CST2
COMPUTER SCIENCE TRIPOS Part II

Tuesday 8 June 2021 11:30 to 14:30 BST

COMPUTER SCIENCE Paper 8

*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-p8-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*) Suppose you have a randomised approximation algorithm for a maximisation problem such that, for any $\epsilon > 0$ and any problem instance of size n, the algorithm returns a solution with cost C such that

$$\mathbf{Pr}[C \geq (1 - 1/\epsilon) \cdot C^*] \geq 1/n \cdot \exp(-1/\epsilon),$$

where C^* is the cost of the optimal solution. Can you use your algorithm to obtain a PTAS or FTPAS? Justify your answer. [6 marks]

(*b*) We consider the following optimisation problem. Given an undirected graph $G = (V, E)$ with non-negative edge weights $w : E \rightarrow \mathbb{R}^+$, we are looking for an assignment of vertex weights $x : V \rightarrow \mathbb{R}$ such that: (*i*) for every edge $\{u, v\} \in E$, $x(u) + x(v) \geq w(\{u, v\})$, (*ii*) $\sum_{v \in V} x(v)$ is as small as possible.

(*i*) Design a 2-approximation algorithm for this problem. Also analyse the running time and prove the upper bound on the approximation ratio. *Note:* For full marks, your algorithm should run in at most $O(E^2)$ time.

Hint: One way to solve this question is to follow the approach used by the greedy approximation algorithm for the VERTEX-COVER problem. [8 marks]

(*ii*) Can this problem be solved exactly in polynomial-time? Either describe the algorithm (including a justification of its correctness and why it is polynomial time) or prove that the problem is hard via a suitable reduction. [6 marks]

2 Bioinformatics

(*a*) Compute the nearest neighbour phylogeny from the four species (B,M,H,O) distance matrix.

$$\left(\begin{array}{ccccc} & B & M & H & O \\ B & 0 & 5 & 6 & 4 \\ M & 5 & 0 & 3 & 2 \\ H & 6 & 3 & 0 & 2 \\ O & 4 & 2 & 2 & 0 \end{array}\right)$$

[6 marks]

(*b*) Can we always build a phylogenetic tree from a distance matrix? [2 marks]

(*c*) Derive the Burrows-Wheeler (BWT) transform of the string 'TAGTATA'. How can the transform be reversed? Comment on the use of BWT for a genome sequence that has many repeated substrings. [4 marks]

(*d*) Three analysis techinques for gene expression data (microarray) are hierarchical clustering, *k*-means and Markov clustering. Describe the structure of a set of experimental results that could be analysed by all three techniques and state what each form of analysis might identify and any additional inputs required. [4 marks]

(*e*) Discuss how a Hidden Markov Model can be used to identify different gene parts and how many sequences might be needed to compute reliable transition probabilities. [4 marks]

3 Comparative Architectures

(*a*) Some VLIW processors exploit fine-grain multithreading and SIMD execution units.

(*i*) What benefits could adding support for fine-grain multithreading to a VLIW processor provide? [4 marks]

(*ii*) Why might a simple round-robin thread schedule be inefficient and how could we improve the schedule? [2 marks]

(*iii*) Assuming the VLIW processor has taken full advantage of fine-grain multithreading with a simple round-robin thread schedule, what changes to the processor might an optimised thread schedule require to ensure programs continue to execute correctly? [4 marks]

(*iv*) Why might it be useful to include SIMD functional units when a VLIW processor can already specify independent operations to be executed in parallel? [4 marks]

(*b*) Some VLIW processors support variable-length bundles of independent instructions.

(*i*) Why is this a useful feature and how could it be supported? [2 marks]

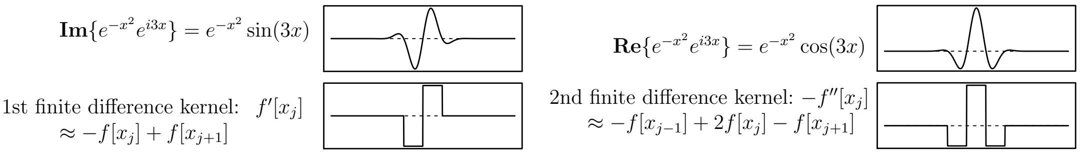
(*ii*) What costs would be incurred and additional logic needed to support this feature? [4 marks]

