SUPPLEMENTARY MATERIAL
DOES THE HALF ADVERSARIAL ROBUSTNESS REPRESENT THE WHOLE? IT DEPENDS ... A THEORETICAL PERSPECTIVE OF SUBNETWORK ROBUSTNESS

Anonymous authors
Paper under double-blind review

1 SEMIROBUSTNESS GUARANTEES

1.1 PROOF OF THEOREM 1

First, we show the leftward implication, that if the layers $f^{(j)}, f^{(j-1)}, \ldots, f^{(1)}$ are semirobust, then $F^{(j)}$ is semirobust. This is proved becasue $F^{(j)}$ is $f^{(j)}(x^{(j-1)})$ where $x^{(j-1)} = f^{(j-1)} \circ \ldots \circ f^{(1)}$. Therefore, if $f^{(j)}$ is semirobust, regardless of whether any of $f^{(j-1)}, \ldots, f^{(1)}$ is semirobust, then $F^{(j)}$ is also semirobust.
Next, we show the rightward implication, that if $F^{(j)}$ is semirobust, then $f^{(j)}, f^{(j-1)}, \ldots, f^{(1)}$ are semirobust. If $F^{(j)}$ is semirobust then

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_j \circ F^{(j)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_j \circ f^{(j)}(\mathbf{X}+\delta)\right] \geq \gamma_j \tag{1}$$

This implies that $f^{(j)}$ is semirobust. Now let $G_{j-1} = G_j \circ f^{(j)}$, then

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_j \circ F^{(j)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_j \circ f^{(j)} \circ F^{(j-1)}(\mathbf{X}+\delta)\right] \tag{2}$$

$$= \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_{j-1} \circ F^{(j-1)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_{j-1} \circ f^{(j-1)}(\mathbf{X}+\delta)\right] \geq \gamma_j, \tag{3}$$

This implies that $f^{(j-1)}$ is semirobust. By induction, it's shown that the other layers $f^{(j-2)}, \ldots, f^{(1)}$ are also semirobust.

1.2 PROOF OF LEMMA 1

Let $f^{(n-1)} = g \in \mathcal{L}_{n-1}$ and $f^{(n)} = h \in \mathcal{L}_n$. As $f^{(n-1)}$ is semirobust,

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_{n-1} \circ g(\mathbf{X}+\delta)\right] \geq \gamma_{n-1} \text{ and } \sum_y \pi(y) I\left(g_\delta; h_\delta | y\right) \geq \rho, \tag{4}$$

and after simplification, we have

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y \cdot G_{n-1} \circ g(\mathbf{X}+\delta)\right] = \sum_y y \cdot \pi(y) \int \inf_{\delta\in S} D(x|y) \cdot G_{n-1} \circ g(x+\delta) d\mathbf{x} \tag{5}$$

with $\pi(y)$ being the prior of y, and $D(x|y)$ being the probability density function of x and y. Let $g_\delta = g(\mathbf{x}+\delta) \in \mathcal{L}_{n-1}$, with components $g_\delta^{(i)} = g_i(\mathbf{x}+\delta)$ such that $\mathbf{x} = (\mathbf{x^{(1)}}, \ldots, \mathbf{x^{(d)}}) \in \mathcal{R}^{\mathbf{d}}$, and $g = (g_1, \ldots, g_m) \in \mathcal{L}_{n-1}$. Note that the multivariate transformation g_i is one to one; hence, the transformation is invertible and can be solved for the equation $x^{(i)} + \delta^{(i)} = g_i^{-1}(g_\delta)$.

Thus, the last line in (5) equals

$$\sum_y y \cdot \pi(y) \int \inf_{\delta \in S} D(g^{-1}(g_\delta) - \delta | y) \cdot G_{n-1} \circ g_\delta \; |J| \; dg_\delta$$

where $|J|$ denotes the absolute value of the determinant of the Jacobian J. In addition, using the probability density for a function of a random variable, we can write:

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta \in S} y \cdot G_{n-1} \circ g(\mathbf{X}+\delta)\right] = \sum_y y \cdot \pi(y) \int \inf_{\delta \in S} p(g_\delta|y) \cdot G_{n-1} \circ g_\delta \; dg_\delta \tag{6}$$

By simplifying mutual information $I(g_\delta; h_\delta|y)$, using $\log(x) \leq x+1$, and recalling assumption **B1**:

$$\sum_y \pi(y) \iint \inf_{\delta \in S} p(g_\delta, h_\delta|y) \left[\frac{p(g_\delta, h_\delta|y)}{p(g_\delta|y)p(h_\delta|y)}\right] dg_\delta dh_\delta + 1 \geq \rho \tag{7}$$

To show that $f^{(n)}$ is γ_n-semirobust, we prove

$$\gamma_n \leq \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta \in S} y \cdot h(\mathbf{X}+\delta)\right].$$

Set $\gamma_n \leq \gamma_{n-1} + \rho$, hence we need to show that

$$\gamma_{n-1} + \rho \leq \sum_y y \cdot \pi(y) \int \inf_{\delta \in S} p(h_\delta|y) \cdot h_\delta dh_\delta \tag{8}$$

