AIS5103 homework 4 (due 13 Nov 2026)

1. Protein folding is one of the most difficult problems in computational biology. Recent advances in deep learning have led to significant progress in protein structure prediction. The protein folding problem has largely been solved in terms of structure prediction. Read the 2021 Nature paper, vol. 596, page 583, entitled "Highly accurate protein structure prediction with AlphaFold," from the Google DeepMind team. Write a summary. Focus on the transformer-like architecture known as Evoformer. Discuss the multiple sequence alignment (MSA) representation, pair representation, and attention operations. Limit your summary to half to three-quarters of an A4 page.

2. In this programming problem, we implement an Ising model configuration generating network using the transformer architecture (using the PyTorch transformer implementation). Consider first the one-dimensional Ising model, infinitely long, or long enough that boundary effects can be ignored. The joint probability of a configuration is given by the Boltzmann distribution $P(s) = \frac{1}{Z}\exp(-\frac{E(s)}{kT})$, where the energy function is given by $E(s) = -J\sum_i s_i s_{i+1}$, where $s = \{\cdots, s_i, s_{i+1}, s_{i+2,} \cdots\}$ is the set of spin configurations. The Ising spin takes the value $+1$ and -1 only. Z is the normalization constant called the partition function. For simplicity, we can choose the constant $J = 1$. $kT = 1$ for numerical calculation. Our problem is to predict the next spin value s_{i+1}, given the previous values of $\cdots, s_{i-1}, s_i$, as a semi-infinite sequence. You can say this is the language model for the 1D Ising spin. You can generate the training data by Monte Carlo sampling of the Ising model, for example, using the more efficient Swendsen-Wang algorithm.
    a. Design the decoder-only transformer network, consider how many heads are a good choice, and what length of input sequence is a good size.
    b. Train your network to learn the correct conditional probability of the next spin value, using cross-entropy as the training loss.
    c. Evaluate the performance of your trained network by comparing its predicted conditional probabilities with the exact results for the one-dimensional Ising model. Discuss how the input sequence length affects the prediction accuracy.
    d. If the 1D Ising is changed to 2D model with a strip $N \times \infty$ ($N = 1$ for the 1D model), discuss how you would modify the network architecture, training data, and evaluation procedure in parts (a)–(c). No implementation is required.

3. Bishop 20.14, page 605. The score function is defined by $\mathbf{s}(\mathbf{x}) = \nabla_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{w})$ and is therefore a vector of the same dimensionality as the input vector $\mathbf{x}$. Consider a matrix whose elements are given by

$$M_{ij} = \frac{\partial s_i}{\partial x_j} - \frac{\partial s_j}{\partial x_i}.$$

Show that if the score function is defined by taking the gradient $\mathbf{s} = \nabla_{\mathbf{x}} \phi(\mathbf{x})$ of the output of a neural network with a single output variable $\phi(\mathbf{x})$, then all the matrix elements $M_{ij} = 0$ for all pairs i, j. Note that if the score function $s(\mathbf{x}) = \nabla_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{w})$ is instead represented directly by a deep neural network with the same number of outputs as inputs, then only the diagonal matrix elements $M_{ii} = 0$, and so the output of the network does not in general correspond to the gradient of any scalar function.

4. Bishop 20.3, page 603 (equation numbers mentioned below refer to Bishop's book.) In this exercise, we use proof by induction to show that the marginal distribution of $\mathbf{z}_t$ for the forward process of the diffusion model, as defined by (20.4), is given by (20.6), where α_t is defined by (20.7). First, verify that (20.6) holds when $t = 1$. Now assume that (20.6) is true for some particular value of t and derive the corresponding result for the value $t + 1$. To do this, it is easiest to write the forward process using the representation (20.3) and to make use of the result (3.212), which shows that the sum of two independent Gaussian random variables is itself a Gaussian in which the means and covariances are additive.

- The end -