A APPENDIX

A.1 PROOF OF THEOREM 1

Since $\mathcal{Z}_L \in \hat{\mathcal{Z}}_L$ it trivially holds that

$$\min_{\mathbf{z}\in\hat{\mathcal{Z}}_L} \quad \max\{\mathbf{c}_{y,i}^\top \mathbf{z}, \mathbf{c}_{a,i}^\top \mathbf{z}\} \leq \min_{\mathbf{z}\in\mathcal{Z}_L} \quad \max\{\mathbf{c}_{y,i}^\top \mathbf{z}, \mathbf{c}_{a,i}^\top \mathbf{z}\} \tag{20}$$

The lower bound is now a convex minimization, which can be rewritten as

$$\min_{\mathbf{z}\in\hat{\mathcal{Z}}_L} \quad \max\{\mathbf{c}_{y,i}^\top \mathbf{z}, \mathbf{c}_{a,i}^\top \mathbf{z}\} = \min_{\tau,\mathbf{z}\in\hat{\mathcal{Z}}_L} \quad \tau \qquad \text{s. t.} \quad \mathbf{c}_{y,i}^\top \mathbf{z} \leq \tau \quad , \mathbf{c}_{a,i}^\top \mathbf{z} \leq \tau.$$

Defining the slack variables $\eta_a \geq 0$ and $\eta_y \geq 0$ for the inequality constraints, the Lagrangian can be written as

$$\mathcal{L}(\tau, \mathbf{z}, \eta_a, \eta_y) = \tau + \eta_a(\mathbf{c}_{a,i}^\top \mathbf{z} - \tau) + \eta_y(\mathbf{c}_{y,i}^\top \mathbf{z} - \tau)$$

and minimizing $\mathcal{L}(\tau, \mathbf{z}, \eta_a, \eta_y)$ with respect to the primal variable τ, yields $\eta_a + \eta_y = 1$. Defining $\eta := \eta_a = 1 - \eta_y$, and using the fact that the dual maximization always serves as a lower bound on the primal we get

$$\max_{0\leq\eta\leq 1} \min_{\mathbf{z}\in\hat{\mathcal{Z}}_L} \quad \Big(\eta\ \mathbf{c}_{a,i} + (1-\eta)\ \mathbf{c}_{y,i}\Big)^\top \mathbf{z} \leq \min_{\mathbf{z}\in\mathcal{Z}_L} \quad \max\{\mathbf{c}_{y,i}^\top \mathbf{z}, \mathbf{c}_{a,i}^\top \mathbf{z}\}. \square$$

A.2 PROOF OF THEOREM 2

Following on the statement of Theorem 1 and by substituting $\mathbf{z} = \mathbf{W}_L^\top \mathbf{z}_{L-1} + \mathbf{b}_L$, we get

$$\max_{0\leq\eta\leq 1} \min_{\mathbf{z}_{L-1}\in\hat{\mathcal{Z}}_{L-1}} \quad \Big(\eta\ \mathbf{c}_{a,i} + (1-\eta)\ \mathbf{c}_{y,i}\Big)^\top \Big(\mathbf{W}_L^\top \mathbf{z}_{L-1} + \mathbf{b}_L\Big) \leq \min_{\mathbf{z}\in\hat{\mathcal{Z}}_L} \quad \max\{\mathbf{c}_{y,i}^\top \mathbf{z}, \mathbf{c}_{a,i}^\top \mathbf{z}\} \tag{21}$$

which can be reordered as

$$\max_{0\leq\eta\leq 1} \min_{\underline{\mathbf{z}}_{L-1}\leq\mathbf{z}_{L-1}\leq\bar{\mathbf{z}}_{L-1}} \quad (\boldsymbol{\omega}_1 + \eta\boldsymbol{\omega}_2)^\top \mathbf{z}_{L-1} \tag{22}$$

where $\boldsymbol{\omega}_1 := \mathbf{W}_L \mathbf{c}_{y,i}$ and $\boldsymbol{\omega}_2 := \mathbf{W}_L(\mathbf{c}_{a,i} - \mathbf{c}_{y,i})$, which then equals

$$\max_{0\leq\eta\leq 1} \quad (\boldsymbol{\omega}_1 + \eta\boldsymbol{\omega}_2)^\top \hat{\mathbf{z}}_{L-1} \tag{23}$$