Following (7) and semirobustness for $f^{(n-1)}$, the inequality (8) can be transformed into

$$\sum_y \pi(y) \iint \inf_{\delta \in S} p(g_\delta, h_\delta|y) \left(y \cdot G_{n-1} \circ g_\delta + \frac{p(g_\delta, h_\delta|y)}{p(g_\delta|y)p(h_\delta|y)} - y \cdot h_\delta + 1\right) dg_\delta dh_\delta \leq 0$$

and subsequently into

$$\sum_y \pi(y) \mathbb{E}_{p(g_\delta, h_\delta|y)}[\inf_{\delta \in S} y \cdot (G_{n-1} \circ g_\delta - h_\delta)] \leq -(1+U), \tag{9}$$

which holds true recalling the assumption **B2**. This concludes the proof of the lemma.

1.3 Proof of Theorem 2

To prove Theorem 2, given that

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta \in S} y \cdot G_a \circ h_a(\mathbf{X}+\delta)\right] \geq \gamma_a \;\text{ and }\; \sum_y \pi(y) I\left(h_{\delta,a}; h_{\delta,a+1}|y\right) \geq \rho_{a+1}, \tag{10}$$

We need to show from the inequalities above that for $\gamma_{a+1} \leq \gamma_a + \rho_{a+1}$

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta \in S} y \cdot G_{a+1} \circ h_{a+1}(\mathbf{X}+\delta)\right] \geq \gamma_{a+1} \tag{11}$$

Under assumptions **A1** and **A2** for $j = a+1$, we can simplify (10) and (11). We then need to show

$$\sum_y \pi(y) \mathbb{E}_{p(g_\delta, h_{\delta,a+1}|y)}[\inf_{\delta \in S} y \cdot (G_a \circ h_{\delta,a} - G_{a+1} \circ h_{\delta,a+1})] \leq -(1+U_{a+1}) \tag{12}$$

The above holds true recalling the assumption **A2** and $\pi(y)$ being non-negative. Hence, $f^{(a+1)}$ is γ_{a+1}-semirobust. And because $f_a = F^{(a)}$ is γ_a-semirobust, then according to Theorem 1, $F^{(a+1)}$ is also γ_{a+1}-semirobust.

Similarly, since $f_a = F^{(a+1)}$ is γ_{a+1}-semirobust, and by assumptions **A1** and **A2** for $j = a+2$, it is implied that $f^{(a+2)}$ is γ_{a+2}-semirobust. Recursively, it can be shown that all layers in f_b, i.e. $f^{(a+1)}, \ldots, f^{(n)}$, are γ_j-semirobust for $j = a+1, \ldots, n$ respectively. Then, according to Theorem 1, f_b is γ_b-semirobust where $\gamma_b \leq \gamma_a + \sum_{j=a+1}^{b} \rho_j$, proving Theorem 2.

1.4 PROOF OF LEMMA 2

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.f^{(n)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\sum_{i=1}^{n-1}\lambda_i^T.f^{(i)}(\mathbf{X}+\delta)\right] \tag{13}$$

$$\sum_{i=1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_i^T f^{(i)}(\mathbf{X}+\delta)\right] = \sum_{i=1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_i^T.F^{(i)}(\mathbf{X}+\delta)\right]. \tag{14}$$

The last equality holds true because the noises are added to the input $\mathbf{X}$ and since in feedforward network, each layer is a function of the previous layer therefore $f^{(i)}(\mathbf{X}+\delta) = F^{(i)}(\mathbf{X}+\delta)$.

Next, by letting $G_i = \lambda_i^T$, then we have

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.f^{(n)}(\mathbf{X}+\delta)\right] = \sum_{i=1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_i\circ F^{(i)}(\mathbf{X}+\delta)\right] \geq \sum_{i=1}^{n-1}\gamma_i := \gamma_n. \tag{15}$$

The first inequality is true because of Theorem 1. This concludes that $f^{(n)}$ is γ_n-semirobust.

1.5 PROOF OF THEOREM 3

In Theorem 3, note that if $f_b = f^{(n-1)}$ and $f_a = f_{n-1} = F^{(1,n-1)}$, then it turns to Lemma 2. We prove the theorem where $f_b = F^{(n-1,n)}$ and $f_a = F^{(1,n-2)}$, and the general case f_a and f_b can be shown similarly by extension. Let $G_b : \mathcal{L}_b \mapsto \mathcal{Y}$ be a function that maps layer f_b to the output y. Proof of the case where $f_b = F^{(n-1,n)}$ and $f_a = F^{(n-2)}$:

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.F^{(n-1,n)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.f^{(n)}(\mathbf{X}+\delta)\right] \tag{16}$$

