**Exercise 1: Class Prototypes (25 P)**

Consider the linear model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ mapping some input $\boldsymbol{x}$ to an output $f(\boldsymbol{x})$. We would like to interpret the function $f$ by building a prototype $\boldsymbol{x}^\star$ in the input domain which produces a large value $f$. Activation maximization produces such interpretation by optimizing

$$\max_{\boldsymbol{x}} \left[ f(\boldsymbol{x}) + \Omega(\boldsymbol{x}) \right].$$

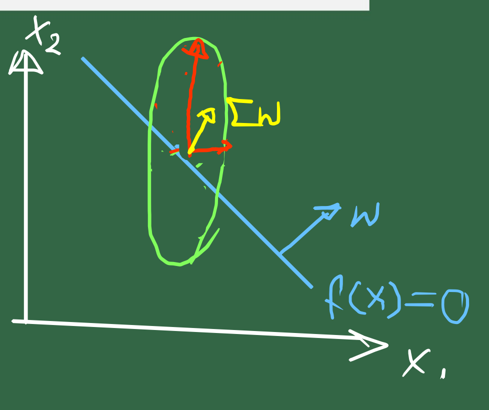
*Find* the prototype $\boldsymbol{x}^\star$ obtained by activation maximization subject to $\Omega(\boldsymbol{x}) = \log p(\boldsymbol{x})$ with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ and $\Sigma$ are the mean and covariance.

$$\frac{\partial}{\partial x}\left(w^\top x + b - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \text{const}\right).$$

$$= w - \Sigma^{-1}(x-\mu) \overset{!}{=} 0$$

$$\Leftrightarrow \quad w = \Sigma^{-1}(x-\mu) \qquad | \Sigma(\cdot)$$

$$\Leftrightarrow \quad \Sigma w = x - \mu \quad \Leftrightarrow \quad x^\star = \Sigma w + \mu$$




**Exercise 2: Shapley Values (25 P)**

Consider the function $f(\boldsymbol{x}) = \min(x_1, \max(x_2, x_3))$. Compute the Shapley values $\phi_1, \phi_2, \phi_3$ for the prediction $f(\boldsymbol{x})$ with $\boldsymbol{x} = (1, 1, 1)$. (We assume a reference point $\widetilde{\boldsymbol{x}} = 0$, i.e. we set features to zero when removing them from the coalition).

$\phi_1$:

| $S$ | $\alpha_S$ | | $f(x_{S \cup \{1\}}) - f(x_S)$ | |
|---|---|---|---|---|
| $\{\}$ | $1/3$ | $\cdot$ | $0$ | $= 0$ |
| $\{2\}$ | $1/6$ | $\cdot$ | $1 - 0$ | $= 1$ |
| $\{3\}$ | $1/6$ | $\cdot$ | $1 - 0$ | $= 1$ |
| $\{2,3\}$ | $1/3$ | $\cdot$ | $1 - 0$ | $= 1$ |

$$\phi_1 = \frac{1}{3}\cdot 0 + \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 1 + \frac{1}{3}\cdot 1 = \frac{2}{3}$$

Conservation:
$$\phi_2 + \phi_3 = \frac{1}{3}$$

Symmetry:
$$\phi_2 = 1/6 \qquad \phi_3 = 1/6$$

**Exercise 3: Taylor Expansions (25 P)**

Consider the simple radial basis function

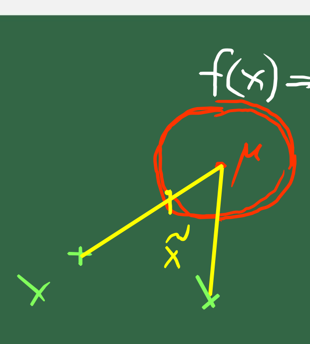
$$f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta$$

with $\theta > 0$. For the purpose of extracting an explanation, we would like to build a first-order Taylor expansion of the function at some root point $\widetilde{\boldsymbol{x}}$. We choose this root point to be taken on the segment connecting $\boldsymbol{\mu}$ and $\boldsymbol{x}$ (we assume that $f(\boldsymbol{x}) > 0$ so that there is always a root point on this segment).

*Show* that the first-order terms of the Taylor expansion are given by

$$\phi_i = \frac{(x_i - \mu_i)^2}{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2} \cdot (\|\boldsymbol{x} - \boldsymbol{\mu}\| - \theta)$$

$$f(x) = f(\tilde{x}) + \sum_i [\nabla_x f(\tilde{x})]_i \cdot (x_i - \tilde{x}_i) + \ldots$$




$$\tilde{x} = \mu + t(x - \mu)$$

$$x - \tilde{x} = (1 - t)(x - \mu)$$

$$\phi_i = \frac{\tilde{x}_i - \mu_i}{\|\tilde{x} - \mu\|} \cdot (1 - t)(x_i - \mu_i)$$

$$= \frac{\mu_i + t(x_i - \mu_i) - \mu_i}{\|\mu + t(x - \mu) - \mu\|} \cdot (1 - t)(x_i - \mu_i) = \frac{\cancel{t}(x_i - \mu_i)}{\|\cancel{t}(x - \mu)\|} \cdot (1 - t)(x_i - \mu_i)$$

