ELV832: Special Module in Machine Learning

Major Test, Maximum marks: 36

1. Recall that in the original rate-distortion formulation, with a pre-determined distortion function $d(\mathbf{x}, \tilde{\mathbf{x}})$, the rate minimisation depends on the choice of the set of codewords $\tilde{X}$. To do a joint inference on the codewords and the optimal mapping from inputs to codewords, we set up an EM-like iteration between the following two steps:

   - For a fixed set of codewords $\tilde{X} = \{\tilde{\mathbf{x}}\}$, use the Blahut-Arimoto algorithm to find the mapping $p(\tilde{\mathbf{x}}|\mathbf{x})$ which minimises the rate for a given choice of β.
   - For a fixed mapping to a set of codewords $p(\tilde{\mathbf{x}}|\mathbf{x})$ (but where the actual value of each codeword $\tilde{\mathbf{x}}$ is now treated as a variable), choose the actual set of codewords $\tilde{X} = \{\tilde{\mathbf{x}}\}$ so as to minimise the expected distortion $\langle d(\mathbf{x}, \tilde{\mathbf{x}}) \rangle_{p(\tilde{\mathbf{x}}, \mathbf{x})}$.

   In this regard, answer the following:

   (a) The first step gives us $p(\tilde{\mathbf{x}}|\mathbf{x})$, but the calculation of the expectation in the second step requires $p(\tilde{\mathbf{x}}, \mathbf{x})$. How do we obtain the latter from the former? **[1.5]**

   (b) Recall the K-means algorithm, which also iterates between two steps: given some cluster centres, assigning each point to the nearest one; and then given the assignments, updating the cluster centres to the mean of all points assigned to that cluster. Do these two steps correspond in some way to the two steps mentioned above? Explain how. **[4]**

   (c) When we set up the iterative algorithm (similar to Blahut-Arimoto) for the information bottleneck method, we said that the selection of the actual codewords $\{\tilde{\mathbf{x}}\}$ *no longer remains an independent problem*, and that the optimisation over these codeword choices now effectively happens within the Blahut-Arimoto iteration itself (which just corresponds to the first step above). Explain why this is different from the above rate-distortion setting. **[4]**

2. In the rate-distortion framework, the basic optimisation problem was forumated as follows:

   $$R(D) = \min_{\langle d(\mathbf{x}, \tilde{\mathbf{x}}) \rangle \leq D} I(X; \tilde{X}). \tag{1}$$

   Where in the information bottleneck framework, we had:

   $$R'(M) = \min_{I(\tilde{X}; Y) \geq M} I(X; \tilde{X}). \tag{2}$$

   In the second case, we know that the *effective distortion* emerges as:

   $$d_{IB}(\mathbf{x}, \tilde{\mathbf{x}}) = D_{KL} \left(p(y|\mathbf{x}) || p(y|\tilde{\mathbf{x}}) \right). \tag{3}$$

   Given that the expectation of the above can be written as $\langle d_{IB}(\mathbf{x}, \tilde{\mathbf{x}}) \rangle = I(X; Y) - I(\tilde{X}; Y)$, use this to convert the information bottleneck formulation (2) into an equivalent rate-distortion formulation of the form (1). What is the corresponding value of the distortion bound D? **[5]**

