbf{fn}\ n : nat.\, n) \cong_{\text{ctx}} \big(\mathbf{fn}\ n : nat.\, \mathbf{succ}(\mathbf{pred}(n))\big) : nat \to nat$$

[2 marks]

(*c*) For every pair of closed PCF expressions M, N of type nat, let $F_{M,N}$ be the closed PCF expression of type $(nat \to nat) \to (nat \to nat)$ given by

$$\begin{array}{l} \mathbf{fn}\ f : nat \to nat.\, \mathbf{fn}\ n : nat. \\ \quad \mathbf{if}\ \mathbf{zero}(n)\ \mathbf{then}\ M \\ \quad \mathbf{else}\ \mathbf{if}\ \mathbf{zero}(\mathbf{pred}(n))\ \mathbf{then}\ N \\ \qquad\quad \mathbf{else}\ \mathbf{succ}(f(\mathbf{pred}(n))) \end{array}$$

(*i*) Give an explicit description of $[\![\mathbf{fix}(F_{M,N})]\!] \in (\mathbb{N}_\perp \to \mathbb{N}_\perp)$ in terms of $[\![M]\!], [\![N]\!] \in \mathbb{N}_\perp$. Justify your answer. [8 marks]

(*ii*) Prove or disprove that there are closed PCF expressions M, N of type nat such that $\mathbf{fix}(F_{M,N}) \cong_{\text{ctx}} \big(\mathbf{fn}\ n : nat.\, \mathbf{pred}(n)\big) : nat \to nat$. You may use any standard results provided that you state them clearly. [4 marks]

6 Digital Signal Processing

(*a*) A discrete-time LTI filter can be described through the locations of zeros and poles in the z-transform $H(z)$ of its impulse response. Consider an IIR filter of order 2 with $H(c_1) = H(c_2) = 0$ and $|H(z)| \to \infty$ for $z \to d_1$ and $z \to d_2$.

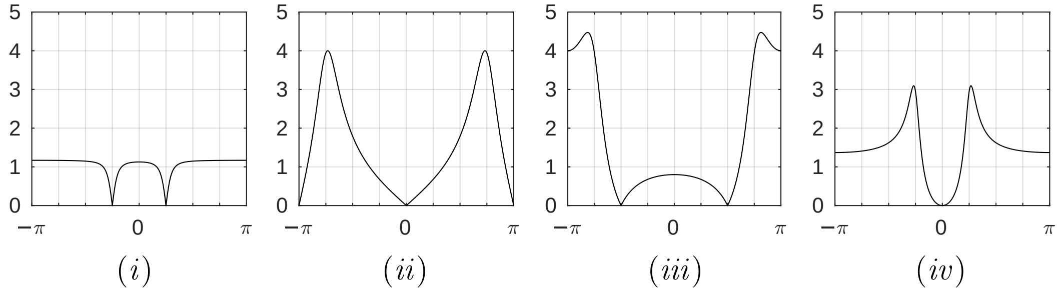
(*i*) What is the z-transform $H(z)$ of its impulse response? [2 marks]

(*ii*) What additional parameter (beyond c_1, c_2, d_1, d_2) is required to fully describe the impulse response of this filter? [1 mark]

(*iii*) What is the magnitude of the discrete-time Fourier transform (DTFT) of the impulse response of this filter? [2 marks]

(*iv*) Under what condition on c_1, c_2, d_1, and d_2 is the impulse-response of this filter real-valued? [2 marks]

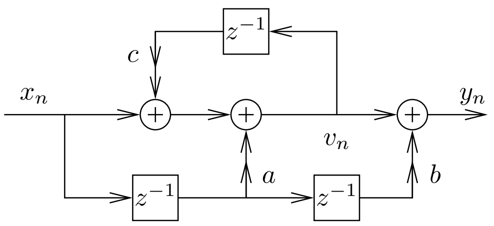
(*b*) The following plots show the magnitude of the DTFT of the real-valued impulse response of four different IIR filters:



The z-transform of each impulse response has two zeros and two poles. Each zero or pole is at one of these 12 possible locations: $e^{\pi jk/4}$ with $k \in \{0, \ldots, 7\}$ or $0.6 \pm 0.6j$ or $-0.5 \pm 0.5j$.