where minimization w.r.t. $\mathbf{z}_{L-1}$ is solved by the (this is under the setting for most networks with positive activations, and thus lower bound $\underline{\mathbf{z}}_l$ is always non-negative)

$$[\hat{\mathbf{z}}_{L-1}]_j = \begin{cases} [\bar{\mathbf{z}}_{L-1}]_j & \text{if} \quad [\boldsymbol{\omega}_1 + \eta\boldsymbol{\omega}_2]_j \leq 0 \\ [\underline{\mathbf{z}}_{L-1}]_j & \text{if} \quad [\boldsymbol{\omega}_1 + \eta\boldsymbol{\omega}_2]_j \geq 0 \end{cases} \tag{24}$$

and can be rewritten as

$$\max_{0\leq\eta\leq 1} \quad \sum_{j=1}^{n_{L-1}} \Big[\boldsymbol{\omega}_1 + \eta\boldsymbol{\omega}_2\Big]_j \left(1_{\{\eta\leq -\frac{\omega_{1,j}}{\omega_{2,j}}\}} [\bar{\mathbf{z}}_{L-1}]_j + 1_{\{\eta\geq -\frac{\omega_{1,j}}{\omega_{2,j}}\}} [\underline{\mathbf{z}}_{L-1}]_j\right) \tag{25}$$

and can be rewritten as

$$\max_{0\leq\eta\leq 1} \quad \sum_{j=1}^{n_{L-1}} \left(1_{\{\eta\leq -\frac{\omega_{1,j}}{\omega_{2,j}}\}} [\boldsymbol{\omega}_1 \circ \bar{\mathbf{z}}_{L-1} + \eta\boldsymbol{\omega}_2 \circ \bar{\mathbf{z}}_{L-1}]_j + 1_{\{\eta\geq -\frac{\omega_{1,j}}{\omega_{2,j}}\}} [\boldsymbol{\omega}_1 \circ \underline{\mathbf{z}}_{L-1} + \boldsymbol{\omega}_2 \circ \underline{\mathbf{z}}_{L-1}]_j\right) \tag{26}$$

where "$\circ$" denotes the elementwise multiplication. Thus, due to the concavity of the dual, optimal η can be found by evaluationg the objective in between the break points which are given by $\boldsymbol{\zeta} := [\zeta_1, ..., \zeta_{n_L}] := -\boldsymbol{\omega}_1/\boldsymbol{\omega}_2$ with element-wise division.

To do this, let us use $\mathbf{s}$ to denote the n_L-ary tuple of indices that sorts $\boldsymbol{\zeta}$. That is

$$\tilde{\boldsymbol{\zeta}} = [\tilde{\zeta}_1, ..., \tilde{\zeta}_{n_L}] := \Pi_{\mathbf{s}}(\boldsymbol{\zeta}) := [\zeta_{s_1}, ..., \zeta_{s_{n_L}}] \quad \text{s.t.} \quad \zeta_{s_1} \leq ... \leq \zeta_{s_{n_L}}$$

with operator $\Pi_s(.)$ denoting the permutation of its arguments according to $\mathbf{s}$, such that $\tilde{\zeta}_i = \zeta_{s_i} \forall i$, and $\tilde{\boldsymbol{\zeta}}$ is sorted in the increasing order .

We can also rewrite the problem by summing over the indices in the sorting set $\mathbf{s}$ instead, as

$$\max_{0\leq\eta\leq 1} \quad \sum_{j=1}^{n_{L-1}} \left(1_{\{\eta\leq -\frac{\omega_{1,s_j}}{\omega_{2,s_j}}\}} [\boldsymbol{\omega}_1 \circ \bar{\mathbf{z}}_{L-1} + \eta \boldsymbol{\omega}_2 \circ \bar{\mathbf{z}}_{L-1}]_{s_j} + 1_{\{\eta\geq -\frac{\omega_{1,s_j}}{\omega_{2,s_j}}\}} [\boldsymbol{\omega}_1 \circ \underline{\mathbf{z}}_{L-1} + \boldsymbol{\omega}_2 \circ \underline{\mathbf{z}}_{L-1}]_{s_j} \right). \tag{27}$$

