A Robustness and Generalizaion

Here, we start by following the definition of *robustness* and derivation of the generalization bound in Xu and Mannor (2012):

Definition 1. *Denote $\mathcal{Z}$ as the set from which each sample is drawn and S as training sample set consisting of n training samples $(\mathbf{x}_1, ..., \mathbf{x}_n)$, respectively. Algorithm $\mathcal{A}$ is mapping from $\mathcal{Z}^n$ to hypothesis set $\mathcal{H}$. Then algorithm $\mathcal{A}$ is (K, $\epsilon(S)$) robust, for $K \in \mathbb{N}$ and $\epsilon(S) : \mathcal{Z}^n \mapsto \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{C_i\}_{i=1}^K$, such that the following holds for all $S \in \mathcal{Z}^n$:*

$\forall \mathbf{x} \in S$, $\forall \mathbf{z} \in \mathcal{Z}$, $\forall i = 1, ..., K$: if $\mathbf{x}, \mathbf{z} \in C_i$, then

$$|\mathcal{L}(\mathcal{A}_S; \mathbf{z}) - \mathcal{L}(\mathcal{A}_S; \mathbf{x})| \leq \epsilon(S), \tag{1}$$

where $\mathcal{L}$ is the loss function.

Theorem 1. *If a learning algorithm $\mathcal{A}$ is (K, $\epsilon(S)$) robust, $\mathcal{L}$ is upper bounded by M and the training sample set S is generated by n IID draws from a distribution μ, then for any $0 < p \leq 1$, with probability at least $1 - p$ we have*

$$|\mathbb{E}_{z\sim\mu}[\mathcal{L}(\mathcal{A}_S; \mathbf{z})] - \frac{1}{n}\sum_{\mathbf{x}_i \in S} \mathcal{L}(\mathcal{A}_S; \mathbf{x}_i)| \leq \epsilon(S) + M\sqrt{\frac{2K\ln 2 + 2\ln(1/p)}{n}}$$

We slightly modify the aforementioned notions for describing the generalization ability in DNNs. In Definition 1, robustness adopts the upper bounds of the difference between loss functions of two data points, given that they are on the same partition C_i. When selecting any input $\mathbf{x} \in \mathbb{R}^{n\times 1}$ belonging to C_i, let $\boldsymbol{\delta}^{\mathbf{x}} \in \mathbb{R}^{n\times 1}$ be any perturbation on the input satisfying $(\mathbf{x} + \boldsymbol{\delta}^{\mathbf{x}}) \in C_i$. Further, denote $\mathbf{w} \in \mathbb{R}^d$ as the vectorized weight of the model and $\mathcal{L}(\mathbf{w}; \mathbf{x})$ as the loss function of the neural network. Then, we obtain a method for measuring *robustness* in DNNs.

$$|\mathcal{L}(\mathbf{w}; \mathbf{x} + \boldsymbol{\delta}^{\mathbf{x}}) - \mathcal{L}(\mathbf{w}; \mathbf{x})| < \epsilon, \tag{2}$$

which has equivalent meaning with (1).

B Extension of Lemma 1 and Proposition 1

Lemma 2. *Let $\boldsymbol{\delta}^{\mathbf{x}}$ and $\boldsymbol{\delta}^{\mathbf{W}_i}$ be input and weight perturbations at i-th layer, respectively. For a n-layer neural network, the input perturbations $\boldsymbol{\delta}^{\mathbf{x}}$ can be transferred to the combinations of weight perturbations $\boldsymbol{\delta}^{\mathbf{W}_i}$s. More formally, suppose we have weight matrix which is between $(i-1)$-th layer and i-th layer $\mathbf{W}_i \in \mathbb{R}^{l_i \times l_{i-1}}$ where l_i is the number of features in the i-th layer and assume $l_i \leq l_{i-1}$. Then for any $\boldsymbol{\delta}^{\mathbf{x}} \neq 0$, there exists $\boldsymbol{\delta}^{\mathbf{W}_i}$s such that $|\mathcal{L}(\mathbf{w}+\boldsymbol{\delta}^{\mathbf{w}}; \mathbf{x}) - \mathcal{L}(\mathbf{w}; \mathbf{x})| \simeq |\mathcal{L}(\mathbf{w}; \mathbf{x}+\boldsymbol{\delta}^{\mathbf{x}}) - \mathcal{L}(\mathbf{w}; \mathbf{x})|$.*

Proof. Here, we ignore the bias term and activation function. We start with a two-layer neural network $\mathbf{y} = \mathbf{W_2 W_1 x}$, then generalize to n-layer DNNs. In case of $n = 2$, the proof can be accomplished by finding $\boldsymbol{\delta}^{\mathbf{W}_2}$ and $\boldsymbol{\delta}^{\mathbf{W}_1}$ which satisfies

$$(\mathbf{W}_2 + \boldsymbol{\delta}^{\mathbf{W}_2})(\mathbf{W}_1 + \boldsymbol{\delta}^{\mathbf{W}_1})\mathbf{x} = \mathbf{W}_2\mathbf{W}_1(\mathbf{x} + \boldsymbol{\delta}^{\mathbf{x}}) \tag{3}$$

$$(\mathbf{W}_2\boldsymbol{\delta}^{\mathbf{W}_1} + \boldsymbol{\delta}^{\mathbf{W}_2}\mathbf{W}_1)\mathbf{x} \simeq \mathbf{W}_2\mathbf{W}_1\boldsymbol{\delta}^{\mathbf{x}} \tag{4}$$