4 Computer Vision

(*a*) In the context of automated detection and interpretation of affective expressions using FACS (Facial Action Coding System), define the following concepts:

(*i*) facial muscle action unit (AU)
(*ii*) action descriptor (AD)
(*iii*) valence
(*iv*) arousal
(*v*) "Pan-Am smile"
(*vi*) cross-cultural universal [6 marks]

(*b*) Gabor wavelets offer a unifying framework for many tasks in computer vision, including edge detection, finding facial features, and pattern matching. The complex wavelet components (upper row) are parameterised to approximate the convolution kernels for computing 1st and 2nd derivatives (lower row, given as functions of x but rotatable in images $f(x, y)$ into functions of y as well):



$\mathbf{Im}\{e^{-x^2}e^{i3x}\} = e^{-x^2}\sin(3x)$

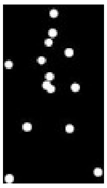
$\mathbf{Re}\{e^{-x^2}e^{i3x}\} = e^{-x^2}\cos(3x)$

1st finite difference kernel: $f'[x_j] \approx -f[x_j] + f[x_{j+1}]$

2nd finite difference kernel: $-f''[x_j] \approx -f[x_{j-1}] + 2f[x_j] - f[x_{j+1}]$

Explain how Gabor wavelets can estimate the gradient vector field $\vec{\nabla} f(x, y)$ in edge detection, extracting both edge strength and edge direction. Also describe how they can be used in a demodulation network to localise facial features. Identify one application of Gabor wavelets in pattern matching. [5 marks]

(*c*) What can we learn from the perceptual experiments of the Swedish psychologist Johansson, involving sparse dot patterns such as shown on the right? How might his findings be useful in computer vision for data fusion, integration of motion cues in object recognition, and general aspects of scene understanding?



[4 marks]

(*d*) Biological neurones are notoriously noisy, are apparently random in their connections and their firing patterns, and sluggish, with maximum firing rates around 100 Hz. Yet biological vision systems are wonderfully capable. Is there really any need for computer vision systems to use double precision arithmetic and GHz clock speeds? Give three examples of tasks in machine vision whose execution appears to require double precision arithmetic and high FLOPS, and for each example, explain this contrast with biological solutions. [5 marks]

5 Cryptography

(*a*) Consider the following two alternative definitions of a MAC function, which receives as input an $(n \cdot L)$-bit long message of the form $M = M_1 \| M_2 \| \dots \| M_L$ with $M_i \in \{0,1\}^n$ and a private key $K \in \{0,1\}^n$ picked uniformly at random, returning a tag $T \in \{0,1\}^n$. Show how neither definition provides the security property of *existential unforgeability.*

(*i*) Let F be an n-bit to n-bit pseudo-random function. Return the message tag $T = F_K(M_1) \oplus F_K(M_2) \oplus \cdots \oplus F_K(M_L)$. [4 marks]

(*ii*) Let F be a $(2n)$-bit to n-bit pseudo-random function. Return the message tag $T = F_K(\langle 1 \rangle \| M_1) \oplus F_K(\langle 2 \rangle \| M_2) \oplus \cdots \oplus F_K(\langle L \rangle \| M_L)$. [6 marks]

[*Notation:* $\|$ = concatenation of bit strings, $\oplus$ = bit-wise XOR, $\langle i \rangle$ = n-bit binary representation of non-negative integer i.]

(*b*) Your colleague proposes to construct an authenticated encryption scheme that encrypts a plain-text message M by first calculating the message authentication code $\text{CMAC}_K(M) = T$, and then forms the ciphertext by encrypting $M \| T$ using CFB mode with initial vector $IV = E_K(T)$, using the same key and blockcipher E_K. Does this construction offer CCA security? Why or why not? [5 marks]

(*c*) Given a block cipher E_K with n-bit block size, where $n \geq 64$ is a power of two, how can you use E_K to construct a strong pseudo-random permutation for $\frac{n}{2}$-bit blocks? [5 marks]

6 Denotational Semantics

A *right adjoint* of a monotone function $f : P \to Q$ between posets is a monotone function $g : Q \to P$ such that $\mathrm{id}_P \sqsubseteq g \circ f$ and $f \circ g \sqsubseteq \mathrm{id}_Q$.