Given that $f^{(n)}$ is a linear combination of all the other layers, with λ_{in}^T mapping $f^{(i)}$ to y,

$$= \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\sum_{i=1}^{n-1}\lambda_{in}^T.f^{(i)}(\mathbf{X}+\delta)\right] = \sum_{i=1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{in}^T.f^{(i)}(\mathbf{X}+\delta)\right] \tag{17}$$

$$= \sum_{i=1}^{n-2}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{in}^T.f^{(i)}(\mathbf{X}+\delta)\right] + \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{n-1(n)}^T.f^{(n-1)}(\mathbf{X}+\delta)\right] \tag{18}$$

Let $G_i = \lambda_{in}^T$, and let α be the second term in (18). Then, using Theorem 1,

$$= \sum_{i=1}^{n-2}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_i\circ f^{(i)}(\mathbf{X}+\delta)\right] + \alpha \geq \sum_{i=1}^{n-2}\gamma_i + \alpha = \gamma_a + \alpha \tag{19}$$

where $\gamma_a = \sum_{i=1}^{n-2}\gamma_i$. Now simplifying α, given that $f^{(n-1)}$ is a linear combination of the layers before it, with $\lambda_{i(n-1)}^T$ mapping $f^{(i)}$ to $f^{(n-1)}$:

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{n-1(n)}^T.f^{(n-1)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{n-1(n)}^T.\sum_{i=1}^{n-2}\lambda_{i(n-1)}^T.f^{(i)}(\mathbf{X}+\delta)\right] \tag{20}$$

$$= \sum_{i=1}^{n-2}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda_{n-1(n)}^T.\lambda_{i(n-1)}^T.f^{(i)}(\mathbf{X}+\delta)\right]. \tag{21}$$

Let $\tilde{G}_i = \lambda^T_{n-1(n)}.\lambda^T_{i(n-1)}$. Then, using Theorem 1,

$$= \sum_{i=1}^{n-2} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\tilde{G}_i \circ f^{(i)}(\mathbf{X}+\delta)\right] \geq \sum_{i=1}^{n-2}\gamma_i = \gamma_a \tag{22}$$

With both terms simplified, $\gamma_a + \alpha \geq \gamma_a + \gamma_a = \gamma_b$. Therefore, f_b is semirobust. This proof can be extended to any other combination of f_a and f_b. Let's show the case where $f_b = F^{(a+1,n)}$ and $f_a = F^{(a)}$:

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.F^{(a+1,n)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.f^{(n)}(\mathbf{X}+\delta)\right] \tag{23}$$

Given that $f^{(n)}$ is a linear combination of all the other layers, with λ^T_{in} mapping $f^{(i)}$ to y,

$$= \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\sum_{i=1}^{n-1}\lambda^T_{in}.f^{(i)}(\mathbf{X}+\delta)\right] = \sum_{i=1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.f^{(i)}(\mathbf{X}+\delta)\right] \tag{24}$$

$$= \sum_{i=1}^{a}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.f^{(i)}(\mathbf{X}+\delta)\right] + \sum_{i=a+1}^{n-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.f^{(i)}(\mathbf{X}+\delta)\right] \tag{25}$$

Let $G_i = \lambda^T_{in}$, and let

$$\alpha_i = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.f^{(i)}(\mathbf{X}+\delta)\right], i = a+1,\ldots,n-1. \tag{26}$$

Then, by using Theorem 1 again we have,

$$= \sum_{i=1}^{a}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_i \circ f^{(i)}(\mathbf{X}+\delta)\right] + \sum_{i=a+1}^{n-1}\alpha_i \geq \sum_{i=1}^{a}\gamma_i + \sum_{i=a+1}^{n-1}\alpha_i = \gamma_a + \sum_{i=a+1}^{n-1}\alpha_i \tag{27}$$

where $\gamma_a = \sum_{i=1}^{a}\gamma_i$. Next we show that $\sum_{i=a+1}^{n-1}\alpha_i \geq \gamma_a\Big((n-1-a)(n-a)/2\Big)$, and conclude the proof by setting $\gamma_b := \gamma_a + \gamma_a\Big((n-1-a)(n-a)/2\Big)$. Now by the assumption that

$$f^{(i)} = \sum_{\ell=1}^{i-1}\lambda^T_{\ell i}.f^{(\ell)}, \tag{28}$$

with $\lambda^T_{\ell i}$ mapping $f^{(\ell)}$ to $f^{(i)}$, then simplifying α_i yields

$$\alpha_i = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.\sum_{\ell=1}^{i-1}\lambda^T_{\ell i}.f^{(\ell)}(\mathbf{X}+\delta)\right] = \sum_{\ell=1}^{i-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.\lambda^T_{\ell i}.f^{(\ell)}(\mathbf{X}+\delta)\right], \tag{29}$$

for $i = a+1,\ldots,n-1$. Therefore we have

$$\sum_{i=a+1}^{n-1}\alpha_i = \sum_{i=a+1}^{n-1}\sum_{\ell=1}^{i-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.\lambda^T_{in}.\lambda^T_{\ell i}.f^{(\ell)}(\mathbf{X}+\delta)\right]$$
$$= \sum_{i=a+1}^{n-1}\sum_{\ell=1}^{i-1}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell n} \circ f^{(\ell)}(\mathbf{X}+\delta)\right], \tag{30}$$

where $G_{\ell n} := \lambda^T_{in}.\lambda^T_{\ell i}$. Under the assumption (28), we know that for $i = a+1,\ldots,n-1$,

$$\sum_{\ell=1}^{a}\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell n} \circ f^{(\ell)}(\mathbf{X}+\delta)\right] \geq \gamma_a, \tag{31}$$

