STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7042
EXAMINATION
13 Mar 2023

Exam in Statistical Deep Learning
13 Mar 2023, time 14:00-19:00

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

NOTE: The exam consists of 4 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 (Feedforward neural networks, total 25p)

a) Consider a feedforward neural network for a K class outcome with the cross entropy cost function. Show that if no nonlinear hidden layers are added, this is the same as a multinomal logistic regression model. **(6p)** Show that the decision boundary of the logistic regression is linear. **(4p)**

b) Compute expressions for the gradient of the weights and biases of the feedforward neural network used in the XOR example introduced in Sec 6.1 of the course book, with respect to the MSE cost function. **(8p)** When training the network using full-batch gradient descent, describe the computations of the forward pass and backward pass parts of the training. **(2p)**

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

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

where the number of components N 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)**

Problem 2 (Regularization, total 19p)

a) State one advantage and disadvantage for each of the early stopping and L^2 regularization techniques, and justify your answers. **(5p)**

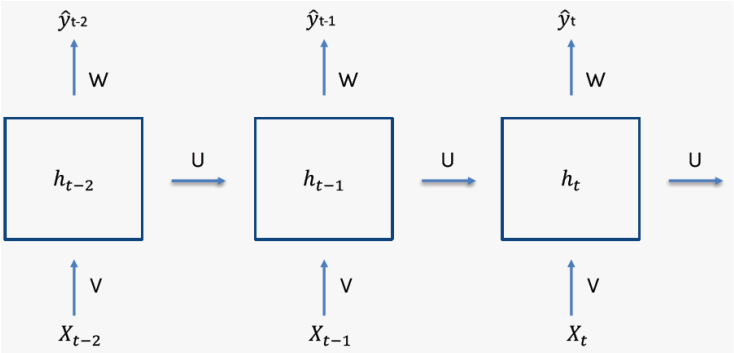
b) Under what assumptions on the cost function $J(\theta)$ and under what condition on the learning rate ϵ, number of epochs τ and norm penalty parameter α is early stopping equivalent to L^2 regularization? **(4p)**

c) In Eq. 7.54-55 of the course book, it was proposed that the geometric mean can be used to compute the ensemble prediction $p_{\text{ensemble}}(y|x)$. Show that $p_{\text{ensemble}}(y|x) = \text{softmax}\left(\frac{1}{N}\sum_{i=1}^{N} \ln p^{(i)}(y|x)\right)$, where N is the number of models and $p^{(i)}(y|x)$ is the prediction from the ith model. **(6p)** Discuss the advantages and disadvantages of using the geometric mean compared to the arithmetic mean for ensemble prediction. **(4p)**

Problem 3 (Optimization, RNN and back propagation thru time, total 31p)

a) In the Adam algorithm (Algorithm 8.7 in course book), the accumulated 1st and 2nd moments are normalized ($\hat{s} \leftarrow s/(1-\rho_1^t)$ and $\hat{r} \leftarrow r/(1-\rho_2^t)$). Explain concisely what is the purpose of these normalizations. **(8p)**

b) Referring to Fig. 10.13 in the course book, draw the unfolding graphs for the RNNs in the panels (b) and (c). **(2p)** Explain what are the benefits to introduce the additional hidden states and links that are absent in the "Vanilla" RNN in panel (a). **(4p)**



c) 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 $\partial L_t/\partial V$. **(12p)**

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

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

Problem 4 (Attentions and transformers, total 25p)

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, where n is the number of tokens in the sequence and d is the dimension of the embedding vector.

a) Write down the explicit expression 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. **(7p)** NOTE: Please clearly define your weight matrices.

b) What is the total number of trainable weights? **(4p)**

c) Based on the results in part a and b, discuss if these trainable weights are independent and propose a way to reduce the number of trainable weights. Justify your answers **(6p)**

d) Referring to Eq. 11.6.3 in the "Dive into deep learning" book, show that the matrix relating the vectors $(p_{i,2j}, p_{i,2j+1})$ and $(p_{i+\delta,2j}, p_{i+\delta,2j+1})$ is an orthogonal matrix. **(5p)** What does Eq. 11.6.3 tell us about the relative positional information by using the positional encoding in Eq. 11.6.2? **(3p)**

Good Luck!