Let $f : P \to Q$ be a monotone function with a right adjoint $g : Q \to P$.

(*a*) For $p \in P$ and $q \in Q$, prove that $f(p) \sqsubseteq_Q q$ if, and only if, $p \sqsubseteq_P g(q)$. [4 marks]

Let $h : P \to P$ and $\ell : Q \to Q$ be monotone functions such that $f \circ h = \ell \circ f : P \to Q$.

(*b*) Prove that if h has a least pre-fixed point $\mathit{fix}(h)$ then $f(\mathit{fix}(h))$ is a least pre-fixed point of ℓ. [8 marks]

Further assume that $g \circ f = \mathrm{id}_P$, in which case f is said to be an *embedding* and g a *projection*.

(*c*) Prove that if ℓ has a least pre-fixed point $\mathit{fix}(\ell)$ then $g(\mathit{fix}(\ell))$ is a least pre-fixed point of h. [8 marks]

7 E-Commerce

Over the years you developed a friendship with Alice, the owner-operator of a high quality coffee shop. As a result of recent government regulations in relation to managing the current pandemic, Alice has decided to close the retail aspect of her business and move it online. She turns to you for help as you are the friendly neighbourhood Computer Scientist.

(*a*) Discuss four aspects, from a legal standpoint, of moving her business online that Alice needs to consider that would be different from operating a retail location. [4 marks]

(*b*) Moving online enables Alice to consider alternative business models to the traditional e-commerce merchant model. Discuss two business models that might be relevant to her coffee business highlighting both the business activity and financial model. [6 marks]

(*c*) Choose one of the business models you identified above and discuss what activities Alice might need to use to implement the model, along with the metrics she could use to measure if she is being successful or not. [10 marks]

8 Hoare Logic and Model Checking

Consider commands C composed from assignments $X := E$ (where X is a program variable, and E is an arithmetic expression), heap dereference $X := [E]$, heap assignment $[E_1] := E_2$, the no-op `skip`, sequencing $C_1; C_2$, conditionals `if` B `then` C_1 `else` C_2 (where B is a boolean expression), and loops `while` B `do` C. `null` is 0. We write $\texttt{align}(t, s)$ for the smallest multiple of s larger than t.
Let $\mathsf{block}(t, 0) = \mathsf{emp}$, $\quad \mathsf{block}(t, n+1) = (\exists t'.\, t \mapsto t') * \mathsf{block}(t+1, n)$.

(*a*) Explain why the following postcondition for an allocator that returns aligned blocks is incorrect, and propose a fix.

```
{$\mathsf{block}(B, E-B) * 1 \leq S$}
if $\texttt{align}(B, 2^S) + 2^S < E$
then ($R := \texttt{align}(B, 2^S); B := B + 2^S$) else $R := 0$
{ $\mathsf{block}(B, E-B) *$
  $(R \neq 0 \Longrightarrow (\mathsf{block}(R, 2^S) * R = \texttt{align}(R, S)))$ }
```

[3 marks]

(*b*) With this specification, allocations cannot be chained, as in $C_{\text{alloc}}; Y := X; C_{\text{alloc}}$. Explain why, and propose a fix. [2 marks]

(*c*) Strengthen the precondition just enough to guarantee the success of allocation (so that $R \neq 0 \Longrightarrow$ is not needed anymore). [2 marks]

(*d*) Consider the following representation predicate for lists of free blocks of size 2^S:

$$\texttt{freelist}(t, S) = (t = \texttt{null} * \texttt{emp}) \vee \left(\exists t'. \begin{pmatrix} t = \texttt{align}(t, 2^S) \wedge \\ \begin{pmatrix} t \mapsto t' * \\ \texttt{block}(t+1, 2^S - 1) * \\ \texttt{freelist}(t', S) \end{pmatrix} \end{pmatrix} \right)$$

Give a loop invariant, and precisely but informally explain why it is preserved, for this "add the contents of a block into a free list" triple:

```
{$B = \texttt{align}(B, 2^S) * \mathsf{block}(B, E-B) * 1 \leq S * L = \texttt{null}$}
while $B + 2^S < E$ do
  ($[B] := L; L := B; B := B + 2^S$)
{$\mathsf{block}(B, E-B) * \texttt{freelist}(L, S)$}
```

[7 marks]