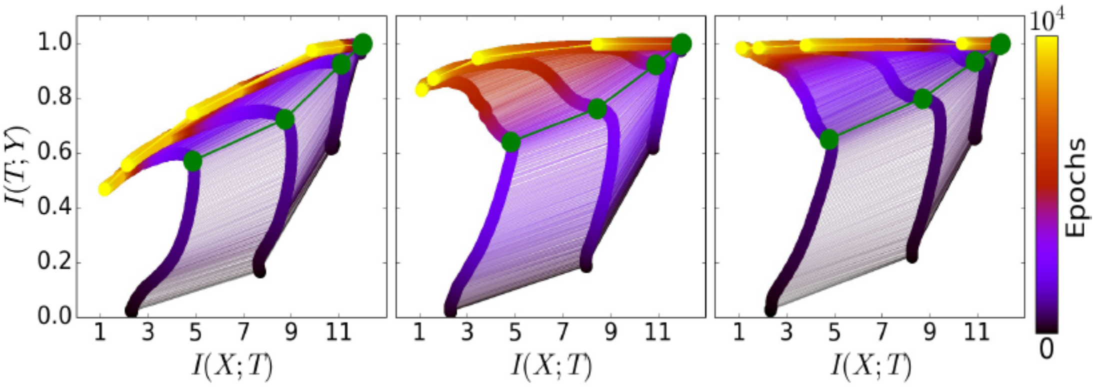
3. Consider mappings of the form $y = f(x_1, x_2)$, where (x_1, x_2) is sampled from a random variable X with range $\{(0,0), (0,1), (1,0), (1,1)\}$, and y can be thought of as a sample from a random variable Y with range $\{0, 1\}$. Suppose that the distribution over Y is known to be uniform.

   (a) If $f(x_1, x_2) = x_1$ AND x_2, what is $I(X; Y)$? **[1.5]**

   (b) If $f(x_1, x_2) = x_1$ XOR x_2, what is $I(X; Y)$? **[1.5]**

   (c) What do these values tell us about the relation between $I(X; Y)$ and the complexity of the mapping from X to Y? Under what conditions does the mutual information become informative of this complexity? Give a simple example to illustrate such a setting. **[6]**

4. Consider the below figure from Schwartz-Ziv & Tishby (2017).



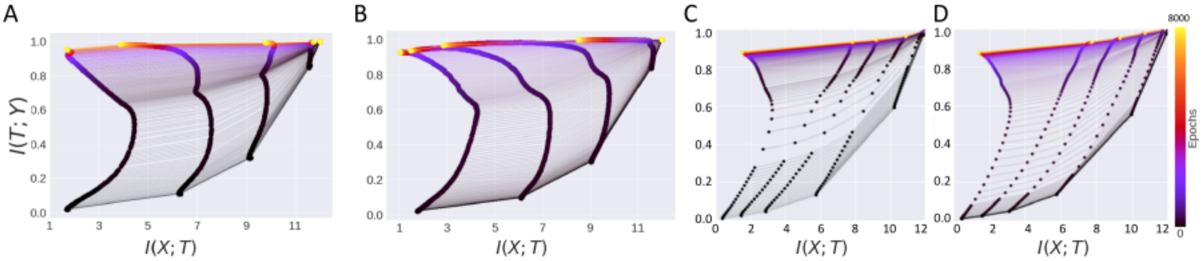
Here the plots correspond to 5%, 45%, and 85% of the data set respectively being used for training; however the mutual information estimates are over the entire data set in all cases.

(a) In the left plot, and to a lesser extent in the middle one, there appear to be some epochs towards the end where both $I(X;T)$ and $I(Y;T)$ are decreasing. Given that the neural network is supposed to be learning a mapping from X to Y, it seems strange that the training process should be throwing away information about both in the intermediate representations. Why do you think this is happening in these cases? **[2.5]**

(b) Draw a rough sketch of what you think the left plot would look like (explaining your thinking), if the mutual information estimates were based *only on the training data*. Highlight the final positions of the layers in your plot. **[2]**

(c) Draw a rough sketch of what you think the middle plot would look like (explaining your thinking), if the mutual information estimates were based *only on the testing data*. Highlight the final positions of the layers in your plot. **[2]**

5. Consider the below figure from Saxe *et al.* (2018).



The first two plots are for *tanh* networks, and the last two for *ReLU* networks. The first and third are for SGD training, whereas the second and fourth use BGD (gradients calculated on the entire training set).

The results depicted in this figure undermine or weaken two key claims of Schwartz-Ziv & Tishby (2017). Which ones, and why? **[6]**