$$= \frac{(x_i - \mu_i)^2}{\|x - \mu\|} \cdot (1 - t)$$

$$= \frac{(x_i - \mu_i)^2}{\|x - \mu\|^2} \cdot \|x - \mu\| \cdot (1 - t)$$

sum to 1

$$\sum_i \phi_i = \|x - \mu\| - \theta = \sum_i \varphi_i \cdot f(x)$$

$$\sum_i \varphi_i = 1$$

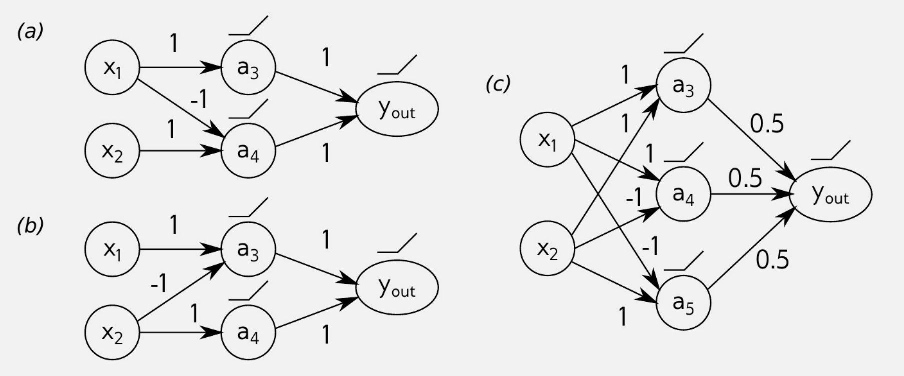
$$= \frac{(x_i - \mu_i)^2}{\|x - \mu\|^2}(\|x - \mu\| - \theta)$$

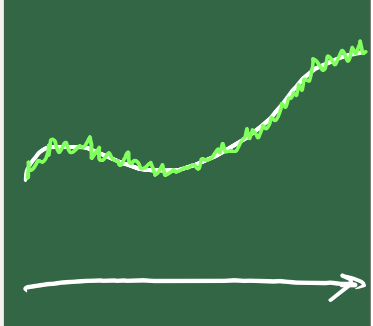
2nd: $f(\tilde{x}) = 0$, solve for $t$.

$$\|x - \mu\|^2 = \sum_i (x_i - \mu_i)^2$$

**Exercise 4: Layer-Wise Relevance Propagation (25 P)**

We would like to test the dependence of layer-wise relevance propagation (LRP) on the structure of the neural network. For this, we consider the function $y = \max(x_1, x_2)$, where $x_1, x_2 \in \mathbb{R}^+$ are the input activations. This function can be implemented as a ReLU network in multiple ways. Three examples are given below.



We consider the propagation rule:

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

where $j$ and $k$ are indices for two consecutive layers and where $()^+$ denotes the positive part. This propagation rule is applied to both layers.

*Give* for each network the computational steps that lead to the scores $R_1$ and $R_2$, and the obtained relevance values. More specifically, express $R_1$ and $R_2$ as a function of $R_3$ and $R_4$ (and $R_5$), and express the latter relevances as a function of $R_{\text{out}} = y$.

$$R_{out} = y$$

$$R_3 = \frac{a_3 \cdot \frac{1}{2}}{a_3 \cdot \frac{1}{2} + a_4 \frac{1}{2} + a_5 \frac{1}{2}} \cdot R_{out}$$

$$R_4 = \frac{a_4 \cdot \frac{1}{2}}{a_3 \cdot \frac{1}{2} + a_4 \frac{1}{2} + a_5 \frac{1}{2}} \cdot R_{out}$$

$$R_5 = \frac{a_5 \cdot \frac{1}{2}}{a_3 \cdot \frac{1}{2} + a_4 \frac{1}{2} + a_5 \frac{1}{2}} \cdot R_{out}$$

$$R_2 = \frac{x_2 \cdot 1}{x_2 \cdot 1 + x_1 \cdot 1} \cdot R_3 + \frac{x_2 \cdot 0}{x_2 \cdot 0 + x_1 \cdot 1} R_4$$

$$+ \frac{x_2 \cdot 1}{x_2 \cdot 1 + x_1 \cdot 0} \cdot R_5$$

$$R_1 = \frac{x_1 \cdot 1}{x_2 \cdot 1 + x_1 \cdot 1} R_3 + \frac{x_1 \cdot 1}{x_2 \cdot 0 + x_1 \cdot 1} R_4$$

$$+ \frac{x_1 \cdot 0}{x_2 \cdot 1 + x_1 \cdot 0} R_5$$

a)

$$R_{out} = y$$

$$R_3 = \frac{a_3 \cdot w_{3out}^+}{a_3 w_{3out}^+ + a_4 w_{4out}^+} \cdot R_{out}$$

$$= \frac{a_3 \cdot 1}{a_3 \cdot 1 + a_4 \cdot 1} \cdot R_{out}$$

$$R_4 = \frac{a_4 \cdot 1}{a_3 \cdot 1 + a_4 \cdot 1} R_{out}$$

$$R_2 = \frac{x_2 \cdot w_{24}^+}{x_2 \cdot w_{24}^+ + x_1 \cdot w_{14}^+} R_4$$

$$= R_4$$

$$R_1 = \frac{x_1 w_{13}^+}{x_1 w_{13}^+} R_3 + \frac{x_1 w_{14}^+}{x_2 w_{24}^+ + x_1 w_{14}^+} R_4$$

$$= R_3$$