VIR 2021 **Name:** ____________________
Midterm test
Variant: C Points ____________________

1. **Convergence rate:** You are given the following function:

$$f(\mathbf{x}, \mathbf{w}) = \frac{1}{2}\|\mathbf{w}^\top \mathbf{x}\|_2^2 = \frac{1}{2}((w_1 x_1)^2 + (w_2 x_2)^2)$$

and a single training example $\mathbf{x} = [\sqrt{3}, 1]^\top$. Consider Stochastic Gradient Descend algorithm, which updates the weights as follows:

$$\mathbf{w}^k = \mathbf{w}^{k-1} - \alpha \left.\frac{\partial f^\top(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^{k-1}},$$

where α denotes its learning rate. For which α the SGD:

- converges (at least slowly) in both dimensions?
  Hint: Derive formula for weight values in k-th iteration

$$w_1^k = \rho_1(\alpha)^k w_1^0$$

$$w_2^k = \rho_2(\alpha)^k w_2^0,$$

  where $\rho_i(\alpha)$ denotes convergence rate in dimension $i = 1, 2$.

- oscillates at least in one dimension?

- diverges in one dimension and converges in another dimension?

- What is the best learning rate α^*, which guarantees the fastest convergence rate for arbitrary weight initialization $\mathbf{w}^0$ and this particular training example.
  Hint: The smaller the $|\rho_i(\alpha)|$, the faster the convergence. Choose alpha, which minimize maximal convergence rate:

$$\alpha^* = \arg\min_{\alpha} \max\{|\rho_1(\alpha)|, |\rho_2(\alpha)|\}$$

2. **Convolution feedforward pass:** Network calculates output of 3D convolution with a single 3D kernel $4 \times 5 \times 3$, padding =1, stride = 1. Input is RGBD image (4 channels: red, green, blue and depth) with spatial resolution of 100×100 pixels. Calculate the amount of operations performed during feedforward pass. 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.

3. **Computational graph and backpropagation:** Consider the following network

$$y = \Big(\sin(\mathbf{w})\Big)^\top \mathbf{x},$$

where $\sin(\mathbf{w})$ denotes element-wise function $[\sin(\mathbf{w}_1), \sin(\mathbf{w}_2)]^\top$. Consider a training set consisting of a single pair: input $\mathbf{x} = [2, 1]^\top$ and label $l = -2$. The network is initialized with weights $\mathbf{w} = [\pi/2, \pi]^\top$. You will minimize L_2-norm $(y - l)^2$ by Stochastic Gradient Descend with Momentum.

- Is there a combination of training parameters α (learning rate), β (momentum), which assures convergence into a global minimum (for the given initial weights and training set)? If so, find them and demonstrate the convergence. If there are no such training parameters, explain why and suggest a solution.
  Hint: Look at (i) the gradient in the computational graph and (ii) loss in a global minimum.

4. **MLE for classification:** Maximum likelihood estimate for two-class classification problem uses the following discrete probability distribution

$$p(y|\mathbf{x}, \mathbf{w}) = \begin{cases} \sigma(f(\mathbf{x}, \mathbf{w})) & y = +1 \\ 1 - \sigma(f(\mathbf{x}, \mathbf{w})) & y = -1 \end{cases},$$

where $\sigma(.)$ is the sigmoid function. In order to avoid explicit usage of the *split* (using different functions for different y-values), the probability distribution is simplified as follows:

$$p(y|\mathbf{x}, \mathbf{w}) = \sigma\Big(y \cdot f(\mathbf{x}, \mathbf{w})\Big)$$

Show that these two expressions are equivalent.