UNIVERSITY OF BRISTOL

January 2019 Examination Period

FACULTY OF ENGINEERING

M Level Examination for the Degree of
Bachelor of Science / Master of Engineering / Masters of Science

COMSM0021
Neural Information Processing

TIME ALLOWED:
2 hours

Answers to COMSM0021: Neural Information Processing

Intended Learning Outcomes:

Section A: short questions - answer all questions

Q1. What is the maximum value Shannon's entropy could have for a random variable that takes eight possible values. What is its probability distribution.

Solution: A uniform distribution $p = 1/8$ for all states would be the maximum, with

$$H(X) = \log_2 8 = 3$$

with other bases allowed if clearly stated. [1 for uniform and 1 more for value]

Q2. Calculate Shannon's entropy for the variable X where $p_X(a) = 1/2$, $p_X(b) = p_X(c) = 1/4$.

Solution:

$$H(X) = \frac{1}{2} + 2\frac{1}{4}2 = 1.5$$

[1 if some awareness fo formula, 1 more for correct value]

Q3. The infomax algorithm for recorded signals

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$$

tries to find M that finds

$$\mathbf{x} = M\mathbf{r}$$

with x_1 and x_2 as independent as possible. It seeks to do this by maximizing $H(X_1, X_2)$; why is this useful?

Solution: Well obviously we need to minimize $I(X_1, X_2)$ and

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$$

so if $H(X_1) + H(X_2)$ remained fixed maximizing $H(X_1, X_2)$ would minimize $I(X_1, X_2)$. [1 mark for understanding $I(X_1, X_2) = 0$, the other for the formula for $I(X1, X_2)$]

Q4. In Bayesian statistics what is meant by a prior?

Solution: It is our prior knowledge of some variable $p_X(x)$ where we have some evidence $p_{E|X}(e|x)$ and wish to update our knowledge by calculating $p_{X|E}(x|e)$. It is sufficient to write down Bayes' rule and point out the prior:

$$p_{X|E}(x|e) = \frac{p_{E|X}(e|x)p_X(x)}{p_E(e)}$$

[1 for mixed up version, 2 for good version]

Q5. In the Eriksen flanker task sketch the accuracy versus reaction time for the consistent, **HHH**, and inconsistient, **HSH**, conditions. The overall scale of the reaction time is not what is being asked for, rather the shape.

Solution: In both cases the line starts at 0.5 but in the consistient case it rises quickly to one, in the inconsistient case it falls below 0.5 before rising more slowly to one. [1 mark for starting at 0.5 and rising to 1, the other for the dip in inconsistient].

Q6. What two pieces of evidence are fused in a Kalman filter?

Solution: The dead reckoning estimate is fused with the sensor information. [1 mark each]

Q7. Give two features specific to convolutional neural networks that are inspired by neuroscience.

Solution:

- Convolutional processing: Each neuron has a receptive field for a specific region of the input (as observed in the sensory cortices and inspired by the work of Hubel and Wiesel).
- Connectivity: Hierarchical processing/receptive fields and multiple stages of pooling/connectivity.
- Pooling: Max pooling as often used in convnets can be implemented in neuroscience with a winner-take-all with lateral inhibition acting as a competition mechanism. There is some evidence for this (Riesenhuber and Poggio 1999).

Q8. What is the key difference between supervised and unsupervised learning in terms of their cost function?

Solution: The existence of a teacher (or teaching signal) in supervised learning, but not in unsupervised learning [2 marks].

Q9. Give an example of a gated recurrent neural networks (GRNN).

Solution: Long short-term memory network (LSTM) or Gated recurrent unit (GRU), etc.

Q10. Give the value update function of temporal difference (TD) learning.

Solution: $\underbrace{V(S_t)}_{\text{value}} = V(S_t) + (\overbrace{\underbrace{R_{t+1}}_{\text{reward}} + \lambda \underbrace{V(S_{t+1})}_{\text{future value}}}^{\text{learned value}} - V(S_t))$

Q11. What is the role of the Dopamine neuromodulation system for reinforcement learning in the brain?

Solution: Dopaminergic neuromodulation has been proposed to signal reward prediction error.

Q12. The pairwise maximum entropy model describes the probability distribution across neural population activity patterns as $P(\mathbf{x}) = \frac{1}{Z} \exp\left[\sum_i h_i x_i + \sum_{i \neq j} \frac{1}{2} J_{ij} x_i x_j\right]$ where $\mathbf{x}$ is the vector of binary neural activities and Z is a normalising constant to make sure $P(\mathbf{x})$ sums to one. What do the parameters h_i and J_{ij} represent?