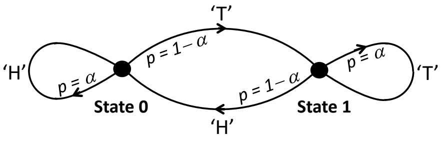
(*e*) Give a loop invariant, and precisely but informally explain why it is preserved, for this "coalesce blocks of a size S free list into a size $S+1$ free list" triple:

```
{$\texttt{freelist}(L1, S) * L2 = \texttt{null} * D = 0$}
while $D = 0$ do
  if $L1 = \texttt{null}$ then $D := 1$
  else
    $X := [L1]$;
    if $X = \texttt{null}$ or $X \bmod 2^{S+1} \neq 0$ then $D := 1$
    else
      $Y := [X]$;
      if $X + 2^S \neq Y$ then $D := 1$
      else ($[X] := L2; L2 := X$)
{$\texttt{freelist}(L1, S) * \texttt{freelist}(L2, S+1)$}
```

[6 marks]

9 Information Theory

(*a*) A "smart" coin (one with memory) is tossed, whose frequencies of coming up heads ('H') or tails ('T') are equal; but with probability α the outcomes repeat the previous one ($0 < \alpha < 1$).



(*i*) Suppose you know $\alpha = 0.75$, and you observe that a particular outcome is the opposite of the previous one. How much information, in bits, is associated with this improbable observation? [1 mark]

(*ii*) Treating α as a free parameter, provide an expression for the entropy $H(\alpha)$ of this two-state Markov process. What is the maximum possible value of $H(\alpha)$, and how is that compatible with your answer in (*i*)? [3 marks]

(*b*) Consider two discrete probability distributions $p(x)$ and $q(x)$ over the same set of four values $\{x\}$ of a random variable:

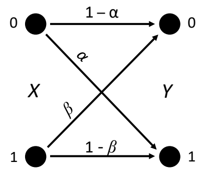
$p(x)$	1/8	1/8	1/4	1/2
$q(x)$	1/4	1/4	1/4	1/4

(*i*) Calculate the cross-entropy $H(p, q)$ between $p(x)$ and $q(x)$. [2 marks]

(*ii*) Calculate their Kullback-Leibler distance $D_{\mathrm{KL}}(p\|q)$. [2 marks]

(*iii*) Comment on the use of metrics $H(p, q)$ and $D_{\mathrm{KL}}(p\|q)$ in machine learning and for calculating the efficiency of codes. [2 marks]

(*c*) Consider an asymmetric binary channel whose input source is the alphabet $X = \{0, 1\}$ with probabilities $(0.5, 0.5)$ and whose outputs are $Y = \{0, 1\}$, but with asymmetric error probabilities. Thus an input 0 is flipped with probability α, but an input 1 is flipped with probability β.



(*i*) Give its channel matrix $p(y_k|x_j)$ and the output probabilities. [3 marks]

(*ii*) Show that the capacity C of this asymmetric binary channel is minimised, $C = 0$, for any combination (α, β) in which $\alpha + \beta = 1$. [2 marks]

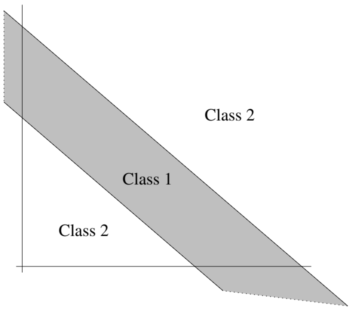
(*d*) In the Information Diagram developed by Dennis Gabor, explain the concept of an "atom" and what is irreducible about it. Draw several atoms in this plane representing different trade-offs, labelling the axes of the plane, and explain what all the atoms have in common despite their differences. Write a parameterised expression $f(t)$ defining atoms as functions of time, and explain what makes atoms an optimal basis for representing information in signals. [5 marks]

10 Machine Learning and Bayesian Inference

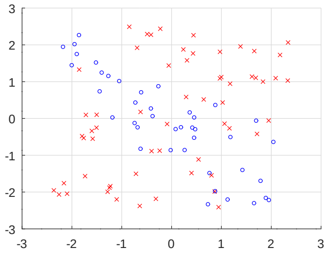
The standard linear classifier for two-class problems models the generation of noisy data as

$$\Pr(\text{Class } 1|\mathbf{w}, b; \mathbf{x}, \theta) = \sigma_\theta\left(\mathbf{w}^T\mathbf{x} + b\right) \tag{1}$$

where $\sigma_\theta(x) = 1/(1 + \exp(-\theta x))$. You are presented with a problem where the data appears not quite linear, in the sense that Class 1 forms a band, with Class 2 on either side:



The data is still noisy however, and looks like this:



We refer to this as the *banded* data.

