STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7042
EXAMINATION
10 Jan 2025

Exam in Statistical Deep Learning
10 Jan 2025, time 14:00-17:00

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 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 (Total 25p)

a) Argue that the number of parameters (weights and bias) in a neural network is not a good measure of model complexity. Specifically, give an example to support or illustrate your reasoning. **(5p)**

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. **(10p)**

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

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

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)**

Problem 2 (Total 40p)

a) Consider modifying the cross entropy loss by introducing a prior $p(\theta)$ on the weights θ, so that the loss is given by $-\log p(y|x,\theta)p(\theta)$, thereby maximizing a Bayesian posterior distribution instead of the likelihood. Show that a Gaussian prior gives rise to L_2 regularization and a Laplace prior to L_1 regularization. **(5p)**

b) In the Adam algorithm (Algorithm 8.7), the accumulated 1st and 2nd moments are normalized ($\hat{s} \leftarrow s/(1-\rho_1^t)$ and $\hat{r} \leftarrow r/(1-\rho_1^t)$). Explain what is the purpose of this normalization. **(10p)**

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.

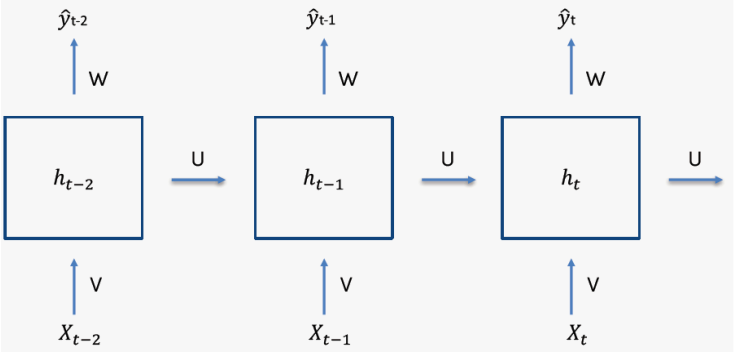
c) Explain in terms of the concept of SGD how the training procedure in the section "2.1 Detailed workflow of dropout" works to minimize the dropout loss $\sum_\mu p(\mu)J(Y, X, \theta, \mu)$, where $\mu \sim p(\mu)$ is the mask vector, $p(\mu)$ is the Bernoulli distribution, and $J(Y, X, \theta, \mu)$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

e) In Fig. 5b of the note, it says that during the test phase, it is the weights, instead of the neurons during the training phase (see Fig. 5a), that are dropped out. Explain why this is done so. **(7p)**

Problem 3 (Total 35p)

a) Show that convolution is translational equivariant **(5p)**. Is convolution reflective equivariant, i.e., equivariant under the transformation $i \to -i$ and/or $j \to -j$ for pixels in a 2D image? **(5p)**



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) From Algorithm 1 in the paper "Diffusion Models for Generative Artificial Intelligence: An Introduction for Applied Mathematicians" by Higham *et al.*, what is the loss function for the training of the neural network with parameters θ? What are the predictor and response variables? **(5p)**

Good Luck!