A PROOFS

A.1 PROOF OF THEOREM 1

According to (8), the update of $\tau(\tilde{\boldsymbol{\delta}}[k])$ for k-th frequency component of l_2 norm FGM adversarial perturbation at the $(t+1)$-th step is proportional to

$$(1-2\bar{\eta}_2^{(t)})|\widetilde{\nabla f}^{(t)}[k]|^2 - 2\tilde{\eta}^{(t)}|\widetilde{\nabla f}^{(t)}[k]||\tilde{\boldsymbol{x}}[k]|\cos\left(\triangle\varphi_k^{(t)}\right). \tag{13}$$

We adopt the following representations to see trends of $\cos\left(\triangle\varphi_k^{(t)}\right)$ through the training process:

$$\widetilde{\nabla f}^{(t)}[k] := \overrightarrow{c_{(t)}} = (|c_{(t)}|\cos\theta, |c_{(t)}|\sin\theta)$$
$$\tilde{\boldsymbol{x}}[k] := \overrightarrow{x_{(k)}} = (|x_{(k)}|\cos\zeta, |x_{(k)}|\sin\zeta),$$

then $|\widetilde{\nabla f}^{(t+1)}[k]||\tilde{\boldsymbol{x}}[k]|\cos\left(\triangle\varphi_k^{(t)}\right)$ will have the form of

$$\left\langle \overrightarrow{c_{(t+1)}}, \overrightarrow{x_{(k)}} \right\rangle = \left(1-\bar{\eta}^{(t)}\right)\left\langle \overrightarrow{c_{(t)}}, \overrightarrow{x_{(k)}} \right\rangle - \tilde{\eta}^{(t)}|x_{(k)}|^2 + \mathcal{O}(\eta^2)$$

at the $(t+1)$-th step. At the $(t+1)$-th step, we have

$$\triangle\tau_l^{(t+1)} - \triangle\tau_h^{(t+1)} = (1-\bar{\eta}^{(t)})(\triangle\tau_l^{(t)} - \triangle\tau_h^{(t)}) + \tilde{\eta}^{(t)}(L_x - H_x), \tag{14}$$

which means that if $\triangle\tau_l^{(t)} > \triangle\tau_h^{(t)}$ then we must have $\triangle\tau_l^{(t')} > \triangle\tau_h^{(t')}$ for all $t' \geq t$. Besides, there must be a step which leads to this condition since $L_x > H_x$. Let $L^{(t)} = \sum_{k=0}^{k_c}|\widetilde{\nabla f}[k]|^2$ and $H^{(t)} = \sum_{k=k_c+1}^{(d-1)/2}|\widetilde{\nabla f}[k]|^2$. At the $(t+1)$-th step, the changed amount of $\tau(\tilde{\boldsymbol{\delta}} \in S_l)$ is

$$\begin{aligned}
&\frac{(1-2\bar{\eta}^{(t)})L^{(t)} + \tilde{\eta}^{(t)}\triangle\tau_l^{(t)}}{(1-2\bar{\eta}^{(t)})(L^{(t)}+H^{(t)}) + \tilde{\eta}^{(t)}(\triangle\tau_l^{(t)} + \triangle\tau_h^{(t)})} - \frac{L^{(t)}}{L^{(t)}+H^{(t)}} \\
=&\tilde{\eta}^{(t)}\frac{\triangle\tau_l^{(t)}\tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_h\right) - \triangle\tau_h^{(t)}\tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_l\right)}{\sum_{k=0}^{(d-1)/2}|\widetilde{\nabla f}^{(t+1)}[k]|^2}.
\end{aligned} \tag{15}$$

If $\triangle\tau_l^{(t)} > \triangle\tau_h^{(t)}$ and $\tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_h\right) > \tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_l\right)$ then the above changed amount will be positive and gradient descent will increase $\tau\left(\tilde{\boldsymbol{\delta}} \in S_l\right)$ at this step.