Now let us define $\underline{\mathbf{u}}_1 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_1 \circ \underline{\mathbf{z}}_{L-1})$, $\bar{\mathbf{u}}_1 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_1 \circ \bar{\mathbf{z}}_{L-1})$, $\underline{\mathbf{u}}_2 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_2 \circ \underline{\mathbf{z}}_{L-1})$, $\bar{\mathbf{u}}_2 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_2 \circ \bar{\mathbf{z}}_{L-1})$, we get

$$\max_{0\leq\eta\leq 1} \quad \sum_{j=1}^{n_{L-1}} \left(1_{\{\eta\leq\tilde{\zeta}_j\}} (\bar{u}_{1,j} + \eta \bar{u}_{2,j}) + 1_{\{\eta\geq\tilde{\zeta}_j\}} \left(\underline{u}_{1,j} + \eta \underline{u}_{2,j}\right) \right). \tag{28}$$

By breaking the objective of maximization into piece-wise terms, and by imposing the feasible set $0 \leq \eta \leq 1$ by finding indices

$$m = \min_{\zeta_{s_\nu}\geq 0} \nu \quad \text{and} \quad M = \max_{\zeta_{s_\nu}\leq 1} \nu$$

we can reduce the problem down to piece-wise maximizations for $m \leq \nu \leq M-1$ as

$$\max_{\tilde{\zeta}_\nu\leq\eta\leq\tilde{\zeta}_{\nu+1}} \quad \sum_{j=1}^{\nu} \left(\underline{u}_{1,j} + \eta \underline{u}_{2,j}\right) + \sum_{j=\nu+1}^{n_{L-1}} (\bar{u}_{1,j} + \eta \bar{u}_{2,j}) \tag{29}$$

which will be maximized with lower bound of $\eta = \tilde{\zeta}_\nu$ if the coefficient of η is negative, and with the upper bound $\eta = \tilde{\zeta}_{\nu+1}$ otherwise.

So, the overall maximization boils down to obtaining $\alpha_\nu + \beta_\nu$ for $\nu = m, ..., M$ where

$$\alpha_\nu := \sum_{i=1}^{\nu} \underline{u}_{1,i} + \sum_{i=\nu+1}^{n_{L-1}} \bar{u}_{1,i} = \alpha_{\nu-1} + \underline{u}_{1,\nu} - \bar{u}_{1,\nu}$$

and

$$q_\nu = \sum_{i=1}^{\nu} \underline{u}_{2,i} + \sum_{i=\nu+1}^{n_{L-1}} \bar{u}_{2,i} = q_{\nu-1} + \underline{u}_{2,\nu} - \bar{u}_{2,\nu}$$

and

$$\beta_\nu = \left(\tilde{\zeta}_\nu 1_{\{q_\nu\leq 0\}} + \tilde{\zeta}_{\nu+1} 1_{\{q_\nu > 0\}} \right) \times q_\nu.$$

Values $\alpha_\nu + \beta_\nu$ can be efficiently computed by a forward cumulative sum of $\underline{\mathbf{u}}_1$ and $\underline{\mathbf{u}}_2$, and forward-backward cumulative sum of $\bar{\mathbf{u}}_1$ and $\bar{\mathbf{u}}_2$, thus imposing the overall complexity which is dominated by the sorting at $\mathcal{O}(n_{L-1}\log(n_{L-1}))$. .□

A.3 Description of Algorithm 1

Here is a step-by-step walk-through for Algorithm 1, with insight on how these steps are performed and why.

It is important to notice that the optimization in Theorem 2 could also be solved alternatively via bi-section which maybe simpler, however Alg. 1 solves it analytically.

1. Form vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, which are the last layer values as $\boldsymbol{\omega}_1 = \mathbf{W}_L \mathbf{c}_{y,i}$ and $\boldsymbol{\omega}_2 = \mathbf{W}_L(\mathbf{c}_{a,i} - \mathbf{c}_{y,i})$
2. Define $\boldsymbol{\zeta} = [\zeta_1, ..., \zeta_{n_L}] := -\boldsymbol{\omega}_1/\boldsymbol{\omega}_2$ and get the vector of indices $\mathbf{s}$ that sorts it, i.e., $\zeta_{s_1} \leq \cdots \leq \zeta_{s_{n_{L-1}}}$

