STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7042
EXAMINATION
16 Mar 2026

Exam in Statistical Deep Learning
16 Mar 2026, time 8:00-11:00

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the exam, you may use any book or note. However, **Electronic devices are NOT allowed**

NOTE: The exam consists of 3 problems with 100 points in total. Logical explanation and steps leading to the final solution must be clearly shown in order to receive full marks.

NOTE: Your answers and explanations must be to the point, **redundant writing irrelevant to the solution will result in point deduction**.

Problem 1 (Basic Concepts Total 34p)

a) Suppose a feedforward neural network is trained to infer the parameters of a one-dimensional conditional mixture model

$$p(y \mid x) = \sum_{i=1}^{m} p^{(i)}(x)\lambda^{(i)}(x)e^{-\lambda^{(i)}(x)y}, \tag{1}$$

where the number of components m is known, $\lambda^{(i)}(x) > 0$ is the rate parameter of the ith exponential component at point x and $p^{(i)}(x)$ is the mixture probability of component i given x. Design a suitable output layer of the network and motivate the choice. **(5p)** Write down the corresponding loss function? **(5p)**

b) We have discussed the intuition why a "Barrel" form of NN shape is used for classification (i.e., expand dimensions to disentangle the classes and then compress dimensions irrelevant to classification). Provide an intuition of using a "Barrel" form for regression. Please draw a schematic to help your explanation. **(6p)**

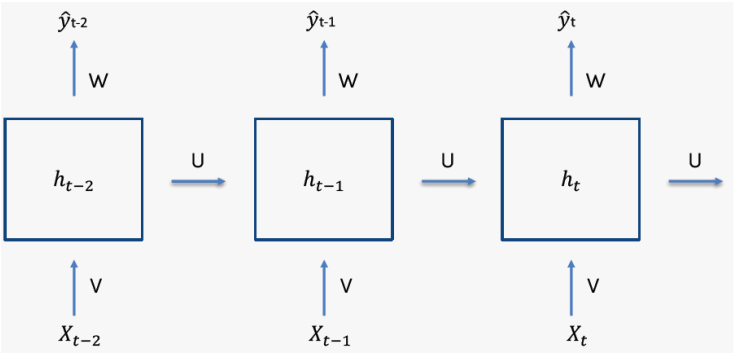
The following parts refer to the "Supplementary reading for Dropout". NOTE: In these parts, p is the probability of setting the neuron variable to zero.

c) Explain in terms of the concept of SGD how the training procedure in the section "2.1 Detailed workflow of dropout" works to minimize the dropout loss $\sum_\mu p(\mu)J(Y, X, \theta, \mu)$, where $\mu \sim p(\mu)$ is the mask vector, $p(\mu)$ is the Bernoulli distribution, and $J(Y, X, \theta, \mu)$ is the usual cross entropy loss. **(10p)**

d) It says in the note that "$\cdots$ it is imperative to rescale the vector $y_1, y_2, \cdots, y_{1000}$ by multiplying it with $1/(1-p)\cdots$", explain why this rescaling is needed. **(8p)**

Problem 2 (CNN, RNN and Attentions Total 38p)

a) Is convolution reflective equivariant, i.e., equivariant under the transformation $i \to -i$ and/or $j \to -j$ for pixels in a 2D image? Justify your answer. **(8p)**



b) Consider the recurrent neural network in the figure above where the matrices W, U and V are defined in Eq. 2, and $g_y(\cdot)$ and $g_h(\cdot)$ represent the activation functions. Write down with CLEAR STEPS the back propagation through time derivative $\partial L_t/\partial V$. **(15p)**

$$\begin{aligned} \text{Outputs} \quad & \hat{y}_t = g_y(Wh_t + b) \\ \text{Hidden units} \quad & h_t = g_h(Vx_t + Uh_{t-1} + b') \\ \text{Loss function} \quad & L = \sum_t L_t(\hat{y}_t) \end{aligned} \tag{2}$$

c) From the expression in part b, discuss if the problems of vanishing/exploding gradient and learning long time dependence exist for $\partial L_t/\partial V$. **(5p)**

d) Let $X \in \mathbb{R}^{N\times D}$ be the input to a multi-head self attention with h heads and with the scaled dot-product attention score function. Here N is the number of tokens in the sequence and D is the embedding dimension. Derive the explicit equations for the output $Y \in \mathbb{R}^{N\times D}$ of the multi-head self attention in terms of the input and the appropriate trainable weight matrices. Note: No need to include the position-wise FF part of the attention block. **(10p)**

Problem 3 (Short Unseen Questions Total 28p)

a) In the training stage, the sample mean and variance ($\hat{\mu}$ and $\hat{\sigma}$) in the Batch Normalization (BN) are estimated using the minibatch samples of the current optimization step. In the validation stage, how are the sample mean and variance ($\hat{\mu}$ and $\hat{\sigma}$) of the BN estimated? **(6p)**

b) Given the following operations: swish activation, batch normalization and FiLM (for conditioning), what is the correct order to apply these three operations consecutively? Justify your answer. **(8p)**

c) A CNN builds a "big picture" understanding of an image hierarchically, layer by layer. A Vision Transformer (ViT), however, has the potential to see the whole image in its very first self attention block. Draw a schematic picture of the self attention block with only one head **(4p)**, then indicate where in the block that ViT could integrate information within each image patch **(3p)** and across different image patches. **(3p)**

d) Describe the difference between a "many-to-one" and a "many-to-many" RNN architecture. Provide a real-world use case for each. **(4p)**

Good Luck!