A.2 PROOF OF THEOREM 2

We provide the proof for ReLU activation function. The update rules of R_k at the $(t+1)$-th step are

$$R_k^{(t+1)2} = R_k^{(t)2} - 2\tilde{\eta}^{(t)}R_k^{(t)}\cos\left(\triangle\varphi_k^{(t)}\right)$$
$$R_k^{(t)}cos\left(\triangle\varphi_k^{(t)}\right) = R_k^{(t-1)}\cos\left(\triangle\varphi_k^{(t-1)}\right) - \tilde{\eta}^{(t-1)}.$$

For any $k > k' > 0$ with $R_k^{(0)} > R_{k'}^{(0)}$, we have

$$\begin{aligned}
R_k^{(t+1)2} - R_{k'}^{(t+1)2} &= R_k^{(t)2} - R_{k'}^{(t)2} - 2\tilde{\eta}^{(t)}\left[R_k^{(t)}\cos\left(\triangle\varphi_k^{(t)}\right) - R_{k'}^{(t)}\cos\left(\triangle\varphi_{k'}^{(t)}\right)\right] \\
&= R_k^{(0)2} - R_{k'}^{(0)2} - 2\sum_{t'=0}^{t}\tilde{\eta}^{(t')}\left[R_k^{(0)}\cos\left(\triangle\varphi_k^{(0)}\right) - R_{k'}^{(0)}\cos\left(\triangle\varphi_{k'}^{(0)}\right)\right]
\end{aligned}$$

at the $(t+1)$-th step. Considering the case when $\cos\left(\triangle\varphi_k^{(0)}\right) = -\cos\left(\triangle\varphi_{k'}^{(0)}\right) = 1$ and $\tilde{\eta}^{(t')} = \tilde{\eta}_{\text{max}}$ and the condition that the left L.H.S of the above equation is larger than 0 gives us Eq.(10). On the other hand, if $R_k^{(0)}\cos\left(\triangle\varphi_k^{(0)}\right) - R_{k'}^{(0)}\cos\left(\triangle\varphi_{k'}^{(0)}\right) < 0$, then $R_k^{(t)} > R_{k'}^{(t)2}$ for all $t \geq 0$.

A.3 PROOF OF THEOREM 3

We consider the case for l_2 FGM perturbations, the case for PGD perturbations is similar. If the condition in Eq.(11) is satisfied, then at the $(t+1)$-th step,

$$\tau(\tilde{\boldsymbol{\delta}}^{(t+1)}) = \tau(\tilde{\boldsymbol{x}} \in S_l) - \zeta + \tilde{\eta}^{(t)} \frac{\triangle\tau_l^{(t)}\tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_h\right) - \triangle\tau_h^{(t)}\tau\left(\tilde{\boldsymbol{\delta}}^{(t)} \in S_l\right)}{\sum_{k=0}^{(d-1)/2} |\widetilde{\nabla f}^{(t+1)}[k]|^2} < \tau(\tilde{\boldsymbol{x}} \in S_l). \quad (16)$$

B PGD PERTURBATIONS

B.1 l_2-PGD PERTURBATIONS

The PGD update rule for finding perturbations of $\boldsymbol{x}$ with learning rate ξ at step $j+1$ is:

$$\boldsymbol{\delta}^{(j+1)} = \mathcal{P}_{\mathcal{B}(0,\epsilon)}\left[\boldsymbol{\delta}^{(j)} + \xi \frac{\partial \ell}{\partial f}^{(j)} \nabla_{\boldsymbol{x}} f^{(j)}\right], \quad (17)$$

where $\mathcal{B}(0,\epsilon)$ is a ball centered at 0 with radius ϵ in Euclidean space and $\mathcal{P}$ is the projection operator defined as

$$\mathcal{P}_{\mathcal{B}(0,\epsilon)}\left[\boldsymbol{\delta}\right] = \underset{\boldsymbol{\delta}' \in \mathcal{B}(0,\epsilon)}{\operatorname{argmin}} \|\boldsymbol{\delta}' - \boldsymbol{\delta}\|^2.$$