3. Form the element-wise product of $(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ with $(\underline{\mathbf{z}}_{L-1}, \bar{\mathbf{z}}_{L-1})$), and sort them according to the index set s.
   $\underline{\mathbf{u}}_1 = \Pi_{\mathbf{s}}(\boldsymbol{\omega}_1 \circ \underline{\mathbf{z}}_{L-1})$, $\bar{\mathbf{u}}_1 = \Pi_{\mathbf{s}}(\boldsymbol{\omega}_1 \circ \bar{\mathbf{z}}_{L-1})$, $\underline{\mathbf{u}}_2 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_2 \circ \underline{\mathbf{z}}_{L-1})$, $\bar{\mathbf{u}}_2 := \Pi_{\mathbf{s}}(\boldsymbol{\omega}_2 \circ \bar{\mathbf{z}}_{L-1})$.
4. Get the lowest and highest indexes (m,M) such that the sorted ζ vector value at those indices are between 0 and 1.
5. Now, at this point the goal is to iterate over the index $\nu = m, ..., M$, and evaluate the objective (which can be expressed as $\alpha_\nu + \beta_\nu$) for each ν, and select the optimal value.
   However, this can be done in an intelligent way to save computation. To this end initialize these values at

$$\alpha_m = \sum_{i=1}^{m} \underline{u}_{1,i} + \sum_{i=m+1}^{n_{L-1}} \bar{u}_{1,i}$$
$$q_m = \sum_{i=1}^{m} \underline{u}_{2,i} + \sum_{i=m+1}^{n_{L-1}} \bar{u}_{2,i}$$
$$\beta_m = \left(\tilde{\zeta}_\nu 1_{\{q_\nu \leq 0\}} + \tilde{\zeta}_{\nu+1} 1_{\{q_\nu > 0\}}\right) \times q_\nu$$

6. Iterate over $\nu = m+1, ..., M$ and set

$$\alpha_\nu = \alpha_{\nu-1} + \underline{u}_{1,\nu} - \bar{u}_{1,\nu}$$
$$q_\nu = q_{\nu-1} + \underline{u}_{2,\nu} - \bar{u}_{2,\nu}$$
$$\beta_\nu = q_\nu \left(\zeta_{s_\nu} 1_{\{q_\nu \leq 0\}} + \zeta_{s_{\nu+1}} 1_{\{q_\nu > 0\}}, \right)$$

7. Return the maximum value of $\alpha_\nu + \beta_\nu$ over $\nu = m, ..., M-1$

A.4 Proof of Theorem 3

Let us start by splitting the feasible set into disjoint sets of

$$\hat{\mathcal{Z}}^a_{L-1} := \{\mathbf{z}_{L-1} \mid z_{L-1,a} \geq z_{L-1,y}\}, \text{ and } \hat{\mathcal{Z}}^y_{L-1} := \{\mathbf{z}_{L-1} \mid z_{L-1,a} < z_{L-1,y}\}$$

where

$$\hat{\mathcal{Z}}_{L-1} = \hat{\mathcal{Z}}^y_{L-1} \cup \hat{\mathcal{Z}}^a_{L-1}, \text{ and } \hat{\mathcal{Z}}^y_{L-1} \cap \hat{\mathcal{Z}}^a_{L-1} = \emptyset.$$

Proof is carried out by considering $\mathbf{z} \in \hat{\mathcal{Z}}^y_{L-1}$ and $\mathbf{z} \in \hat{\mathcal{Z}}^a_{L-1}$, separately.

Restricting $\mathbf{z} \in \hat{\mathcal{Z}}^y_{L-1}$ we have $\ell_{\text{xent}\backslash a}(f_\theta(\mathbf{x}+\boldsymbol{\delta}), y) \leq \ell_{\text{xent}\backslash y}(f_\theta(\mathbf{x}+\boldsymbol{\delta}), a)$ which leads to

$$L^{\text{abstain}}_{\text{robust}}(\mathbf{x}, y; \theta) = \max_{\boldsymbol{\delta} \in \Delta} \min \left\{\ell_{\text{xent}\backslash a}(f_\theta(\mathbf{x}+\boldsymbol{\delta}), y), \ell_{\text{xent}\backslash y}(f_\theta(\mathbf{x}+\boldsymbol{\delta}), a)\right\} \tag{30}$$
$$\leq \max_{\mathbf{z}_{L-1} \in \hat{\mathcal{Z}}^y_{L-1}} \ell_{\text{xent}\backslash a}(\mathbf{z}_L, y) \quad \text{s.}t. \quad \mathbf{z}_L = \mathbf{W}_L^\top \mathbf{z}_{L-1} + \mathbf{b}_L \tag{31}$$

