VIR 2022 **Name:** Chytrý Student

Exam test

Variant: A Points 20

1. Let us consider gradient learning of the linear regressor $y = \mathbf{w}^\top \mathbf{x}$. Given the single training example $(\mathbf{x} = [\sqrt{3}, 1]^\top, y = 0)$, the least squares learning reduces to the minimization of the following criterion

$$\frac{\partial f}{\partial w_1}\Big|_{x_1} = w_1 x_1^2 \Big|_{x_1} = 3w_1$$

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

$$\frac{\partial f}{\partial w_2}\Big|_{x_2} = w_2 x_2^2 \Big|_{x_2} = w_2$$

TASK 1.1 Derive the recurrent formula for values of weights in the k-th iteration

$$w_1^k = \rho_1(\alpha)^k w_1^0 = (1 - 3\alpha)^2 w_1^0$$
$$w_2^k = \rho_2(\alpha)^k w_2^0, = (1 - \alpha)^2 w_2^1$$

$$w_1^k = w_1^{k-1} - \alpha \cdot 3 w_1^{k-1}$$
$$w_2^k = w_2^{k-1} - \alpha \cdot w_2^{k-1}$$

TASK 1.2 For which learning rate α the gradient descent converges (at least slowly) in both dimensions?
Hint: The smaller the $|\rho_i(\alpha)|$, the faster the convergence. Find α for which both formulas converge to zero.

$\alpha^{\text{convergent}} \in (0, 2/3)$

$|1 - 3\alpha| < 1 \wedge |1 - \alpha| < 1$

TASK 1.3 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: Choose alpha, which minimizes the maximum of both convergence rates:

$$\alpha^* = \arg\min_\alpha \max\{|\rho_1(\alpha)|, |\rho_2(\alpha)|\} = \frac{1}{2}$$

$$1 - \alpha^* = 3\alpha^* - 1$$
$$\alpha^* = \frac{1}{2}$$

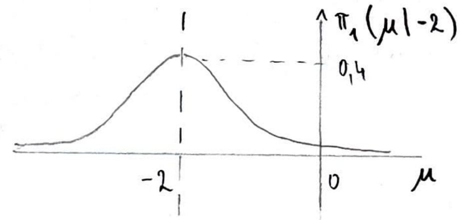
2. Consider stochastic continuous policy, that selects the action $\mathbf{u} \in \mathbb{R}$ in the state $\mathbf{x} \in \mathbb{R}$ according to the following probability distribution:

$$\pi_\theta(\mathbf{u}|\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\theta\mathbf{x} - \mathbf{u})^2)$$

with scalar parameter $\theta = 1$. This policy maps one-dimensional state $\mathbf{x}$ on the Gaussian probability distribution (with the unit variance) of possible actions $\mathbf{u}$.

TASK 2.1 Let us assume that the robot/agent is in state $\mathbf{x}_1 = -2$. Sketch the shape of probability distribution $\pi_\theta(\mathbf{u}|\mathbf{x}_1 = -2)$ from which the actions are drawn.

$$\pi_1(\mu|-2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\mu+2)^2\right) \rightarrow \mathcal{N}(-2,1)$$



TASK 2.2 The policy performs the action $\mathbf{u}_1 = 1$ (that has been randomly generated from the probability distribution), and the robot ends up in the state $\mathbf{x}_2 = +3$. The reward function for the resulting training trajectory $\tau = [\mathbf{x}_1, \mathbf{u}_1, \mathbf{x}_2]$ is $r(\tau) = 2$. Estimate REINFORCE policy gradient:

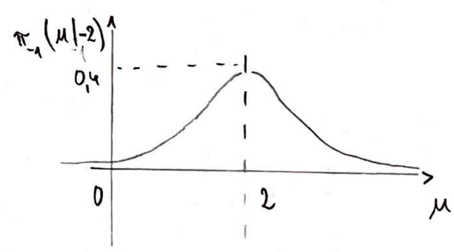
$$\left.\frac{\partial \log \pi_\theta(\mathbf{u}|\mathbf{x})}{\partial\theta}\right|_{\substack{\mathbf{x}=\mathbf{x}_1\\ \mathbf{u}=\mathbf{u}_1}} \cdot r(\tau) = \frac{\partial}{\partial\theta} \log\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta x - \mu)^2\right)\right]\Bigg|_{\substack{x_1\\ \mu_1}} \cdot r(\tau) =$$

$$= -(\theta x - \mu)x \Big|_{\substack{x_1\\ \mu_1}} \cdot r(\tau) = -12$$

TASK 2.3 Update policy parameters by the gradient ascent method with $\alpha = 1/6$ and sketch the shape of the updated distribution $\pi_{\theta^{\text{updated}}}(\mathbf{u}|\mathbf{x}_1 = -2)$

$$\theta^{\text{updated}} = \theta + \alpha(-12) = -1$$

$$\pi_{-1}(\mu|-2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\mu-2)^2\right) \rightarrow \mathcal{N}(2,1)$$



