STOCKHOLM UNIVERSITY	MT 7042
DEPT. OF MATHEMATICS	EXAMINATION
Div. of Mathematical statistics	21 Mar 2022

Exam in Statistical Deep Learning
21 Mar 2022, time 14:00-19:00

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the home exam, you may use any literature. 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 34p)

a) Suppose a feedforward neural network is used for classification of N classes, where the softmax output function is used together with the cross entropy cost function. Show that the softmax function saturates only when the classification is correct. **(10p)**

b) When using the softmax output function, the neural network outputs N classification probabilities. We can however save some computational cost by removing 1 output unit, making use of the fact that the probabilities must sum to 1. How would you modify the output layer to achieve this? **(4p)** Show that this modification preserves the saturation property in part a. **(4p)**

c) Suppose that $p_{\text{model}}(y \mid x) = \lambda(x)e^{\lambda(x)y}$ for $\lambda(x) > 0$. We want to train a neural network $\hat{\lambda}(x;\theta)$ to infer $\lambda(x)$. What would be the form of the cost function if we use the cross entropy loss? **(4p)** What would be a suitable output function in this case? Motivate your answer. **(4p)**

d) Provide a choice of activation function, output function, loss function and model that would make a deep feedforward neural network $f(x;\theta)$ equivalent to an ordinary least squares regression. **(4p)** Show that for these design choices the network $f(x;\theta)$ is a linear function in the input x. **(4p)**

Problem 2 (Regularization, total 24p)

a) Consider modifying the cross entropy loss by introducing a prior $p(\theta)$ on the parameters θ of the neural network, so that the cross entropy 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 **(4p)** and a Laplace prior to L^1 regularization. **(4p)**

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? **(6p)**

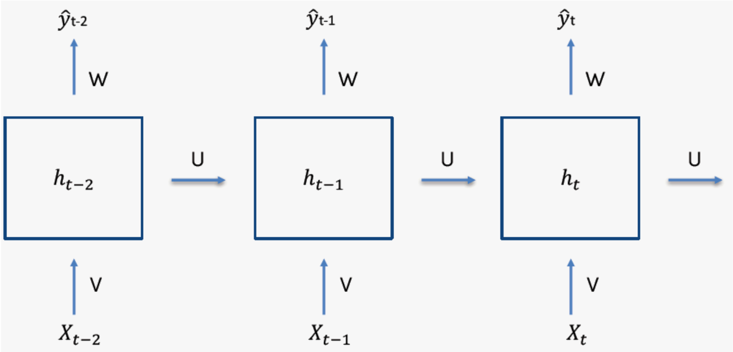
c) Let the number of hidden units be the same for a shallow (with only one wide hidden layer) and deep (with many hidden layers) feedforward neural network, give an example to demonstrate which of the two has more parameters (weights and bias). **(4p)** Reason with the help of example that the number of parameters of a feedforward network is NOT a good measure of model complexity. **(6p)**

Problem 3 (Optimization and back propagation, total 26p)

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) Consider the recurrent neural network in the figure below where the matrices W,U and V are defined in Eq. 1, 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)** From the expression, discuss if the problems of vanishing/exploding gradient and learning long time dependence exist for $\partial L_t / \partial V$. **(6p)**

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



Problem 4 (Autoencoder and variational autoencoder (VAE), total 16p)

a) Derive Eq. 11 in the 2018 paper “A practical tutorial on autoencoders for nonlinear feature fusion” by Charte *et al.* **(6p)**

b) Consider the loss function of VAE (Eq. 23 in the 2021 paper “An introduction to deep generative modeling” by Ruthotto *et al.*), Explain why one cannot drop the prior term $\log p_Z(z)$ that does not depend on the parameters Ψ and θ explicitly. **(4p)**

c) There is one hyperparameter σ that needs to be fixed by us in the likelihood (or decoder) $p_\theta(x|z)$ (see Eq. 3 of Ruthotto paper). What are the problems if σ is too small or too big? **(3p)** Propose with reasoning how this hyperparameter can be chosen. **(3p)**

Good Luck!