Note that in this part quantities removing the step script j (*e.g.* $\frac{\partial \ell}{\partial f}$ and $\nabla_{\boldsymbol{x}} f$) refers to not involving perturbations. For convenience, we use the same learning rate ξ for all steps j and instead explore the update of

$$\boldsymbol{\kappa}^{(j)} = \frac{\boldsymbol{\delta}^{(j)}}{\xi} \quad (18)$$

to the order of δ to see trends of PGD perturbations in frequency domain alongside PGD iterations since $\tau(\tilde{\boldsymbol{\delta}}[k]) = \tau(\tilde{\boldsymbol{\kappa}}[k])$. For ease of notation, we adopt the following representations: Let $\triangle\phi_k^{(j)}$ denote the difference of phases between $\widetilde{\nabla f}[k]$ and $\tilde{\boldsymbol{\kappa}}^{(j)}[k]$; let $\bar{\beta}^{(j)} = \partial\ell/\partial f + \boldsymbol{\delta}^{(j)} \cdot \nabla_{\boldsymbol{x}} f$ where $\bar{\beta}^{(0)} > 0$ and one can drive similar result for $\bar{\beta}^{(0)} < 0$; we denote

$$\triangle\bar{\tau}_l^{(j)} \triangleq 2\sum_{k=0}^{k_c} |\tilde{\boldsymbol{\kappa}}^{(j)}[k]||\widetilde{\nabla f}[k]| \cos\left(\triangle\phi_k^{(j)}\right) \text{ and } \triangle\bar{\tau}_h^{(j)} \triangleq 2\sum_{k=k_c+1}^{(d-1)/2} |\tilde{\boldsymbol{\kappa}}^{(j)}[k]||\widetilde{\nabla f}[k]| \cos\left(\triangle\phi_k^{(j)}\right),$$

where $\bar{\beta}^{(j)}\triangle\bar{\tau}_l^{(j)}$ and $\bar{\beta}^{(j)}\triangle\bar{\tau}_h^{(j)}$ are changed amounts of LFC and HFC of $|\boldsymbol{\kappa}|^2$ at the $(j+1)$-th step of PGD iteration. We provide below our result on frequency spectrum of l_2-norm PGD perturbations.

Theorem 4 (The spectral trajectory of l_2 PGD perturbation) *Iteration of the $(j+1)$-th step of PGD will change the ratio of LFC of l_2 norm PGD adversarial perturbation for a neural network (1) which satisfies $|\partial\ell/\partial f| = \epsilon^{1+\nu}$ with $0 < \nu < 1$ as follows,*

$$\tau(\tilde{\boldsymbol{\delta}} \in S_l) \leftarrow \tau(\tilde{\boldsymbol{\delta}} \in S_l) + \bar{\beta}^{(j)} \frac{\triangle\bar{\tau}_l^{(j)}\tau\left(\tilde{\boldsymbol{\delta}}^{(j)} \in S_h\right) - \triangle\bar{\tau}_h^{(j)}\tau\left(\tilde{\boldsymbol{\delta}}^{(j)} \in S_l\right)}{\sum_{k=0}^{(d-1)/2} |\tilde{\boldsymbol{\kappa}}^{(j+1)}[k]|^2}. \quad (19)$$

Remark If the two-layer neural network in (1) is trained with at least $t \geq t_0$ steps (t_0 determined by theorem 1) such that $\tau(\widetilde{\nabla f} \in S_l) > \tau(\widetilde{\nabla f} \in S_h)$, then, according to theorem 4, there exists

$$j_0 = \max\left\{0, \frac{\triangle\bar{\tau}_h^{(0)} - \triangle\bar{\tau}_l^{(0)}}{\tilde{\beta}\left(\sum_{k=0}^{k=k_c} |\widetilde{\nabla f}[k]|^2 - \sum_{k=k_c}^{k=(d-1)/2} |\widetilde{\nabla f}[k]|^2\right)}\right\}, \quad (20)$$

where $\tilde{\beta} = \max_{i\in[0,j_0]}\bar{\beta}^{(i)}$, such that $\tau(\tilde{\boldsymbol{\delta}}^{(j+1)} \in S_l) > \tau(\tilde{\boldsymbol{\delta}}^{(j)} \in S_l)$ for all $j > j_0$ if $\tau(\tilde{\boldsymbol{\delta}}^{(j)} \in S_l) < \tau(\tilde{\boldsymbol{\delta}}^{(j)} \in S_h)$.