Therefore,

$$\sum_{i=a+1}^{n-1} \alpha_i \geq (n-1-a)\gamma_a + \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{(a+1)n} \circ f^{(a+1)}(\mathbf{X}+\delta)\right]$$
$$+ \sum_{\ell=a+1}^{a+2} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell n} \circ f^{(\ell)}(\mathbf{X}+\delta)\right] + \ldots + \sum_{\ell=a+1}^{n-2} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell n} \circ f^{(\ell)}(\mathbf{X}+\delta)\right] \tag{32}$$

Without loss of generality, assume that $G_{\ell n} = G_\ell$ for all $\ell = a+1, \ldots, n-2$. This simplifies (32) as

$$\sum_{i=a+1}^{n-1} \alpha_i \geq (n-1-a)\gamma_a + \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{(a+1)} \circ f^{(a+1)}(\mathbf{X}+\delta)\right]$$
$$+ \sum_{\ell=a+1}^{a+2} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell} \circ f^{(\ell)}(\mathbf{X}+\delta)\right] + \ldots + \sum_{\ell=a+1}^{n-2} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell} \circ f^{(\ell)}(\mathbf{X}+\delta)\right]. \tag{33}$$

Below we show that if $f_a = F^{(1,a)}$ is semi-robust and $f^{(a+1)}$ is a linear combination of layers $f^{(1)}, \ldots, f^{(a)}$, with $\lambda_{\ell(a+1)}$ mapping $f^{(\ell)}$ to $f^{(a+1)}$, then $f_{a+1} = F^{(1,a+1)}$ is a semi-robust feature:

$$\mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{(a+1)} \circ f^{(a+1)}(\mathbf{X}+\delta)\right] = \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{(a+1)} \circ \sum_{\ell=1}^{a} \lambda^T_{\ell(a+1)} f^{(\ell)}(\mathbf{X}+\delta)\right]$$
$$= \sum_{\ell=1}^{a} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{(a+1)} \circ \lambda^T_{\ell(a+1)} f^{(\ell)}(\mathbf{X}+\delta)\right] = \sum_{\ell=1}^{a} \mathbb{E}_{(\mathbf{X},y)\sim D}\left[\inf_{\delta\in S} y.G_{\ell} \circ f^{(\ell)}(\mathbf{X}+\delta)\right], \tag{34}$$

where $G_\ell = G_{(a+1)} \circ \lambda^T_{\ell(a+1)}$. Since f_a is semi-robust, all layers $f^{(1)}, \ldots, f^{(a)}$ are semi-robust. Hence, the right-hand side in (34) is greater than or equal to $\sum_{\ell=1}^{a} \gamma_\ell = \gamma_a$.

Consequently, with the same methodology, this can be extended to the following: if $f_a = F^{(1,a)}$ is semi-robust, and $f^{(a+\ell)}$ is a linear combination of layers $f^{(1)}, \ldots f^{(a+\ell-1)}$, then $f_{a+\ell} = F^{(1,a+\ell)}$ for $\ell = 1, \ldots, n-2-a$ is semi-robust. This implies that the Ineq. (32) is lower-bounded by

$$(n-1-a)\gamma_a + \gamma_a + \underbrace{\sum_{\ell=a+1}^{a+2} \gamma_a}_{2\times\gamma_a} + \underbrace{\sum_{\ell=a+1}^{a+3} \gamma_a}_{3\times\gamma_a} + \ldots + \underbrace{\sum_{\ell=a+1}^{n-2} \gamma_a}_{(n-2-a)\times\gamma_a}, \tag{35}$$

which is equal to

$$(n-1-a)\gamma_a + \gamma_a \sum_{j=1}^{n-2-a} j = (n-1-a)\gamma_a + \gamma_a\Big((n-2-a)(n-1-a)\Big/2\Big). \tag{36}$$

This proves that $\sum_{i=a+1}^{n-1} \alpha_i \geq \gamma_a\Big((n-1-a)(n-a)/2\Big) = \gamma_b$. This completes the proof.

2 Experimental Setup Details

2.1 Attacks

Adversarial attacks for Algorithm 1 of the main paper were produced using the Adversarial Robustness Toolbox (ART) library Nicolae et al. (2018) using the default parameters with the exception of those for which specific values are provided here and in the Experimental Setup section. Notably we use a simplified approach for the attacks, applying perturbations across the full dataset rather than per-batch. The adversarial data is then stored so that we can compare different hyperparameter settings on the same perturbed data. With this setting we still observe significant drops in accuracy on non-robust networks and observe the notable behavior of semi-robust networks.