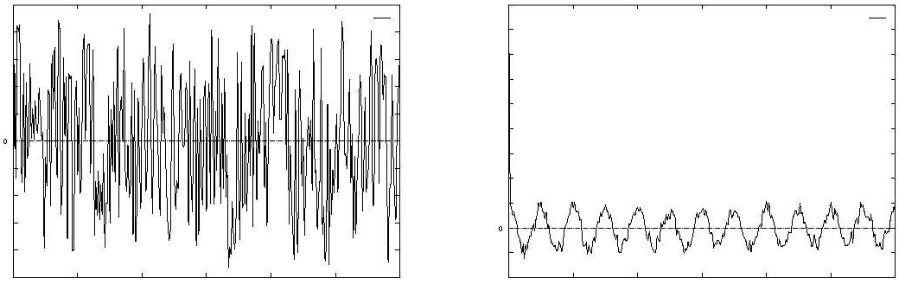
For each filter, state the location of both zeros and both poles. Explain the reasoning behind your choice. [8 marks]

(*c*) What is the z-transform $H(z)$ of the impulse response of the following filter? [5 marks]



7 Information Theory

(*a*) The left panel shows a very noisy cosmic signal $f(t)$, within which is buried a coherent quasi-periodic signal emitted by a pulsar (a collapsed neutron star). Write an auto-correlation integral which, when applied to the noisy signal $f(t)$, allows the clean quasi-periodic signal in the right panel to be detected in it. Explain why this works, and how computing the Fourier transform $F(\omega)$ of $f(t)$ can make the auto-correlation operation efficient. [5 marks]



(*b*) X and Y are discrete random variables described by entropies $H(X)$ and $H(Y)$, mutual information $I(X;Y)$, conditional entropies $H(X|Y)$ and $H(Y|X)$, and joint entropy $H(X,Y)$. Using these quantities as the labels for sets and subsets, with $\cup$ and $\cap$, evaluate the following into simpler quantities: [5 marks]

(*i*) $H(X|Y) \cup I(X;Y)$

(*ii*) $(H(X) \cup H(Y)) \cap I(X;Y)$

(*iii*) $(H(X|Y) \cup H(Y|X)) \cap I(X;Y)$

(*iv*) $H(X,Y) - H(X|Y)$

(*v*) $H(X|Y) \cap H(Y|X)$

(*c*) Explain how dictionary coding as a data compression strategy exploits the sparseness of strings, that is, the fact that the space of possible combinations is only very sparsely populated with those that actually occur. In the case of encoding English text with a coding budget of just 15 bits, contrast the richness of encoding letter-by-letter versus defining 15-bit pointers to a lexicon. How much of English vocabulary can be captured by such a vector quantisation strategy? What added cost does this strategy incur? [5 marks]

(*d*) Continuous random variables X and Y both have uniform probability density distributions on some interval. For X, $p(x) = \frac{1}{2}$ if $x \in [0,2]$ else 0, while for Y, $p(y) = \frac{1}{8}$ if $y \in [0,8]$ else 0. Calculate the differential entropies $h(X)$ and $h(Y)$. Provide an upper bound on the joint entropy $h(X,Y)$, and state the condition for reaching it. [5 marks]

8 Mobile and Sensor Systems

GoGo Ltd, a travel agent, wants to build a mobile phone app to support their customers' travel needs and to monitor their behaviour while travelling.

(*a*) The company has built an app component that uses the phone microphone to detect customer context (e.g., noisy place, cafe ambience) as well as emotions from voice pitch. However, when used in practice by the customers, the accuracy of the detection is much worse than the one obtained in the laboratory environment. Give reasons of why this might happen and possible solutions to improve accuracy. [4 marks]

(*b*) Discuss the privacy issues and possible alleviating mechanisms that can be employed when using the phone microphone in the app. [4 marks]

(*c*) On mobile devices, power always comes at a premium. Describe how you would design the microphone sensing and inferencing to limit energy usage. [6 marks]

(*d*) Customers are generally not keen to use data services abroad as roaming can be expensive. GoGo Ltd would like its app to offer a localised chat feature, allowing customers to send messages to each other without assuming any infrastructure.