B.1.1 PROOF OF THEOREM 4

At the (j + 1)-th step of PGD update for $\boldsymbol{\kappa}^{(j+1)}$, if

1. $\mathcal{P}^{(j)}_{\mathcal{B}(0,\epsilon)} = I$:

$$\boldsymbol{\kappa}^{(j+1)} = \boldsymbol{\kappa}^{(j)} + \frac{\partial \ell}{\partial f}\nabla_{\boldsymbol{x}} f + \frac{\partial \ell}{\partial f}\sum_r a_r \sigma'' \boldsymbol{\delta}^{(j)} \cdot \boldsymbol{W}_{:,r}\boldsymbol{W}_{:,r} + \boldsymbol{\delta}^{(j)} \cdot \nabla_{\boldsymbol{x}} f \nabla_{\boldsymbol{x}} f + \mathcal{O}(\boldsymbol{\delta}^2);$$

2. $\mathcal{P}^{(j)}_{\mathcal{B}(0,\epsilon)} \neq I$:

$$\boldsymbol{\kappa}^{(j+1)} = \frac{\epsilon}{\|\boldsymbol{\kappa}^{(j)} + \nabla_{\boldsymbol{x}} \ell^{(j)}\|}\left(\boldsymbol{\kappa}^{(j)} + \frac{\partial \ell}{\partial f}^{(j)} \nabla_{\boldsymbol{x}} f^{(j)}\right).$$

In either case, the quantity $\tau(\tilde{\boldsymbol{\kappa}}[k])$ is proportional to

$$\left|\mathcal{F}\left(\boldsymbol{\kappa}^{(j)} + \frac{\partial \ell}{\partial f}^{(j)} \nabla_{\boldsymbol{x}} f^{(j)}\right)\right|^2 = \left|\mathcal{F}\left(\boldsymbol{\kappa}^{(j)} + \frac{\partial \ell}{\partial f}\nabla_{\boldsymbol{x}} f + \boldsymbol{\delta}^{(j)} \cdot \nabla_{\boldsymbol{x}} f \nabla_{\boldsymbol{x}} f\right)\right|^2$$

since the term $\boldsymbol{\delta}\partial\ell/\partial f < \delta^2$ and can be dropped. Therefore, we now consider

$$\tilde{\boldsymbol{\kappa}}^{(j+1)}[k] = \tilde{\boldsymbol{\kappa}}^{(j)}[k] + \left(\frac{\partial \ell}{\partial f} + \boldsymbol{\delta}^{(j)} \cdot \nabla_{\boldsymbol{x}} f\right)\widetilde{\nabla f}[k] \tag{21}$$

at the $(j+1)$-th step of PGD in the frequency domain to explore ratio of frequency k to the whole frequency spectrum of perturbations

$$\tau\left(\tilde{\boldsymbol{\delta}}^{(j+1)}[k]\right) \propto |\tilde{\boldsymbol{\kappa}}^{(j)}[k]|^2 + 2\left(\beta' + \boldsymbol{\delta}^{(j)} \cdot \nabla_{\boldsymbol{x}} f\right)|\tilde{\boldsymbol{\kappa}}^{(j)}[k]||\widetilde{\nabla f}[k]|\cos(\triangle\phi_k^{(j)}). \tag{22}$$

Lemma 1 (Dynamics of l_2 Norm PGD in the Frequency Domain) *If the initialization of l_2 norm PGD adversarial perturbation satisfies $\frac{\partial \ell}{\partial f} + \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f > 0$, then it will be positive at every iteration of PGD. In this case, $\widetilde{\nabla f}$ will increase its k-th amplitude after*

$$j = \begin{cases} 0 & \text{if } \cos(\triangle\phi_k^{(0)}) \geq 0, \\ -\frac{|\tilde{\boldsymbol{\kappa}}^{(0)}[k]|\cos(\triangle\phi_k^{(0)})}{\left(\beta' + \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f\right)|\widetilde{\nabla f}[k]|} & \text{if } \cos(\triangle\phi_k^{(0)}) < 0 \end{cases} \tag{23}$$

iterations of PGD [3].

Similar to Eq.(15), one can derive the changed amount of $\tau(\tilde{\boldsymbol{\delta}} \in S_l)$ in theorem 4 at the $(j+1)$-th step and find the condition of increasing.