Solution: h_i are the individual neuron biases. J_{ij} are the couplings between pairs of neurons. [1 mark for identifying each parameter.]

Q13. Draw a diagram of the structure of a Restricted Boltzmann Machine labeling the components.

Solution: Bipartite graph of hidden and visible units with no within-class connections. [1 mark for any bipartite graph with hidden and visible units named, 1 more mark for correctly restricted connections.]

Q14. What is the neural manifold hypothesis?

Solution: The idea that neural population activity is much lower dimensional than the number of neurons. [2 marks for reasonable answer. They should specifically evoke the concept of "low dimensional".]

Q15. Name two limitations of Principal Component Analysis for analysing neural population data.

Solution: Up to any two of

- Assumes linearity, linear subspace, etc.
- Is sensitive only to pairwise correlations, not higher-order.

- Is insensitive to temporal ordering of data/dynamics.
- Assumes gaussian noise / minimises squared error
- Ignores heterogeneity in mean neural firing rates (although listens to variances)

mark for each with a maximum of two marks.

Section B: long questions - answer two questions

Q1. This question is about sensory fusion.

(a) If we have a Markov chain

$$V \rightarrow X \rightarrow H$$

what can we say about $p_{V,H|X}(v, h|x)$? *[4 marks]*

(b) In the Ernst and Banks experiment participants are asked to assess the height of a block using visual and haptic input. If the visual estimate is v and the haptic estimate is h what is meant by the maximum likelihood estimate of the actual height x? *[4 marks]*

(c) Assumption that the visual and haptic estimates are conditionally independent, conditioned on the true value, and normally distributed about the true value with variances σ_v^2 and σ_h^2 respectively. Show the mean of $p(x|v, h)$ is

$$\mu = \frac{\sigma^2}{\sigma_v^2} v + \frac{\sigma^2}{\sigma_h^2} h$$

where

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_v^2} + \frac{1}{\sigma_h^2}$$

[8 marks]

(d) In David and Burr (2004) there is a discussion of the ventriloquist effect, whereby we perceive sound as coming from a visually cued location. An example is when we watch a film and perceive a voice as coming from the person who is speaking on screen rather than the audio speaker located elsewhere in the cinema. In their experiment David and Burr present a sound and, visually, a Gaussian blob somewhere along a line on a screen. If σ_s^2 is the variance in our perception of the location of a sound source and σ_b^2 is the variance of the Gaussian blob, where would you speculate, from a Bayesian point-of-view that the participants perceive the sound as coming from? *[4 marks]*

Solution: a)

$$p_{V,H|X}(v, h|x) = p_{V|X}(v|x)p_{H|X}(h|x)$$

[4 for correct, 2 for some attempt]
b)
It is the x that gives the maximum value of $p(x|v, h)$. [4 for correct, 2 for some attempt]
c)
So this is the classic sensory fusion calculation, with v and h conditionally independent and normal, $p(x|v, h)$ is normal and you can calculate the mean and variance by multiplying out the densities and messing around with the exponent. [3 marks for multiplication of two gaussians, 5 for calculating the mean, 2 for an attempt.]]
d) If b is the location of the blob and s of the sound then the perceived location is x above with (s, b) replacing (v, h). [4 for correct]

Q2. This question is about backpropagation in the brain.

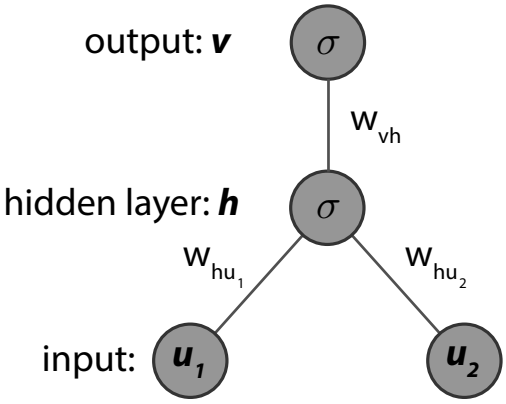

Figure 1: **Schematic of simple feedfoward neural network, with sigmoidal units $\sigma(x)$.**

(a) In what way does the backpropagation algorithm solve the credit assignment problem? *[5 marks]*

(b) Explain three key features that have been suggested to make the backpropagation algorithm used in supervised learning biologically implausible? You should use a simple two layer neural network (with one hidden neuron h, one output neuron v, two input weights, and without biases, see Figure 1) to derive the weight updates and make a schematic of the network with the backprop to help illustrate your answer. Assume the cost function (or error) to be $E = (v - y)^2$, where y is the desired target. *[11 marks]*

(c) Which biologically implausible feature of backprop does *feedback alignment* address? And how does *feedback alignment* address it? *[4 marks]*