Loss function $\ell_{\text{xent}\backslash a}$ is the cross entropy loss defined on the K-dimensional vector $[z_{L,1}, \cdots, z_{L,K}]$ and class y, and thus following Wong & Kolter (2018) given its transnational invariance equals

$$\max_{\mathbf{z}_{L-1} \in \hat{\mathcal{Z}}^y_{L-1}} \ell_{\text{xent}\backslash a}(\mathbf{z}_L, y) = \max_{\mathbf{z}_{L-1} \in \hat{\mathcal{Z}}^y_{L-1}} \ell_{\text{xent}\backslash a}(\mathbf{z}_L - z_{L,y}\mathbf{1}, y) \quad \text{s.}t. \quad \mathbf{z}_L = \mathbf{W}_L^\top \mathbf{z}_{L-1} + \mathbf{b}_L \tag{32}$$

with $\mathbf{1}$ denoting the $(K+1)$-dimensional vector of all ones. Given the invariance of $\ell_{\text{xent}\backslash a}$ with respect to $z_{L,a}$, it can finally be upperbounded by taking the upperbound for all i indices where $i = 1, ..., K, i \neq a, y$ and lowerbound at index $i = y$. Note that for $i = y$, value $[\mathbf{z}_L - z_{L,y}\mathbf{1}]_i = 0$, and a lower bound on other entries $i = 1, ..., K, i \neq a, y$ can be obtained by

$$z_{L,i} - z_{L,y} = -\max\{z_{L,y} - z_{L,i}, z_{L,a} - z_{L,i}\} = -\max\{\mathbf{c}^\top_{y,i}\mathbf{z}, \mathbf{c}^\top_{a,i}\mathbf{z}\} \tag{33}$$
$$\leq -\min_{\mathbf{z}_L \in \mathcal{Z}_L} \max\{\mathbf{c}^\top_{y,i}\mathbf{z}, \mathbf{c}^\top_{a,i}\mathbf{z}\} \leq -J_i(\mathbf{x}, y) \leq -J_i^{\underline{\eta}, \bar{\eta}}(\mathbf{x}, y) \tag{34}$$

where the first equality holds since $\hat{\mathcal{Z}}^y_{L-1} := \{\mathbf{z}_{L-1} \mid z_{L-1,a} < z_{L-1,y}\}$ for $\mathbf{z} \in \hat{\mathcal{Z}}^y_{L-1}$, second inequality is due to Theorem 2, and third inequality is given by Eq. 15.

Thus, for $\mathbf{z} \in \hat{\mathcal{Z}}^y_{L-1}$ the loss term is now upperbounded by

$$L^{\text{abstain}}_{\text{robust}}(\mathbf{x}, y; \theta) \leq \ell_{xent\setminus a}(-\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y), y)$$

where

$$[\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y)]_i = \begin{cases} 0 \quad \text{if} & i = a, y \\ J_i^{\underline{\eta},\bar{\eta}}(\mathbf{x}, y) & \text{otherwise.} \end{cases} \tag{35}$$

Similarly, it can be shown that for Thus, for $\mathbf{z} \in \hat{\mathcal{Z}}^a_{L-1}$ the loss term is now upperbounded by

$$L^{\text{abstain}}_{\text{robust}}(\mathbf{x}, y; \theta) \leq \ell_{xent\setminus y}(-\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y), a).$$

The equality of $\ell_{xent\setminus y}(-\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y), a) = \ell_{xent\setminus a}(-\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y), y)$ trivially follows from the fact that $[\mathbf{J}_{\epsilon,\theta}(\mathbf{x}, y)]_i = 0$ for $i = a, y$.