Describe how you would design this component with considerations on MAC and networking layers. [6 marks]

9 Natural Language Processing

The goal of automatic summarisation is to produce a short version of a text that contains the most important or relevant information. In multi-document summarisation, we need to aggregate content from multiple documents into one cohesive summary.

(*a*) Describe two ways in which multi-document summarisation is more challenging than single-document summarisation. [4 marks]

(*b*) In query-focused multi-document summarisation, sentences can be selected for the summary based on maximal marginal relevance (MMR). Give the formula for MMR and explain the intuition behind it. [5 marks]

(*c*) The two sentences in each sentence pair below are linked by a particular rhetorical relation. Which rhetorical relation does each sentence pair exhibit?

(*i*) *The use of diesel in transport has come under increasing scrutiny in recent years. According to WHO, around three million deaths every year are linked to exposure to outdoor air pollution.*

(*ii*) *Nitrogen oxides can help form ground level ozone. This can exacerbate breathing difficulties.*

(*iii*) *Paris has already taken a series of steps to cut the impact of diesel cars and trucks. Vehicles registered before 1997 have already been banned from entering the city.*

[1 mark each]

(*d*) Briefly discuss how each of the following NLP techniques can be used in extractive summarisation.

(*i*) morphological processing;

(*ii*) syntactic parsing;

(*iii*) lexical and distributional semantics;

(*iv*) discourse parsing, i.e. identification of rhetorical relations.

[2 marks each]

10 Optimising Compilers

(*a*) Liveness and available expression analyses are instances of a general data-flow analysis framework. Describe this framework and contrast its use in these two analyses. [5 marks]

(*b*) Describe how liveness analysis can be used to identify two types of data-flow anomaly. [2 marks]

(*c*) Describe the difference between semantic and syntactic expression availability, giving example pseudo-code. Explain why available expression analysis is safe. [5 marks]

(*d*) Available expression analysis can be used to inform common subexpression elimination. Explain why it might be useful to run copy propagation after the this. [3 marks]

(*e*) Considering the following code, show that five registers are sufficient to hold its variables. Transform the code to determine the minimum number of registers required.

```
a = func(1);
b = func(2);
c = a * b;
print(b, c);
d = c - a;
print(a, d);
e = d - 1;
b = c + a;
d = e + b;
print(d, e);
f = e - 5;
print(b, f);
return;
```

[5 marks]

11 Principles of Communications

(*a*) What are the key differences between the data centre network environment and the broader, general Internet eco-system? How do these lead to simpler choices for offering performance guarantees for traffic? [10 marks]

(*b*) Network coding can be used to combine packets redundantly to provide error protection, but also to reduce the number of transmissions and retransmissions necessary. It is used in the transport layer, from each source, and in the network layer, combining data from multiple sources. How might a combination of these techniques simplify buffering in simple wireless devices that you might find in an Internet of Things (IoT) environment? [10 marks]

12 System-on-Chip Design

We require a hardware accelerator to compute

$$f(x) = \sum_{i=0..7} C[i] \times D[i+x]$$

where all values are 20-bit signed integers and the array $C[i]$ contains design-time constants. The array $D[]$ will contain 1024 values. We need to evaluate $f(x)$ as fast as possible. The values of x are unpredictable and one word of D gets updated by a separate process after every 50 or so evaluations of f. Ignore overflow.

(*a*) An early design holds both C and D in a common, single-ported RAM memory (i.e. one with one address bus). Given that single-cycle multiply-accumulate blocks are available, give a rough estimate of the performance of our system in clocks per evaluation. [3 marks]

(*b*) Suggest a small change to the early design that improves its performance and estimate the resulting performance. Suggest one or more further sensible improvements and indicate the new performance. [6 marks]

(*c*) The output from a High-Level Synthesis (HLS) compiler is generally RTL which is then fed to a logic synthesis compiler. Identify four tasks performed by the combined flow, explaining which stage does which. [4 marks]

(*d*) Using a block-structured high-level language like C or Java, or using pseudocode, briefly write an implementation of $f(x)$ that would be amenable to HLS compilation. State three properties of your implementation that make it likely to be acceptable to an HLS compiler and/or give good performance. What might influence the choice of design from the solution space in part (*b*)? You may ignore the mechanism by which D is updated. [7 marks]