where $\boldsymbol{\delta}^{\mathbf{W}_2}\boldsymbol{\delta}^{\mathbf{W}_1} \approx 0$ because $\boldsymbol{\delta}$ are small. Then, the following choice of $\boldsymbol{\delta}^{\mathbf{W}_2}$ and $\boldsymbol{\delta}^{\mathbf{W}_1}$ satisfies (4) :

$$\mathbf{W}_2\boldsymbol{\delta}^{\mathbf{W}_1} + \boldsymbol{\delta}^{\mathbf{W}_2}\mathbf{W}_1 \simeq \frac{\mathbf{W}_2\mathbf{W}_1\boldsymbol{\delta}^{\mathbf{x}}}{\mathbf{x}^\top\mathbf{x}}\mathbf{x}^\top \tag{5}$$

$$\boldsymbol{\delta}^{\mathbf{W}_2}\mathbf{W}_1 \simeq (\frac{\mathbf{W}_2\mathbf{W}_1\boldsymbol{\delta}^{\mathbf{x}}}{\mathbf{x}^\top\mathbf{x}}\mathbf{x}^\top - \mathbf{W}_2\boldsymbol{\delta}^{\mathbf{W}_1}) \tag{6}$$

In (6), $\boldsymbol{\delta}^{\mathbf{W}_2}$ exists for $\forall\boldsymbol{\delta}^{\mathbf{W}_1}$ because $rank(\mathbf{W}_1) \leq l_1$. If $rank(\mathbf{W}_1) = l_1$ then $\mathbf{W}_1$ has a right inverse. Otherwise, $rank(\mathbf{W}_1) < l_1$ so the equation is under-determined, resulting in infinitely many solutions.

Now, Lemma 1 can be easily generalized to any n-layer neural networks by simply replacing (1) of Lemma 1 with the following :

$$\sum_{i=1}^{n}\Big((\prod_{k=1}^{n-i}\mathbf{W}_{n-k+1})\boldsymbol{\delta}^{\mathbf{W}_i}(\prod_{k=1}^{i-1}\mathbf{W}_{i-k})\Big) \simeq \frac{(\prod_{i=1}^{n}\mathbf{W}_{n-i+1})\boldsymbol{\delta}^{\mathbf{x}}}{\mathbf{x}^\top\mathbf{x}}\mathbf{x}^\top \tag{7}$$

Then, we have

$$\boldsymbol{\delta}^{\mathbf{W}_n}(\prod_{k=1}^{i-1}\mathbf{W}_{i-k}) \simeq \frac{(\prod_{i=1}^{n}\mathbf{W}_{n-i+1})\boldsymbol{\delta}^{\mathbf{x}}}{\mathbf{x}^\top\mathbf{x}}\mathbf{x}^\top - \sum_{i=1}^{n-1}\Big((\prod_{k=1}^{n-i}\mathbf{W}_{n-k+1})\boldsymbol{\delta}^{\mathbf{W}_i}(\prod_{k=1}^{i-1}\mathbf{W}_{i-k})\Big) \tag{8}$$

which means $\boldsymbol{\delta}^{\mathbf{W}_n}$ always exists because $rank(\prod_{k=1}^{n-1}\mathbf{W}_{n-k}) \leq l_{n-1}$. □

Proposition 2. *Suppose $|\mathcal{L}(\mathbf{w}+\boldsymbol{\delta}^{\mathbf{w}}; \mathbf{x}) - \mathcal{L}(\mathbf{w}; \mathbf{x})| < \epsilon$ holds for any $\boldsymbol{\delta}^{\mathbf{w}}$ such that*

$$\left\|\sum_{i=1}^{n}\Big((\prod_{k=1}^{n-i}\mathbf{W}_{n-k+1})\boldsymbol{\delta}^{\mathbf{W}_i}(\prod_{k=1}^{i-1}\mathbf{W}_{i-k})\Big)\right\|_F < \delta.$$

Then $|\mathcal{L}(\mathbf{w}; \mathbf{x} + \boldsymbol{\delta}^{\mathbf{x}}) - \mathcal{L}(\mathbf{w}; \mathbf{x})| < \epsilon$ holds for any $\boldsymbol{\delta}^{\mathbf{x}}$ such that $\frac{\|\boldsymbol{\delta}^x\|}{\|\mathbf{x}\|} < \frac{\delta}{\sigma_{max}(\prod_{i=1}^{n}\mathbf{W}_{n-i+1})}$.

Proof. The proof is almost same with Proposition 1 and easily done by replacing (2) of Lemma 1 with (7). Then,

$$\left\|\sum_{i=1}^{n}\Big((\prod_{k=1}^{n-i}\mathbf{W}_{n-k+1})\boldsymbol{\delta}^{\mathbf{W}_i}(\prod_{k=1}^{i-1}\mathbf{W}_{i-k})\Big)\right\|_F^2 = \frac{\|(\prod_{i=1}^{n}\mathbf{W}_i)\boldsymbol{\delta}^{\mathbf{x}}\|^2\|\mathbf{x}^\top\|^2}{(\mathbf{x}^\top\mathbf{x})^2} \leq \frac{\sigma_{max}^2(\prod_{i=1}^{n}\mathbf{W}_{n-i+1})\|\boldsymbol{\delta}^x\|^2}{\|\mathbf{x}\|^2} < \delta^2 \tag{9}$$

□

References

Xu, H. and Mannor, S. (2012). Robustness and generalization. *Machine learning*, 86(3):391–423.