Thus, since $\hat{\mathcal{Z}}_{L-1} = \hat{\mathcal{Z}}^y_{L-1} \cup \hat{\mathcal{Z}}^a_{L-1}$, the proof is complete. .□

B Appendix: Experiment set up

Training parameters and schedules are similar to (Gowal et al., 2018) and (Zhang et al., 2020), and outlined in detail here. For training the classifier network with architecture given in Table 2, for both datasets, Adam optimizer with learning rate of 5×10^{-4} is used. Unless stated differently, κ is scheduled by a linear ramp-down process, starting at 1, which after a warm-up perio,d is ramped down to value $\kappa_{\text{end}} = 0.5$. Value of ϵ during the training is also simultaneously scheduled by a linear ramp-up, starting at 0, and ramped up to the final value of ϵ_{train}, reported in Tabel 1, and networks are trained with a single NVIDIA Tesla V100S GPU.

- For MNIST, the network is trained in 100 epochs with batchsize of 100 (total of $60K$ steps). A warm up period of 3 epochs ($2K$ steps) is used (normal classification training with no robust loss), followed up by a ramp-up duration of 18 epochs ($10K$ steps), and the learning rate is decayed $\times 10$ at epochs 25 and 42. No data augmentation is used. Furthermore, fixed selection of $\bar{\eta} = 0.9$ and $\underline{\eta} = 0.1$ during training is used for this dataset with no ramp-down. Reported numbers in Table 1 corresponds to $\lambda_1 = 1$ and $\lambda_2 = 2$ for $\epsilon = 0.3$, and $\lambda_1 = 0.6$ and $\lambda_2 = 1$ for $\epsilon = 0.4$ respectively.
- For CIFAR10, the network is trained in 3200 epochs with batchsize of 1600 (total of $100K$ steps). A warm up period of 320 epochs ($10K$ steps) is used (normal classification training with no robust loss), followed up by a ramp-up duration of 1600 epochs ($50K$ steps), and the learning rate is decayed $\times 10$ at epochs 2600 and 3040 (60k and 90K steps). Random translations and flips, and normalization of each image channel (using the channel statistics from the train set) is used during training. Furthermore, during training for all ϵ values we have selected $\bar{\eta}_{\text{start}} = 1.0$ and $\bar{\eta}_{\text{end}} = 0.9$. Additionally, $\underline{\eta}_{\text{end}} = 0.1$ is used during training, with $\underline{\eta}_{\text{start}} = 0.1$ for $\epsilon = 2/255$ (no ramp down), $\underline{\eta}_{\text{start}} = 0.3$ for $\epsilon = 8/255$, $\underline{\eta}_{\text{start}} = 0.4$ for $\epsilon = 12/255$, and $\underline{\eta}_{\text{start}} = 0.5$ for $\epsilon = 16/255$. The intuition behind these parameters selection lies in Remark 2, as large η values promote the abstain option more, so for large ϵ, we start with larger $\underline{\eta}_{\text{start}}$ as well. Reported numbers in Tabel 1 correspond to $\lambda_1 = 1$ for all ϵ values, and $\lambda_2 = 3.0$ for $\epsilon = 2/255$, $\lambda_2 = 2.9$ for $\epsilon = 8/255$, and $\lambda_2 = 3.1$ for $\epsilon = 16/255$ to insure similar natural accuracy for fair comparison against other methods.

B.1 Empirical attack success rate using PGD attacks

In order to obtain empirical attack success on the trained networks, adversarial perturbations are sought by solving

$$\max_{\boldsymbol{\delta} \in \Delta_\epsilon} \left(\max_{i \neq a, y} z_{L,i} - \max\{z_{L,y}, z_{L,a}\} \right) \tag{36}$$

Network layers	
Conv 64 $3 \times 3 + 1$	
Conv 64 $3 \times 3 + 1$	
Conv 128 $3 \times 3 + 2$	
Conv 128 $3 \times 3 + 1$	
Conv 128 $3 \times 3 + 1$	
Fully Conn.	512
# hidden	230K
# params.	17M

Table 2: Network architecture. Similar to the Large network used in (Gowal et al., 2018)

This attack is indeed an adaptive attack as it aims at circumventing the detection while trying to cause misclassification (Tramer et al., 2020). Perturbations are sought by maximizing this objective using PGD with 200-steps for mnist and 500-steps for CIFAR-10 Madry et al. (2017), with 10 random restarts. It is interesting to note that the achieved attack success rate in Table 1 is well below the verified robust error, further implying the effectiveness of incorporation of the detection mechanism as the true robustness of the system against practical adaptive PGD attacks are considerably stronger in comparison to robust classification without detection.