(a) Explain the purpose of each element of (1). Explain why each parameter is needed, and how each relates to the modelling of noisy, linear data. [6 marks]

(b) Explain how the model in (1) can be modified to model the banded data, by leaving the linear part of the model unchanged, but modifying the treatment of the noise. [3 marks]

(c) Explain how the model in (1) can be modified to model the banded data, by making a minimal modification to the linear part of the model, while leaving the treatment of the noise unchanged. [5 marks]

(d) Assuming that you have training algorithms suitable for both of the models proposed for the banded data, briefly explain how, in practice, you might choose which is the better one to use, and how you might choose any relevant hyperparameters. [6 marks]

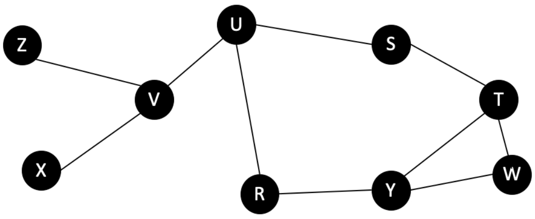
11 Mobile and Sensor Systems

(*a*) Describe the differences between proactive and reactive ad hoc routing protocols. [3 marks]

(*b*) Consider the following routing table for mobile node U in an ad-hoc network:

Dest	Next Hop	Number of Hops
X	V	2
V	V	1
Z	V	2
S	S	1
R	R	1
Y	R	2
T	S	2
W	R	3

For the mobile nodes snapshot illustrated here:



(*i*) How would the routing table for node U be updated by the Destination Sequenced Distance Vector (DSDV) routing protocol, after node Y runs out of battery and disconnects from the network? [Assume no other nodes move in that time interval]. [3 marks]

(*ii*) Now consider the scenario where node Z needs to communicate to node T regularly, but no other nodes communicate. How would you modify the routing protocol? Are there other considerations which would impact the choice of protocol to use? [4 marks]

(*iii*) Consider routing nodes STYW from the original mobile node snapshot moving out of reach of nodes ZVURX. Node S and Y are intermittently moving back in range of the latter cluster. Node Z still needs to communicate regularly with T. What protocol would you use? Illustrate the considerations that would drive your choice of protocol. [5 marks]

(*iv*) Now assume that the nodes depicted in the diagram above are fixed sensors, instead of mobile nodes. Node Y is a sink and all other nodes forward data to it. Explain why the mobile routing solutions of the previous answers are not suitable. Suggest a more appropriate protocol, describing its principles and advantages in this scenario. [5 marks]

12 Optimising Compilers

We wish to use abstract interpretation to analyse the construction, modification and traversal of directed graphs. Graph nodes are represented by the following C-like structure; the root of a graph is a pointer to a `graph_node`:

```
typedef struct graph_node {
  int value;
  struct graph_node *children;
} graph_node;
```

Graph nodes are assumed to have a maximum of two children. Exceptions (for example, caused by trying to add a third child to any node, or a search failing to find a node) cause control to transfer out of the program, and do not need to be considered further in the analysis.

(*a*) The first analysis consists of identifying whether a graph is actually a tree.

(*i*) Create a three-value abstraction for this analysis, describing abstract values and the concrete values that they represent, and why it is safe. [4 marks]

(*ii*) Define the abstract interpretation of the following concrete functions giving a brief explanation for each.

(A) Function *create_child*(g) creates a new `graph_node` and makes it a child of g, returning the new child. [2 marks]

(B) Function *add_child*(g, c) makes an existing node, c, a child of g, returning g. [3 marks]

(C) Function *remove_child*(g, c) removes node c from g's children, returning g. [2 marks]

(D) Function *dfs*(g, v) locates and returns the first node in a depth-first search starting at g that contains the value v. [2 marks]

(*b*) The second analysis consists of calculating the length of the shortest path from a node to any leaf node. Create an abstraction for this analysis and define the abstract interpretation for the four functions in Part (*a*)(*ii*). [*Hint:* consider using a tuple for your abstract values.] [7 marks]

13 Principles of Communications

(*a*) Multipath routing allows flows of packets in the Internet to be split over more than just one source-destination path. As a consequence, the fair allocation of network capacity amongst different flows may need to be re-considered. Inelastic (real time) flows are usually allocated capacity according to max-min fairness. Elastic (non real-time) flows employ end-to-end congestion control to adapt to available capacity, typically targeting proportional fairness over the long run.