13 Hoare Logic and Model Checking

Let AP be a set of atomic propositions, ranged over by p, q, and so on. Recall the grammar of Computation Tree Logic (CTL) path and state formulae:

$$\phi, \psi, \xi ::= \Diamond\Phi \mid \Box\Phi \mid \bigcirc\Phi \mid \Phi \texttt{ UNTIL } \Psi$$
$$\Phi, \Psi, \Xi ::= \top \mid \bot \mid p \mid \Phi \land \Psi \mid \Phi \lor \Psi \mid \Phi \Rightarrow \Psi \mid \neg\Phi \mid \forall\phi \mid \exists\phi$$

(*a*) Fix a CTL model $\mathcal{M} = \langle S, S_0, \rightarrow, L\rangle$. Suppose ϕ is a CTL path formula, Φ is a CTL state formula, s is a state in S, and π is an infinite path of states of S.

Define the two satisfaction relations $\mathcal{M}, \pi \models \phi$ and $\mathcal{M}, s \models \Phi$, explaining fully any notation that you use and any auxiliary definitions that you make. [5 marks]

(*b*) Suppose p, q, and r are atomic propositions taken from the set AP. Suppose also that we define a CTL model $\mathcal{M} = \langle S, S_0, \rightarrow, \mathcal{L}\rangle$, where:

$$S = \{s_0, s_1, s_2, s_3, s_4\} \quad S_0 = \{s_0, s_1\}$$
$$\rightarrow\ = \{(s_i, s_j) \mid i + j \text{ is even, for all } 0 \leq i \leq 4 \text{ and } 0 \leq j \leq 4\}$$
$$\mathcal{L}(s_0) = \mathcal{L}(s_2) = \mathcal{L}(s_4) = \{p\} \quad \mathcal{L}(s_3) = \{q\} \quad \mathcal{L}(s_1) = \{q, r\}$$

For each of the following, identify the set of all states $s \in S$ for which it holds:

(*i*) $\mathcal{M}, s \models \forall\Box p$,

(*ii*) $\mathcal{M}, s \models \exists\Diamond q$,

(*iii*) $\mathcal{M}, s \models \exists \bigcirc (p \land r)$

Explain fully how you computed your answer in each case. [6 marks]

(*c*) Define what it means for two CTL state formulae Φ and Ψ to be semantically equivalent, written $\Phi \equiv \Psi$. [3 marks]

(*d*) Show that $(\Phi \lor \Psi) \land \Xi$ and $(\Phi \land \Xi) \lor (\Psi \land \Xi)$ are semantically equivalent. [6 marks]

14 Topical Issues

Consider a corridor with spotlights embedded along its length in the ceiling. When a user is walking along the corridor the ambient light sensors found in their smartphone will show a peak whenever passing under a spotlight. By thresholding the ambient light sensor we can form a *spotlight detection* sensor, with two states: {under a spotlight, not under a spotlight}. This question considers how this sensor may assist positioning along a corridor.

(*a*) Recursive Bayes filtering is to be used to fuse the sensor measurements to estimate location. Describe the steps and key assumptions made by this filter. Mathematical equations are not required for full marks. [4 marks]

(*b*) A corridor is 21 m in length with 6 spotlights, numbered 1–6. They are spaced evenly at x=3, 6, 9, 12, 15, 18 m, where x runs along the corridor and x=0 is the corridor start. Due to a fault lights 4 and 5 are off.

(*i*) Sketch the measurement model for the spotlight detector. Sketch the belief distribution for $\mathbf{bel}(x)$ for the user's location obtained when the detector is turned on and immediately detects a spotlight. [3 marks]

(*ii*) Describe a suitable motion model assuming the smartphone's inertial sensors are used to perform *Pedestrian Dead Reckoning* (PDR) with a constant step length. Sketch $\mathbf{bel}(x)$ obtained after the phone reports three steps have been taken. [4 marks]

(*iii*) Sketch the new $\mathbf{bel}(x)$ obtained when the light positioning sensor now indicates the user is under a spotlight. Explain its relationship to the previous $\mathbf{bel}(x)$. [3 marks]