Solution:

a) The backpropagation algorithm provides an exact solution for how to adjust (i.e. assign credit) network parameters (e.g. weights and biases) to improve the behaviour (or output) of the network given a desired cost function [3 marks]. It relies on the of use of the differentiation chain rule [2 marks].
b) To illustrate the backprop algorithm we provide the update rule for weight w_{hu_1} (similar for w_{hu_2}) is given by $\frac{\partial E}{\partial w_{hu_1}} = \frac{\partial E}{\partial v}\frac{\partial v}{\partial h}\frac{\partial h}{\partial w_{hu_1}}$ [1 mark]. For the network considered here, these components are [1 mark for each]:

$$\frac{\partial E}{\partial v} = 2(v - y) \tag{1}$$

$$\frac{\partial v}{\partial h} = \sigma'_v w_{vh} \tag{2}$$

$$\frac{\partial h}{\partial w_{hu_1}} = \sigma'_h u_1 \tag{3}$$

which together yields $\frac{\partial E}{\partial w_{hu_1}} = 2(v - y)\sigma'_v w_{vh}\sigma'_h u_1$, where $\sigma' = \sigma(x)(1 - \sigma(x))$ denotes the derivative of the sigmoid activation function. The backwards flow of the error (in orange) is represented in Figure **??** [2 marks].

Three of the following features can be indicated (others related to these ones may be also considered) [2 marks for each]

- *Weight transport problem*: the existence os symmetric weights (i.e. feedback weights derived from backprop are equal to feedforward weights, e.g. w_{vh} (as in Eq. 1). Synaptic weights are unidirectional in the brain, which makes symmetric feedforward and feedback weights implausible.

- *Derivative of activation (input-output) function*: Backprop needs the derivative of the neuronal activation function (as in Eqs. 2 and 3), which is unclear how it is computed biologically.

- *Target or teaching signal*, unclear whether the brain could provide such a teaching signal needed for the cost/error function. However, in principle, different brain areas encoding different aspects of the environment can act as teachers of other brain areas, or such teaching signal could be provided by an external teacher.

- *Weight learning rule is non-local*: synaptic plasticity (i.e. modification of synaptic weights) typically depends only on locally available information (e.g. pre and postsynaptic activity). However, classical backprop relies on non-local information (error signal, as in $\frac{dE}{dw_{hu_1}}$).

- *Separate learning phase*: Learning needs its only separate phase. Backprop relies on first having a forward phase to update activity and then a backward phase to calculate the error and update the neural network parameters. There is no evidence of such a clear separation between activity and learning in neuroscience.

c) Lillicrap et al. 2016 showed that the symmetric feedforward and feedback weights of backprop are not essential for backprop to work [2 marks]. Replacing the feedback symmetric by random weights works well in practice [2 marks].

Q3. This question is partly about sparse coding and partly about multiunit coding.

(a) What types of features are learned by sparse coding algorithms when applied to natural images? *[4 marks]*

(b) Give the classical cost function used in sparse coding and describe its components. *[4 marks]*

(c) Imagine we have recorded the spiking patterns of a population of three neurons at eight sequential points in time. We can represent the data in a 3×8 matrix D as

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

where each row corresponds to the activity time series for one of the three neurons. Write down the time-averaged probability that a neuron is ON p_i for each of the three neurons. *[4 marks]*

(d) Use the same matrix D as in the previous question. Assume a model where each neuron is independent so that the probability of a pattern $\mathbf{x} = x_1, x_2, x_3$ is $p(\mathbf{x}) = \prod_i [p_i x_i + (1 - p_i)(1 - x_i)]$. This model can be used to compute the probability of any activity pattern, even if it was not observed in the original data. Compute the probability of the pattern $0, 1, 1$. *[4 marks]*

(e) Referring to the model in the previous question, the three neurons can jointly make $2^3 = 8$ possible binary patterns. Which of the eight patterns is/are the most probable under the independent model? *[4 marks]*

Solution: Answers
a) Features representing statistics of natural images [2 mark], such as oriented bars/edges [2 mark].
b) cost = $(v - u)^2-$, where v is the activity of output neurons and u is the input activity of input neurons. [3 for formula, 1 for definitions of u and v]
c) $p(x_1) = 1/2$, $p(x_2) = 1/4$ and $p(x_3) = 1/8$. [4 marks, one for general idea, one for each neuron correct]
d) $1/2 \times 1/4 \times 1/8 = 1/64$ [4 marks for correct answer. Do not penalise for any mistakes made in part (a)]
e) The two patterns $0, 0, 0$ and $1, 0, 0$ are the joint most probable. [4 marks total. 2 marks if only identifying one pattern.]

END OF PAPER