2.2 ALGORITHM SETTINGS

The pretraining of models on the normal CIFAR-10 and CIFAR-100 datasets uses settings from DeVries & Taylor (2017), with an initial learning rate of 0.1 which is reduced by a factor of 10 at 60, 120, and 160 epochs for a total of 200 training epochs.
For Imagenette we start with the weights of networks provided in the Torchvision library at https://github.com/pytorch/vision for models pretrained on Imagenet. The output layer is changed to one with 10 classes for Imagenette, and the final layer was finetuned for 60 epochs.
Data is preprocessed by subtracting the mean and dividing by the standard deviation of each channel, and as such our input data has a range slightly larger than $[0, 1]$. This means ϵ values don't directly match with pixel values when processed as $\frac{\epsilon}{255}$. To offset this we have included multiple ϵ values and attacks of differing strengths.
Training of the full model on adversarially attacked datasets is done for 120 epochs at an initial learning rate of 0.01, decreasing by a factor of 10 at 40 and 80 epochs. Training of the subnetwork for trials is done for 20 epochs. All training is done using checkpoints on validation accuracy. A batch size of 512 is used for CIFAR data, while a batch size of 32 is used for Imagenette. Our experiments use $T = 10$ and $k = 1e^{-16}$ for both algorithms.

3 ADDITIONAL EXPERIMENTS

Algorithm 1 Learning Hyperparameter λ

Do regular training of $F^{(n)}$
Do adversarial training of $F^{(n)}$ as (f_a^*, f_b^*)
Store test accuracy of adversarial training (f_a^*, f_b^*) as Acc^*
Get output of $f^{(j)*}$ for each j in $a+1, \ldots, n$
Freeze f_a^*
Replace densely connected layer f_b with the linear-combination $f^{(j)} = \sum_{i=1}^{j-1} \lambda_{ij}^T . f^{(i)}$, $j = a+1, \ldots, n$
Set k to be as small as possible
for $t = 1, \ldots,$ T **do**
  for $e = 1, \ldots,$E **do**
    for each batch **do**
      Loss = $\sum_{j=a+1}^{n} ||f^{(j)*} - f^{(j)}||$
      Solve for lambda using layer outputs of f_a^* and f_b^*
      Store test accuracy of $(f_a^*, \widetilde{f_b})$ as Acc_t^e
    end
    if $Acc^* - Acc_t^e \leq k$ or $e \geq E$ **then**
      Break out of epoch loop and store Acc_t^e
    end
  end
end
$\widetilde{Acc}$ = largest Acc_t^e
Report $\widetilde{Acc}$

Here we display additional results to supplement the ideas presented in the main paper. We first show the impact of applying the linearity constraint on the dependency between layers in the subnetworks. In Algorithm 1 of the SM, each layer in f_b is a linear combination of the layer outputs in f_a^*, and we aim to reach Acc^* on the adversarial data using the frozen f_a^* subnetwork. We demonstrate our success in doing so utilizing a straightforward linear algebra approach to directly solve for λ in one epoch in Table 1 using AlexNet and CIFAR-10 and differing attack types. In order to achieve these results, we record the outputs from the all layers in the frozen f_a^* subnetwork as $\mathbb{F}_a$ and utilize the loss function Loss = $\sum_{j=a+1}^{n} ||f^{(j)*} - f^{(j)}||$ to set $\mathbb{F}^{(j)} = \mathbb{F}^{(j)*}$ and solve for the λ which minimizes the loss in the following steps:

$$\begin{aligned} \mathbb{F}_a^* \cdot \lambda_{i,j} &= \mathbb{F}^{(j)} \\ \mathbb{F}_a^* \cdot \lambda_{i,j} &= \mathbb{F}^{(j)*} \\ \mathbb{F}_a^{*-1} \cdot \mathbb{F}_a^* \cdot \lambda_{i,j} &= \mathbb{F}_a^{*-1} \cdot \mathbb{F}^{(j)} \\ \mathbb{I} \cdot \lambda_{i,j} = \lambda_{i,j} &= \mathbb{F}_a^{*-1} \cdot \mathbb{F}^{(j)} \end{aligned}$$

We describe the steps in performing these experiments in Algorithm 1 of the SM.

Table 1 shows the results of applying Algorithm 1 to AlexNet for a single layer of f_b. Noteably, we limit this experiment to a single layer and use AlexNet, because although directly solving for λ in this way highlights that such a λ exists, the calculation involved would become intractable on larger datasets where you need to take the inverse of a matrix of all intermediate activations of the network. The table shows that we can represent f_b under a linear assumption by replacing it with a

Table 1: Linear combination accuracy with semirobust network on CIFAR-10 with AlexNet

Attack	Acc^* (%)	$\widetilde{Acc}$ (%)	Diff. (%)	$Acc(\lambda_{random})$
FGSM	70.65	70.67	0.02	9.08
I-FGSM	71.22	71.2	-0.02	10.42
PGD	70.9	70.92	0.02	8.62
C&W	67.04	67.01	-0.03	8.9