Discuss how these two fairness policies operate in the presence of multipath routing. [10 marks]

(*b*) Multicast routing provides the delivery of flows of packets in the Internet from a source or set of sources, to a group or set of receivers. Routing protocols construct delivery trees rather than point-to-point paths, and routers have to replicate multicast addressed packets out each interface towards where there are recipients.

Multicast has seen little deployment outside single Internet Service Providers. Barriers include security concerns, but also the lack of a clear business model for inter-domain multicast, even if one were to extend the Border Gateway Protocol to support IP multicast routing and forwarding.

Discuss the security and inter-domain concerns, considering what information is not made visible by IP Multicast, and what information is hidden by the Border Gateway Protocol, paying particular attention to the different cases of customer-provider, and peering relationships.

[10 marks]

14 Quantum Computing

(*a*) A classical bit-flip channel has probability of error p, and a n-bit repetition code is used to suppress the error. If n is even, find the probability that a 'majority vote' decoding returns no answer. [2 marks]

(*b*) A qubit is encoded using a 3-bit repetition code. If it is known that the qubits will only ever encounter noise that can be modelled as independent, identically distributed bit-flips, with the probability of a bit flipping equal to p, then give the threshold of this code. State any assumptions made. [3 marks]

(*c*) A certain error-correction code suppresses the physical qubit error, p, to $\mathcal{O}(p^2)$ and has a threshold of 1%. For a quantum circuit with 20 gates, find the number of layers of concatenation required to achieve an overall error probability of at most 10% when:

(*i*) The gate error-rate is 0.99%.

(*ii*) The gate error-rate is 0.9%.

[5 marks]

(*d*) For a certain implementation of a 3-qubit phase-flip code the *principle of deferred measurement* is invoked to allow the recovery operations to be enacted conditional on qubit states rather than measurement outcomes. Let $|m\rangle$ be the two-qubit state of the parity check qubits, then the recovery circuit must perform the following operations on the three code qubits:

$\lvert m\rangle$	**Recovery Operations**
$\lvert 00\rangle$	$I \otimes I \otimes I$
$\lvert 10\rangle$	$Z \otimes I \otimes I$
$\lvert 11\rangle$	$I \otimes Z \otimes I$
$\lvert 01\rangle$	$I \otimes I \otimes Z$

Design the recovery circuit using only gates from the set: $\{H, T, \text{CNOT}, \text{Toffoli}\}$. [6 marks]

(*e*) How many more gates would be required if only gates from the set $\{H$, T, CNOT$\}$ can be used in the recovery circuit for Part (*d*)? [4 marks]

15 Types

(*a*) In the calculus of proofs and refutations, suppose that $\Gamma; \Delta \vdash A$ $\texttt{true}$ and $\Gamma, A; \Delta \vdash C$ $\texttt{true}$. Show that $\Gamma; \Delta \vdash C$ $\texttt{true}$ is derivable. [*Hint:* Recall that weakening is admissible in this calculus.] [8 marks]

(*b*) In System F, consider an arbitrary type A.

(*i*) Give two terms $f : A \to \forall a.\, (A \to a) \to a$ and $g : (\forall a.\, (A \to a) \to a) \to A$. [3 marks]

(*ii*) Carefully explain what this tells you about the relationship between the types A and $\forall a.\, (A \to a) \to a$. [4 marks]

(*c*) Consider the following piece of Agda code, where `Nat` is the type of natural numbers:

```
X : (P : Nat → Set) →
    P 0 →
    ((n : Nat) → P n → P (1 + n)) →
    (k : Nat) → P k
X P base step zero = base
X P base step (suc n) = step n (X P base step n)
```

(*i*) Explain what the `X` function means in logical terms. [2 marks]

(*ii*) Explain what the `X` function does in terms of functional programming. [3 marks]

END OF PAPER