3. You are given an input feature map (image) $\mathbf{x}$, a convolution layer Conv2d(in_channels=3, out_channels=6, kernel_size=5, stride=1, padding=0, dilation=1), an activation function ReLU, a batch normalization layer BatchNorm2d(6), a max pooling layer MaxPool2d(2, 2) and an output $\mathbf{y}$.

TASK 3.1: Consider the following architecture

nemají vliv na RF

$$\mathbf{x} \rightarrow \text{Conv2d} \rightarrow \text{ReLU} \rightarrow \text{BatchNorm2d} \rightarrow \text{MaxPool2d} \rightarrow \mathbf{y}$$

and compute the receptive field (RF) of the output, i.e., the size of the region in the input $\mathbf{x}$ that produces the feature $\mathbf{y}_{i,i}$:

Conv2D: zajímá nás kernel_size a dilation => 5x5

ks = 5 d = 1

MaxPool2D: +1 k RF (2x2)

RF = 6x6

TASK 3.2: Tick the correct answer (multiple choice).

- ☐ A receptive field depends on the size of the input image.
- ☑ A batch normalization procedure consists of feature-wise operations which do not alter the receptive field of the network.
- ☑ Some linear layers increase the size of the receptive field.
- ☑ The larger the convolutional stride, the larger the receptive field.
- ☑ By adding more convolutional layers, an arbitrarily large receptive field can be achieved.
- ☐ A large receptive field usually negatively impacts the ability of the neural network to understand the context of the input image.

Nebude v testu

4. Consider the composite normalizing flow $f : \mathbb{R}^3 \to \mathbb{R}^3$, $f = f_1 \circ f_2$ of length 2

$$P_Z \sim \boldsymbol{z} \underset{f_1}{\overset{g_1}{\rightleftarrows}} \boldsymbol{y} \underset{f_2}{\overset{g_2}{\rightleftarrows}} \boldsymbol{x} \sim P_X$$

$Z \sim U([-1,1]^3)$, i.e., Z is a real random vector in $\mathbb{R}^3$ with uniform distribution over the cube of edge length 2.

TASK 4.1: g_1 is specified as a linear transformation

$$g_1 : \boldsymbol{y} = A\boldsymbol{z} + \boldsymbol{b},$$

where A is a 3×3 square matrix and $\boldsymbol{b}$ is a 3×1 column vector

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

We have $f_1 = g_1^{-1}$. Calculate the determinant of Jacobian of f_1, i.e., calculate $\det(\mathrm{J}_{f_1}) = \det(\mathrm{J}_{g_1^{-1}})$. Note that you do not need to know the inverse matrix A^{-1} to complete this task.

TASK 4.2: f_2 is a simple coupling flow $\mathbb{R}^3 \to \mathbb{R}^3$ that is specified as follows, $\boldsymbol{y} = f_2(\boldsymbol{x})$:

$$\begin{aligned} y_1 &= x_1, \\ y_2 &= x_2 \cdot \exp(+2x_1) + x_1, \\ y_3 &= x_3 \cdot \exp(-2x_1) + x_1. \end{aligned}$$

Calculate the determinant of the Jacobian f_2.

Nebude v testu

TASK 4.3: Consider the real data point $\boldsymbol{x}^* = (0,1,1)^T$. Assume that $\boldsymbol{x}^*$ was generated from distribution P_X which is further normalized by the flow transformation f to the distribution $P_Z \sim U([-1,1]^3)$.
Calculate the latent representation $\boldsymbol{z}^*$ of $\boldsymbol{x}^*$ under f (Hint: A^{-1} is not required).

$\boldsymbol{z}^* = f(\boldsymbol{x}^*) = f_1 \circ f_2(\boldsymbol{x}^*) =$

TASK 4.4: What is the value of density p_X at this point? Use the change of variable formula

$$p_X(\boldsymbol{x}) = p_Z(f(\boldsymbol{x})) \cdot |\det(\mathrm{J}_f)|$$

and results from TASK 4.1, 4.2, 4.3 to complete the task.

$p_X(\boldsymbol{x}) =$