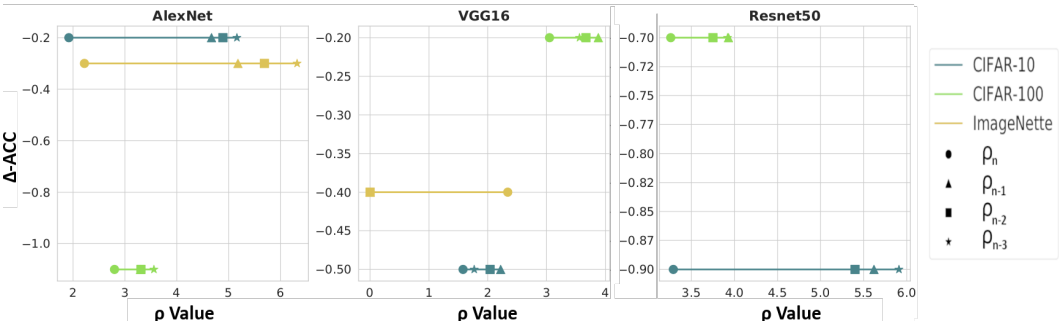

Figure 1: Connectivity values vs performance differences are plotted for differing datasets on each architecture with the CW attack when f_b is 4 trainable layers.

linear combination of the semirobust f_a^*. We provide the accuracies of the fully-robust network Acc^*, the accuracy using the linear combination replacement of f_b ($\widetilde{Acc}$), and the accuracy when using a random linear combination of f_a^* as a negative control $Acc(\lambda_{random})$.

Hyperparameter Analysis Experiments To observe the behavior of ρ when holding certain hyperparameters constant, we ran the experiments in Figs. 1, 2, and 3. For each experiment, f_b = 4 layers and for Fig. 2, we use $\epsilon = \frac{8}{255}$. Observing these results doesn't show any clear pattern except that CIFAR-100 consistently has a much narrower range of ρ values than the other two datasets. varying attack type or network has little to no consistent pattern in the effect on ρ on the other hand.
We extend these results further by running each combination of attack type and network on the CIFAR-10 dataset with an $\epsilon = \frac{8}{255}, \frac{16}{255}$, and $\frac{32}{255}$. The results of these experiments are shown in Tables 2, 3, and 4. We report the accuracy of the non-adversarially-trained network (f_a, f_b) as Acc_{norm}, the same model's accuracy on the adversarially attacked data Acc_{adv}, and the remaining notations are defined in the main text for Table 1. Once again we see little noticeable impact of changing the attack type the ϵ value on ρ. Again, CIFAR-100 shows the narrowest range of ρ values, but also noticeably we observe that for $f_b = 4$ layers, the difference in accuracy from Acc^* is negligible even before

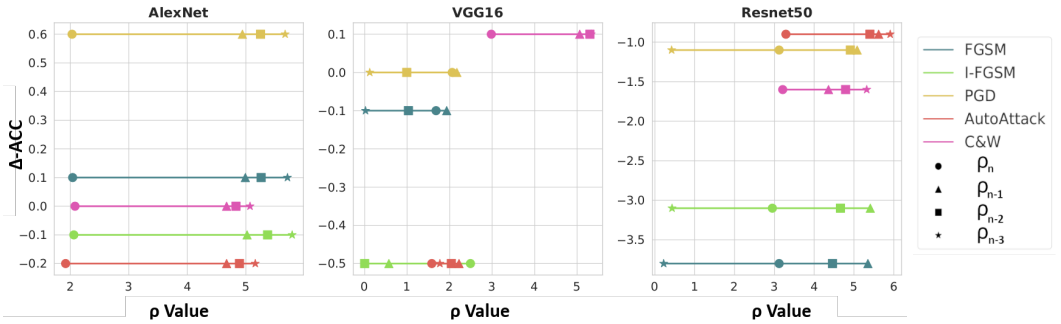

Figure 2: Connectivity values vs performance differences are plotted for differing attacks and networks for CIFAR-10 when f_b is 4 trainable layers.

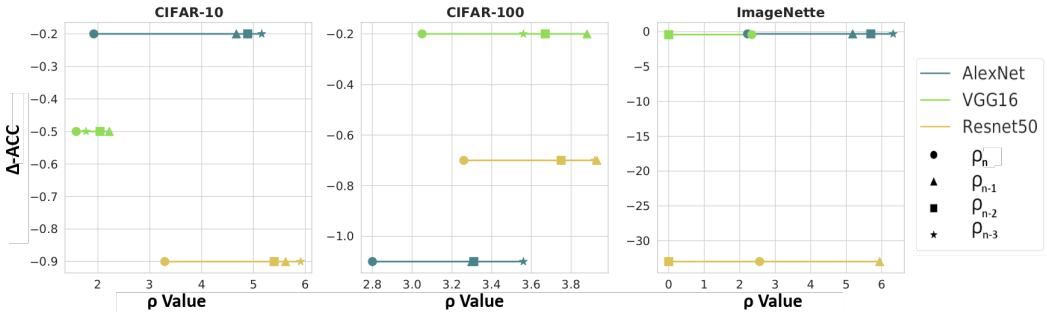

Figure 3: Connectivity values vs performance differences are plotted for differing networks and datasets with the C&W attack when f_b is 4 trainable layers.