(*c*) A *Grid filter* is an implementation of a recursive Bayes filter where the $\mathbf{bel}(x)$ distributions are approximated by histograms. Compare the suitability of such a filter for this problem compared to Kalman and Particle filters, paying particular attention to any parameters and their effects. [6 marks]

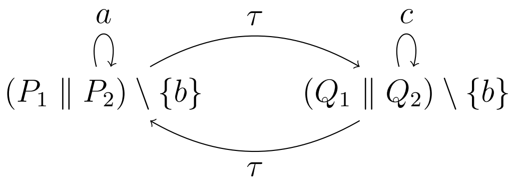
15 Topics in Concurrency

Let the pure CCS processes P_1, P_2, Q_1 and Q_2 be as follows.

$$P_1 \overset{\text{def}}{=} a.P_1 + \overline{b}.Q_1 \qquad Q_1 \overset{\text{def}}{=} b.P_1$$

$$P_2 \overset{\text{def}}{=} b.Q_2 \qquad Q_2 \overset{\text{def}}{=} c.Q_2 + \overline{b}.P_2$$

The transition system from $(P_1 \parallel P_2) \setminus \{b\}$ is as follows.



a
τ
c
$(P_1 \parallel P_2) \setminus \{b\}$
$(Q_1 \parallel Q_2) \setminus \{b\}$
τ

(a) Give full derivations for the two transitions that start from $(P_1 \parallel P_2) \setminus \{b\}$. [5 marks]

(b) The full modal-μ calculus has the syntax

$$A ::= T \mid S \mid \neg A \mid A_1 \wedge A_2 \mid A_1 \vee A_2 \mid \langle a \rangle A \mid [a]A \mid \nu X.A \mid \mu X.A \mid X,$$

where S is an arbitrary set of states. Give a semantics to closed formulas *without* using the abbreviations $\mu X.A \equiv \neg \nu X.\neg A[\neg X/X]$ and $[a]A \equiv \neg\langle a \rangle \neg A$. What condition must be placed on the occurrence of variables and why? [5 marks]

(c) Prove that the operation

$$X \mapsto [a]X$$

is $\bigcap$-continuous. [5 marks]

(d) Give a modal-μ formula that is satisfied by a process if, and only if, it is bisimilar to the process $(P_1 \parallel P_2) \setminus \{b\}$. You may assume that the process is only capable of actions labelled a, c and τ. [5 marks]

16 Types

(*a*) Existential types can be encoded in the Polymorphic Lambda Calculus (PLC) by defining $\exists\,\alpha\,(\tau)$ to be $\forall\beta\,((\forall\alpha\,(\tau\to\beta))\to\beta)$, where $\beta \neq \alpha$ and β does not occur free in the PLC type τ. Give the following definitions, justifying the typings in each case:

(*i*) A closed PLC term `pack` of type $\forall\alpha\,(\tau \to \exists\,\alpha\,(\tau))$. [5 marks]

(*ii*) A PLC term $\texttt{unpack}(M, x, M', \tau')$ satisfying $\Gamma \vdash \texttt{unpack}(M, x, M', \tau') : \tau'$ whenever $\Gamma \vdash M : \exists\,\alpha\,(\tau)$ and $\Gamma, x : \tau \vdash M' : \tau'$ hold, where x is not in the domain of Γ and α does not occur free in Γ or τ'. [5 marks]

(*b*) If $\Gamma \vdash M : \exists\,\alpha\,(\tau)$, $\Gamma, x : \tau \vdash M' : \tau'$ and $\Gamma \vdash N : \tau[\tau'/\alpha]$ hold, where x is not in the domain of Γ and α does not occur free in Γ or τ', to what term does $\texttt{unpack}((\texttt{pack}\,\tau'\,N), x, M', \tau')$ beta-reduce? [2 marks]

(*c*) For each PLC type τ, let $\neg\tau$ be the type $\tau \to \forall\alpha\,(\alpha)$. Give, with justification, closed PLC terms of the following types

(*i*) $\forall\alpha\,(\neg\tau) \to \neg\exists\,\alpha\,(\tau)$ [4 marks]

(*ii*) $\exists\,\alpha\,(\neg\tau) \to \neg\forall\alpha\,(\tau)$ [4 marks]

END OF PAPER