B.2 l_∞-PERTURBATIONS

We now dive into the case for l_∞ PGD perturbations. The PGD update rule at step j now becomes

$$\boldsymbol{\delta}^{(j+1)} = \text{Clip}_{[-\epsilon,\epsilon]}\left[\boldsymbol{\delta}^{(j)} + \xi \cdot \text{sgn}\left(\frac{\partial \ell}{\partial f}^{(j)} \nabla_{\boldsymbol{x}} f^{(j)}\right)\right].$$

Let $\mu \in \{0, 1, ..., d\}$ and recall our definition that $\partial_\mu f$ refers to the μ-th component of $\nabla_{\boldsymbol{x}} f$, we have for the μ-th component of $\boldsymbol{\delta}$ that

$$\delta_\mu^{(j+1)} = \text{Clip}_{[-\epsilon,\epsilon]}\left[\delta_\mu^{(j)} + \xi\text{sgn}\left(\frac{\partial \ell}{\partial f}^{(j)} \partial_\mu f^{(j)}\right)\right]. \tag{24}$$

l_∞ norm PGD adversarial perturbations for the two-layer neural network will have similar frequency spectrum to FGSM adversarial perturbations as stated below by theorem 5.

[3] A similar conclusion exists when $\frac{\partial \ell}{\partial f} + \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f < 0$.

Theorem 5 *l_∞ norm adversarial perturbations generated by PGD for a well normally trained neural network (1) will end up with forms of l_∞ FGSM perturbations after at most $2\epsilon/\xi$ steps.*

For a simple neural network (*e.g.* our model), the above theorem claims that FGSM and PGD with enough iterations are in fact equivalent to generate a l_∞-norm adversarial perturbation. Therefore it is possible to explore frequency spectrum of FGSM adversarial perturbations to provide insights on l_∞ PGD perturbations.

B.2.1 Proof of Theorem 5

For convenience, we explore the case when $\partial\ell/\partial f + \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f > 0$. In the PGD process, we first randomly find a $\boldsymbol{\delta}^{(0)}$. The probability that its components satisfy the condition $\delta_\mu \in [x_\mu - (\epsilon - \xi), x_\mu + (\epsilon - \xi)]$ for all $\mu \in \{0, 1, ..., d\}$

$$\left(\frac{\epsilon - \xi}{\epsilon}\right)^d$$

will be little when d is large, thus we expect that there will be at least some components $\mu \in C_1^{(0)}$ such that

$$\epsilon - \xi \leq \delta_\mu^{(0)} \leq \epsilon$$

and at least some other $\mu \in C_2^{(0)}$ such that

$$-\epsilon \leq \delta_\mu^{(0)} \leq -\epsilon + \xi.$$

Moreover, we let $C_1^{\pm(0)}$ denote

$$C_1^{(0)} \bigcap \{\mu : \text{sgn}(\partial_\mu f) = \pm 1\}$$

and let $C_2^{\pm(0)}$ denote

$$C_2^{(0)} \bigcap \{\mu : \text{sgn}(\partial_\mu f) = \pm 1\}.$$

If $\partial\ell/\partial f + \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f > 0$, the update rule of PGD at the first step will be

$$\begin{aligned}
\delta_\mu^{(1)} &= \text{Clip}_{[-\epsilon,\epsilon]}\left[\delta_\mu^{(0)} + \xi\text{sgn}\left(\partial_\mu f\right)\right] \\
&= \begin{cases} \epsilon & \mu \in C_1^{+(0)}, \\ -\epsilon & \mu \in C_2^{-(0)}, \\ \delta_\mu^{(0)} + \xi & \mu \in C_3^{(0)} \bigcup C_2^{+(0)}, \\ \delta_\mu^{(0)} - \xi & \mu \in C_4^{(0)} \bigcup C_1^{-(0)}, \end{cases}
\end{aligned}$$

where $d = |C_1^{(j)}| + |C_2^{(j)}| + |C_3^{(j)}| + |C_4^{(j)}|$ and $\mu \in C_3^{(j)}(C_4^{(j)})$ denotes $\delta_\mu^{(j)} \in [-\epsilon + \xi, \epsilon - \xi]$ with $\partial_\mu f > 0 (< 0)$. Specifically, $\boldsymbol{\delta}^{(1)} \cdot \nabla_{\boldsymbol{x}} f$ can be bounded by $\boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f$ as the following way:

$$\begin{aligned}
\delta^{(1)} \cdot \nabla_{\boldsymbol{x}} f &= \epsilon\left(\sum_{\rho \in C_1^{+(0)}} \partial_\rho f - \sum_{\rho \in C_2^{-(0)}} \partial_\rho f\right) + \xi\left(\sum_{\rho \in C_3^{(0)} \bigcup C_2^{+(0)}} \partial_\rho f - \sum_{\rho \in C_4^{(0)} \bigcup C_1^{-(0)}} \partial_\rho f\right) \\
&\quad + \sum_{\rho \notin \left(C_1^{+(0)} \bigcup C_2^{-(0)}\right)} \delta_\rho^{(0)} \partial_\rho f \\
&= \boldsymbol{\delta}^{(0)} \cdot \nabla_{\boldsymbol{x}} f + \sum_{\rho \in C_1^{+(0)}} (\epsilon - \delta_\rho^{(0)}) \partial_\rho f - \sum_{\rho \in C_2^{-(0)}} (\epsilon + \delta_\rho^{(0)}) \partial_\rho f \\
&\quad + \xi\left(\sum_{\rho \in C_3^{(0)} \bigcup C_2^{+(0)}} \partial_\rho f - \sum_{\rho \in C_4^{(0)} \bigcup C_1^{-(0)}} \partial_\rho f\right)
\end{aligned}$$

$$\begin{aligned} &\to \boldsymbol{\delta}^{(1)} \cdot \nabla_{\boldsymbol{x}} f \leq \left(\boldsymbol{\delta}^{(0)} + \xi \cdot \mathrm{sgn}(\nabla_x f)\right) \cdot \nabla_x f \\ &\boldsymbol{\delta}^{(1)} \cdot \nabla_x f \geq \boldsymbol{\delta}^{(0)} \cdot \nabla_x f. \end{aligned}$$

Following the iteration at step 2,

$$\begin{aligned} \delta_\mu^{(2)} &= \mathrm{Clip}_{[-\epsilon,\epsilon]}\left[\delta_\mu^{(1)} + \xi \mathrm{sgn}\left(\partial_\mu f\right)\right] \\ &= \begin{cases} \epsilon & \mu \in C_1^{+(1)}, \\ -\epsilon & \mu \in C_2^{-(1)}, \\ \delta_\mu^{(1)} + \xi & \mu \in C_3^{(1)} \bigcup C_2^{+(1)} \\ \delta_\mu^{(1)} - \xi & \mu \in C_4^{(1)} \bigcup C_1^{-(1)}, \end{cases} \end{aligned}$$

where

$$\begin{gathered} C_1^{+(0)} \subset C_1^{+(1)}, C_2^{-(0)} \subset C_2^{-(1)} \\ C_1^{-(1)} = \emptyset, C_2^{+(1)} = \emptyset \\ C_3^{(1)} \subset \left(C_3^{(0)} \bigcup C_2^{+(0)}\right), C_4^{(1)} \subset \left(C_4^{(0)} \bigcup C_1^{-(0)}\right). \end{gathered}$$

The above three sets of relations will be true for any other j—PGD does not reduce the number of components which are $\mathrm{sgn}(\partial_\mu f)\epsilon$ at every iteration. For any randomly chosen $\delta^{(0)}$, based on the above lemma, if there exists at least one component $\delta_\mu^{(0)}$ such that

$$\delta_\mu^{(0)} = -\mathrm{sgn}(\partial_\mu f)\epsilon,$$

it will then move towards the direction of $+\mathrm{sgn}(\partial_\mu f)\xi$ at every iteration j and finally become $\mathrm{sgn}(\partial_\mu f)\epsilon$ after $2\epsilon/\xi$ steps. This is the longest step for any component to move to $\mathrm{sgn}(\partial_\mu f)\epsilon$ after which will it remain unchanged. Therefore, a PGD perturbation now has the form

$$\boldsymbol{\delta} = \epsilon \cdot \mathrm{sgn}\left(\nabla_{\boldsymbol{x}} f\right)$$

which is identical to a l_∞ norm FGSM perturbation for $\frac{\partial \ell}{\partial f} > 0$

$$\boldsymbol{\delta}_{\mathrm{FGSM}} = \epsilon \cdot \mathrm{sgn}\left(\frac{\partial \ell}{\partial f} \nabla_{\boldsymbol{x}} f\right).$$

C SUPPLEMENTARY EXPERIMENT

in this part, we use MSE loss and remain other conditions unchanged in Section 4.

