VIR 2019 **Name:** ____________________

Exam test

Variant: A Points ____________________

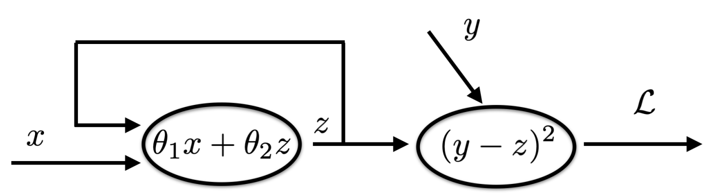
---

1. **ML regression**: You are given probability distribution model $p(y|x, w) = xw \exp(-xwy)$, which models probability of variable $y \in \mathbb{R}^+$, given measurement $x \in \mathbb{R}$ and unknown model parameters $w \in \mathbb{R}$. You are given a training set $\mathcal{D} = \{(x_1, y_1) \dots (x_N, y_N)\}$. Write down the optimization problem, which corresponds to the maximum likelihood estimate of the model parameters w? Simplify resulting optimization problem if possible.

2. **Recurrent network:** Consider linear recurrent neural network with L2 loss depicted on the image below. The network is initialized with parameters $\theta_1 = 1, \theta_2 = -1, z_0 = 1$. You are given the following training sequence:

time=1	time=2
$x_1 = -2$	$x_2 = 1$
$y_1 = 1$	$y_2 = 3$

Estimate gradient of the overall loss (computed over all available outputs y_i for both available times $i = 1, 2$) with respect to θ_1.



Hint: Unroll the network in time, to obtain a usual feedforward network with two loss nodes. Do the backpropagation as usual.

3. **Batch-norm layer:** You are given batch of two one-dimensional training examples $x_1 = 2, x_2 = 4$. The batch-norm layer has two learnable parameters $\gamma = 6$, $\beta = -1$. Compute jacobian of the batch-norm layer with respect to its parameters γ, β. Does the gradient depends on current values of γ, β? How do you update γ, β (using the jacobian) to increase output values of the batch norm layer?

4. Consider the following network.

$$y = \sin(\mathbf{w}^\top \mathbf{x}) - b \tag{1}$$

- Draw the computational graph of the forward pass of this network. Note that **every** operator is a node with a given arity and output. For example, the + operator is a node which has two input arguments and a single output argument, etc...

- Consider an input $\mathbf{x} = [2, 1]\mathbf{w} = [\frac{\pi}{2}, \pi], b = 0$ and label $l = 2$.
  - Compute the forward pass of the network.
  - Use an L_2 loss (Mean square error) to compute the loss value between the forward prediction y and label l. Add this loss to the computation graph.
  - Use the chain rule to compute the gradient $\frac{\partial L(y,l)}{\partial \mathbf{w}}$ and estimate an update of parameters $\mathbf{w}$ with learning rate $\alpha = 0.5$.

5. You are given a batch consisting of positive sample x_1 and negative samples x_2, x_3. Create a augmented version of positive sample $\hat{x}_1$. Augmented version should be created with colourization technique by adding grayscale value of 0.5 for each pixel.

   - Draw a scheme of the contrastive learning of the model and output representation dimensions. Consider an positive sample $\mathbf{x_1} = [2, 1, 3]$, negative samples $\mathbf{x_2} = [1, 2, 1], \mathbf{x_3} = [1, 2, 3]$ and model as a one layer of 2D convolution with weights $\mathbf{w} = [1, 1]$ and bias $\mathbf{b} = 0$ and label $l = 1$.

   - Calculate contrastive loss $\mathcal{L}$ from model representations (model output features) consisting of L2 normalization in all samples. Loss should be minimized and summed across positive and negative training examples.

     Choice: Alternatively, you can use softmax and cross entropy for loss calculation

   - What is the dimension of loss for one sample?

6. You are given an input volume X of dimension $[batch \times width \times height \times channel] = [4 \times 13 \times 13 \times 2]$
   Consider a 2D convolutional filter F of size $[width \times height] = [3 \times 3]$

   - Assuming a stride of [2,2], What is the size of padding, which ensures that the feature map is half the size of the input map? Note: A padding size of 1 for a $[30 \times 30]$ image gives it a resulting size of $[32 \times 32]$, in other words, zeros are added on both sides.

   - Calculate the total memory in bytes of the learnable parameters of the filter, assuming that each weight is a dual-precision float, aka 'double', aka FP64 which takes up 2 bytes each

   - Calculate the amount of operations performed by a single applciation of the filter (just one stamp). Each addition or multiplication counts as a single operation.

     For example: $\alpha x + \beta y + c$ amounts to 2 multiplication and 2 addition operations, totaling 4 operations.

   - Considering the entire input dimensions of X, given a stride of [2,2], no padding and only valid convolutions, calculate the amount of filter applications ("stamps") that you have to perform to process the entire input.