1 Finding the minimal change that will get rid of the network's WaterMark

Given a watermarked trained neural network as described here. We tested what is the minimal change to the network last layer in order to "remove" some watermarks from the network.

1.1 Defining the problem

Given a neural network N with an output size m the network decision for an input x is defined as the coordinate with the maximal value, if the network output is the vector y the decision is $argmax_{i\in[m]}\left\{y_i\right\}$

Given a watermarked network N with a set of K watermarks (A set of inputs to the network) $\{x_1, \cdots, x_K\}$ we'll mark the network last layer L such that L is a $m \times n$ matrix were n is the layer's number of neurons and m is the network output size. The change to the last layer will be a matrix with the same dimension as L we'll mark as ε, such that $\varepsilon_{i,j}$ is the change to the last layer matrix entry $L_{i,j}$. Well measure the overall change to the layer as $\|\varepsilon\|_\infty = max_{i,j}\left\{|\varepsilon_{i,j}|\right\}$.

For a certain input x we're only interested in the input to the last layer we'll mark the input to the last layer v. v is a $n \times 1$ vector. So the original network output $y = Lv$ and the changed network output is $y' = (L + \varepsilon)v$. For a single input x we need to find the minimal ε so that the $argmax_{i\in[m]}\left\{y_i\right\} \neq argmax_{i\in[m]}\left\{y'_i\right\}$

Denote $d := argmax_{i\in[m]}\left\{y_i\right\}$
For some $d' \in [m]\,, d' \neq d$ finding ε with minimal $\|\varepsilon\|_\infty$ such that $y' = (L+\varepsilon)v$ and $d' = argmax_{i\in[m]}\left\{y'_i\right\}$ can be described in a Linear Programming form like so:

$$\begin{aligned} Minimize:&\quad c \\ Subject\ to:&\quad -c \leq \varepsilon_{i,j} \leq c\ \forall i,j \\ &\quad y' = (L+\varepsilon)v \\ &\quad y'_d \leq y'_{d'} \end{aligned}$$