## Machine Learning for Graphs and Sequential Data Exercise Sheet 05

# Robustness of Machine Learning Models II

**Problem 1:** On slide 15 of the robustness chapter, we have defined an optimization problem for untargeted attacks, i.e. we aim to have the sample $\hat{\boldsymbol{x}}$ classified as **any** class other than the correct one:

$$\min_{\hat{\boldsymbol{x}}} \mathcal{D}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + \lambda \cdot L(\hat{\boldsymbol{x}}, y)$$

The loss function is defined as:

$$L(\hat{\boldsymbol{x}}, y) = \left[ Z(\hat{\boldsymbol{x}})_y - \max_{i \neq y} Z(\hat{\boldsymbol{x}})_i \right]_+,$$

where $[\boldsymbol{x}]_+$ is shorthand for $\max(\boldsymbol{x}, 0)$ and $Z(\boldsymbol{x})_i = \log f(\boldsymbol{x})_i$ (i.e. log probability of class $i$. Here, $L(\hat{\boldsymbol{x}}, y)$ is positive if $\hat{\boldsymbol{x}}$ is classified correctly and 0 otherwise.

Provide an alternative loss function to turn this attack into a targeted attack, i.e. we aim to have the sample $\boldsymbol{x}$ classified as a *specific* target class $t$.

**Problem 2:** Recall from slide 41 the MILP constraints expressing the ReLU activation function. Show that a continuous relaxation on $\boldsymbol{a}$ leads to the convex relaxation constraints on slide 54. That is, we replace the constraint $\boldsymbol{a}_i \in \{0, 1\}$ with $\boldsymbol{a}_i \in [0, 1]$.