C.1 SUPPLEMENTARY EXPERIMENT OF CONTRIBUTION 1

We show the supplementary experiment of Section 4.1 in Fig. 7 and Fig. 8. They have similar results to show that the spectrum of adversarial perturbations are more concentrated in the low-frequency domain after a sufficient training.

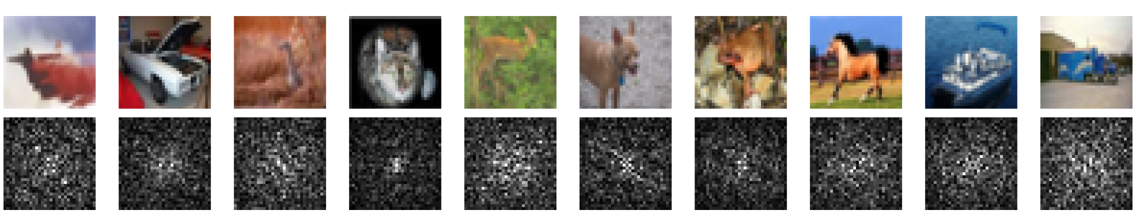

Figure 7: The difference of the spectrum between original and adversarial examples. The model are trained with MSE loss.

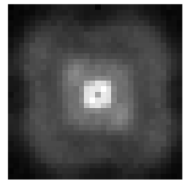

Figure 8: The expectation of the spectrum of adversarial perturbations. The model are trained with MSE loss.

C.2 Supplementary experiment of contribution 2

We show the supplementary experiment of Section 4.2 in Fig. 9 and Fig. 10. They have similar results to show the log-spectrum difference of adversarial examples is generally concentrated around.

In general, there is little empirical gap between MSE loss and CrossEntropy loss.

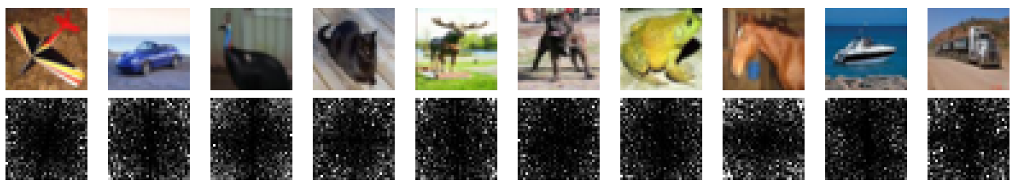

Figure 9: The difference of the log-spectrum between original and adversarial examples. The model are trained with MSE loss.

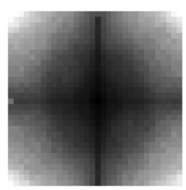

Figure 10: The expectation of the log-spectrum of adversarial perturbations. The model are trained with MSE loss.

D Perturbations of PGD-attack adversarial trained model are more concentrated in LFC

We do use the same setup in Section 4, and the PGD-attack adversarial training is also PGD-attack with $\epsilon = 8/255$, 40 iterations and step size $\xi = 4/255$ in each training step. Then we test the ratio of LFC of original images, perturbations in a normally trained model and perturbations in a PGD-attack adversarially trained model. The results are shown in Fig. 11.

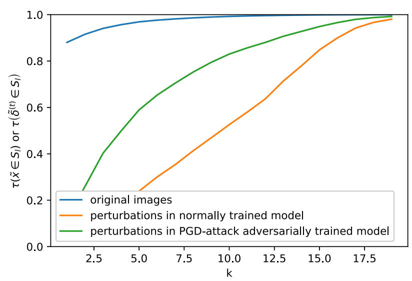

Figure 11: the ratio of LFC of original images, perturbations in a normally trained model and perturbations in a PGD-attack adversarially trained model.