Table 2: Subnetwork training on CIFAR-10 with differing attacks

Model	Attack	ϵ	ACC_{norm}	ACC_{adv}	Acc^*	Acc_{sr}	$\widetilde{Acc}$	Diff.	ρ_n	ρ_{n-1}	ρ_{n-2}	ρ_{n-3}
AlexNet	FGSM	8	75.6	63.0	72.8	34.4	72.9	0.07	2.04	4.99	5.26	5.71
	I-FGSM	8	75.6	62.6	73.0	43.8	72.9	-0.13	2.06	5.02	5.37	5.79
	PGD	8	75.6	62.6	72.5	38.8	73.2	0.61	2.03	4.94	5.25	5.67
	Autoattack	8	75.6	15.1	64.7	19.7	64.5	-0.22	1.92	4.67	4.89	5.16
VGG16	FGSM	8	94.3	79.6	90.0	90.1	89.9	-0.09	1.68	1.93	1.03	0.02
	I-FGSM	8	94.3	78.5	90.5	90.6	90.1	-0.48	2.49	0.57	0.00	0.00
	PGD	8	94.3	78.5	89.3	89.4	89.3	-0.04	2.06	2.17	0.99	0.12
	Autoattack	8	94.3	44.9	79.6	79.7	79.2	-0.46	1.58	2.22	2.04	1.77
ResNet50	FGSM	8	93.8	82.6	88.7	88.7	84.8	-3.82	3.12	5.35	4.46	0.22
	I-FGSM	8	93.8	80.9	88.0	88.0	84.9	-3.10	2.95	5.41	4.66	0.43
	PGD	8	93.8	80.7	88.1	88.1	87.0	-1.09	3.12	5.08	4.91	0.42
	Autoattack	8	93.8	50.4	76.1	76.3	75.2	-0.9	3.29	5.62	5.40	5.91

addition training of f_b. Additionally, many of the experiments have ρ values near or equal to 0 in the first few layers of f_b. We note this behavior in many of the 1 and 4-layer runs we've performed but are unclear as to the underlying reason that these more trivial cases should have diminishing ρ values while more challenging experiments with larger f_b seldom appear to have this behavior as with those in Table 5 and 6.

Varying f_b Size Experiment Data Tables 5 and 6 provide the data for the figures in the paper. For Table 5 we run each network type on datasets perturbed by AutoAttack with $\epsilon = \frac{8}{255}$. The size of f_b varied for each network to ensure that we saw a substantial decrease in accuracy from Acc^* to Acc_{sr}. This way, the value of $\widetilde{Acc}$ couldn't be trivially due to an insufficient disruption of network accuracy by removing the robustness of f_b.
We run a similar setup in Table 6, changing the size of f_b when training ResNet50 on CIFAR-10 data perturbed by AutoAttack with $\epsilon = \frac{8}{255}$. This experiment showed a consistent increase of ρ values deeper into the network from the output layer, with increasingly sharp drops in performance between Acc^* and Acc_{sr}. Despite these challenges to the performance, the subnetwork training was consistently able to reach an $\widetilde{Acc}$ value within $1 - 2\%$ of Acc^*.

REFERENCES

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben

Table 3: Subnetwork training on CIFAR-10 with differing attacks

Model	Attack	ϵ	ACC_{norm}	ACC_{adv}	Acc^*	Acc_{sr}	$\widetilde{Acc}$	Diff.	ρ_n	ρ_{n-1}	ρ_{n-2}	ρ_{n-3}
AlexNet	FGSM	16	75.6	51.1	70.2	36.7	69.8	-0.37	2.00	4.96	5.39	5.80
	I-FGSM	16	75.6	50.0	69.5	45.9	70.0	0.52	2.02	4.97	5.41	5.75
	PGD	16	75.6	50.1	70.6	33.6	70.6	0.00	2.06	5.00	5.35	5.76
	Autoattack	16	75.6	6.7	61.5	13.8	61.2	-0.28	1.90	4.60	4.76	5.08
VGG16	FGSM	16	94.3	72.7	88.0	88.1	87.5	-0.50	2.07	0.59	0.00	0.00
	I-FGSM	16	94.3	67.8	89.4	89.4	89.1	-0.28	1.71	2.75	2.53	1.95
	PGD	16	94.3	67.8	90.0	90.1	89.7	-0.28	2.23	2.08	0.07	0.00
	Autoattack	16	94.3	32.3	75.8	73.6	75.8	0.02	3.22	5.19	5.08	5.20
ResNet50	FGSM	16	93.8	74.4	86.9	86.9	86.0	-0.89	3.09	5.18	4.75	0.00
	I-FGSM	16	93.8	68.7	86.8	86.8	84.8	-1.97	3.02	5.26	4.72	0.00
	PGD	16	93.8	68.8	87.3	87.3	85.4	-1.89	3.10	5.22	4.93	0.45
	Autoattack	16	93.8	36.9	78.2	78.1	76.8	-1.39	3.30	5.32	4.76	5.57

Table 4: Subnetwork training on CIFAR-10 with differing attacks

Model	Attack	ϵ	ACC_{norm}	ACC_{adv}	Acc^*	Acc_{sr}	$\widetilde{Acc}$	Diff.	ρ_n	ρ_{n-1}	ρ_{n-2}	ρ_{n-3}
AlexNet	FGSM	32	75.6	34.4	67.3	21.7	67.2	-0.11	2.07	4.86	5.40	5.93
	I-FGSM	32	75.6	32.4	68.1	18.9	68.6	0.48	2.05	4.93	5.33	5.81
	PGD	32	75.6	32.3	68.1	23.1	67.9	-0.22	2.05	4.92	5.40	5.90
	Autoattack	32	75.6	1.3	61.5	12.1	61.5	-0.02	2.02	4.59	4.67	5.23
	C&W	-	75.6	19.6	66.9	21.7	66.9	-0.02	2.08	4.67	4.83	5.07
VGG16	FGSM	32	94.3	67.1	86.6	86.6	86.3	-0.24	1.73	1.74	1.57	0.72
	I-FGSM	32	94.3	52.3	88.9	88.8	88.3	-0.52	2.15	0.58	0.00	0.00
	PGD	32	94.3	52.3	89.0	89.0	88.6	-0.37	2.50	0.91	0.00	0.00
	Autoattack	32	94.3	20.2	73.6	73.2	73.7	0.07	3.13	5.11	5.11	4.99
	C&W	-	94.3	18.9	80.1	80.3	80.2	0.09	2.98	5.06	5.30	5.33
ResNet50	FGSM	32	93.8	65.5	85.0	85.0	84.0	-1.09	3.01	5.57	4.84	0.54
	I-FGSM	32	93.8	50.1	85.7	85.6	84.9	-0.78	3.16	5.05	5.34	5.03
	PGD	32	93.8	50.1	85.5	85.5	84.0	-1.45	3.16	5.44	4.79	4.31
	Autoattack	32	93.8	26.2	75.6	75.8	75.3	-0.33	3.26	5.42	4.92	6.09
	C&W	-	93.8	15.1	78.3	78.4	76.6	-1.65	3.21	4.36	4.79	5.32

Edwards. Adversarial robustness toolbox v1.2.0. CoRR, 1807.01069, 2018. URL https://arxiv.org/pdf/1807.01069.

Table 5: Subnetwork training on AutoAttack for $\epsilon = \frac{8}{255}$ with differing networks and datasets

Model	Dataset	f_b layers	ACC_{norm}	ACC_{adv}	Acc^*	Acc_{sr}	$\widetilde{Acc}$	ρ_n	ρ_{n-3}	ρ_{n-7}	ρ_{n-11}	ρ_{n-15}
AlexNet	CIFAR-10	4	75.6	15.1	64.7	19.7	64.5	1.92	5.16	-	-	-
	CIFAR-100	4	43.3	7.7	33.6	16.9	32.3	2.79	3.55	-	-	-
	Imagenette	4	71.4	37.8	75.3	67.3	74.0	2.17	6.24	-	-	-
VGG16	CIFAR-10	12	94.3	44.9	79.0	63.9	76.6	3.22	5.27	6.83	7.60	-
	CIFAR-100	12	65.9	34.8	54.2	38.5	51.8	3.05	3.59	4.11	4.28	-
	Imagenette	12	99.5	35.3	91.0	26.1	86.0	2.35	6.82	7.07	7.11	-
ResNet50	CIFAR-10	16	93.8	50.4	75.8	46.7	74.7	3.22	5.93	6.69	6.70	6.48
	CIFAR-100	16	66.0	34.7	56.7	25.8	55.9	3.26	3.92	4.14	3.91	3.95
	Imagenette	16	99.6	48.1	89.2	9.5	82.3	3.05	6.17	6.61	6.58	0.00

Table 6: Subnetwork training of ResNet50 on CIFAR-10 with and with AutoAttack ($\epsilon = \frac{8}{255}$)

# f_b layers	ACC_{norm}	ACC_{adv}	Acc^*	Acc_{sr}	$\widetilde{Acc}$	Diff.	ρ_n	ρ_{n-3}	ρ_{n-7}	ρ_{n-11}	ρ_{n-15}
4	93.8	50.4	76.1	76.3	75.2	-0.9	3.29	5.91	-	-	-
8	93.8	50.4	75.5	73.8	75.1	-0.3	3.32	5.64	6.69	-	-
12	93.8	50.4	77.6	71.0	75.9	-1.7	3.35	6.06	6.45	6.13	-
16	93.8	50.4	75.8	46.7	74.7	-1.1	3.22	5.93	6.69	6.70	6.48