Neural Network Generalization Bounds via Compression

May 2018

Speakers: Thomas Orton and Guillermo Valle

1 Introduction

Neural networks emperically show good generalization when trained on real life data; however, we have been unable to theoretically explain why this is the case. In particular, traditional techniques which try to give generalization bounds based on the number of parameters of a neural network significantly overestimate the number of training samples required for the network to generalize. The paper we're talking about today does two things to help make progress in this area:

1. Introduce a new conceptial "compression" framework for getting generalization bounds (section 2).
2. Give new generalization bounds for neural networks based on metrics of networks which are found to be emperically favourable when trained on real life data (section 3).

2 Compression and Generalization

2.1 Setting and notation

We consider a multiclass classification setting, where labels come from $\{1, ..., k\}$ and for a sample x we map $x \to f(x) \in \mathbb{R}^k$, where the classification loss is defined as

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}\left[f(x)[y] < \max_{i\neq y} f(x)[i]\right]$$

If $\gamma > 0$ is some desired margin, then the expected margin loss is

$$L_\gamma(f) = \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[f(x)[y] \leq \gamma + \max_{i\neq y} f(x)[i]\right]$$

We let $\hat{L}_\gamma$ denote the emperical margin loss, and L_γ the true margin loss. The generalization error is the difference between the two.

2.2 Compressibility

The high level idea of this section is the following: Suppose we have a classifier f from a complicated hypothesis class which has very low emperical loss on m samples. We can try to approximate f by compressing it to a function g, where g belongs to a family of functions with fewer than m effective parameters. This allows us to get generalization bounds on g. The following definitions make this idea formal:

Definition 2.1 ((γ,S)-compressible using helper string s)**.** Suppose $G_{\mathcal{A},s} = \{g_{A,s}|A \in \mathcal{A}\}$ is a class of classifiers indexed by trainable parameters A and fixed strings s. A classifier f is (γ, S)-compressible with respect to $G_{\mathcal{A},s}$ using helper string s if there exists $A \in \mathcal{A}$ such that for any $x \in S$, we have for all y

$$|f(x)[y] - g_{A,s}(x)[y]| \leq \gamma. \tag{1}$$

Theorem 1. *Suppose $G_{\mathcal{A},s} = \{g_{A,s}|A \in \mathcal{A}\}$ where A is a set of q parameters each of which can have at most r discrete values and s is a helper string. Let S be a training set with m samples. If the trained classifier f is (γ, S)-compressible via $G_{\mathcal{A},s}$ with helper string s, then there exists $A \in \mathcal{A}$ s.t. with high probability over the training set,*

$$L_0(g_A) \leq \hat{L}_\gamma(f) + O\left(\sqrt{\frac{q \log r}{m}}\right). \tag{2}$$

Proof. We can use the Chernoff bound to give

$$Pr[L_0(g_A) - \hat{L}_\gamma(g_A) \geq \tau] \leq \exp(-2\tau^2 m) \tag{3}$$

Choosing $\tau = \left(\sqrt{\frac{q \log r}{m}}\right)$, and taking a union bound over all r^q different $A \in \mathcal{A}$, we have that with probability at least $1 - \exp(-q \log r)$, for all $A \in \mathcal{A}$

$$L_0(g_A) \leq \hat{L}_\gamma(g_A) + \left(\sqrt{\frac{q \log r}{m}}\right) \tag{4}$$

Since f is $(\gamma, \mathcal{S})-$compressible with respect to g, we can pick an $A \in \mathcal{A}$ such that for all $x \in S$, and any y, we have

$$|f(x)[y] - g_A(x)[y]| \leq \gamma \tag{5}$$

For each training example, if f has margin at least γ, then g_A also classifies this example correctly. Thus

$$\hat{L}_0(g_A) \leq \hat{L}_\gamma(f) \tag{6}$$

□

Comment: In the setting of Theorem 2.1, if the compression works for $1 - \zeta$ fraction of the training sample, then with high probability

$$L_0(g_A) \leq \hat{L}_\gamma(f) + \zeta + O\left(\sqrt{\frac{q \log r}{m}}\right).$$

2.3 Examples

2.4 Example 1: Linear classifiers with margin

Consider a binary linear classifier $c \in \mathbb{R}^h$, $\|c\| = 1$ with high margin. Namely $c(x) := sgn(c \cdot x)$, where $\forall x, y$ we have $\|x\| = 1$, $y \in \{-1, 1\}$, and for all $(x, y) \in S$ in our training set, we have $y(c^T x) \geq \gamma$. We can now consider how to compress such a classifier to a simpler family of classifiers:

Lemma 2. *For any fixed vector u, Algorithm Vector-Project(c, γ) produces a vector $\hat{c}$ such that with probability at least $1 - \eta$, we have $|\hat{c}^\top u - c^\top u| \leq \gamma$.*

```
Algorithm 1 Vector-Project($\gamma$, $c$)
Require: vector $c$ with $\|c\| \leq 1$, $\eta$.
Ensure: vector $\hat{c}$ s.t. for any fixed vector $\|u\| \leq 1$, with probability at least $1-\eta$, $|c^\top u - \hat{c}^\top u| \leq \gamma$.
  Let $k = 16\log(1/\eta)/\gamma^2$
  Sample $k$ random Gaussian vectors $v_1, ..., v_k \sim N(0, I)$.
  Compute $z_i = \langle v_i, c\rangle$
  (Optional): Round $z_i$ to the closes multiple of $\gamma/2\sqrt{hk}$.
  Return $\hat{c} = \frac{1}{k}\sum_{i=1}^k z_i v_i$
```

Proof. This is in fact a well-known corollary of Johnson-Lindenstrauss Lemma. Observe that

$$\hat{c}^\top u = \frac{1}{k}\sum_{i=1}^{k}\langle v_i, c\rangle\langle v_i, u\rangle \tag{7}$$

The expectation $E[\langle v_i, c\rangle\langle v_i, u\rangle] = E[c^\top v_i v_i^\top u] = c^\top E[v_i v_i^\top]u = c^\top u$. Also

$$\begin{aligned}
Var[\frac{1}{k}\sum_{i=1}^{k}\langle v_i, c\rangle\langle v_i, u\rangle] &= \frac{1}{k}Var[\langle w, c\rangle\langle w, u\rangle] \\
&\leq \frac{1}{k}E\left[((c^T w)(u^T w))^2\right] \\
&= \frac{1}{k}E\left[\sum_{i,j,k,l} c_i c_j u_k u_l w_i w_j w_k w_l\right] \\
&\leq \frac{1}{k}\left[E\left[\sum_{i,k} c_i^2 u_k^2 w_i^2 w_k^2\right] + 2E\left[\sum_{i,k} c_i u_i c_k u_k w_i^2 w_k^2\right]\right] \\
&= O\left(\frac{1}{k}\right)\left[\sum_{i,k} c_i^2 u_k^2 + \sum_{i,k} c_i u_i c_k u_k\right] = O\left(\frac{1}{k}\right)
\end{aligned}$$

Since c, u have norm less than 1, and $\sum_i c_i u_i \leq \|c\|\|u\|$.

Standard concerntration inequalities give that

$$\Pr[|\hat{c}^\top u - c^\top u| > \gamma/2] \leq \exp(-\gamma^2 k/16) = \exp(\log(\eta)) = \eta. \tag{8}$$

It remains to show that by rounding z_i to the closest multiples of $\gamma/2\sqrt{hk}$, our error increases by at most $\gamma/2$.

Claim: a matrix with *i.i.d.* Consider the matrix V with columns $v_1, ..., v_k$. Then with high probability, its spectral norm is at most $2\sqrt{h}$.

Notice that $\hat{c} = Vz$, where z is the vector of the z_i's. If the spectral norm of V is at most $2\sqrt{h}$, then changing each z_i coordinate-wise by at most $\gamma/4\sqrt{hk}$ can change $\hat{c}$ in l_2 norm by at most $\gamma/2$.

□

Lemma 3. *For any number of sample m, there is an efficient algorithm with helper string to generate a compressed vector $\hat{c}$, such that*

$$L(\hat{c}) \leq \tilde{O}(\sqrt{1/\gamma^2 m}). \tag{9}$$

Proof. We will choose $\eta = 1/m$. By Lemma, we know there is a compression algorithm that works with probability $1-\eta$, and has at most $O((\log 1/\eta)/\gamma^2)$ parameters. By Corollary, we know

$$L(\hat{c}) \leq \tilde{O}(\eta + \sqrt{1/\gamma^2 m}) \leq \tilde{O}(\sqrt{1/\gamma^2 m}).$$

□

2.5 Example 2: Existing generalization bounds

Notation: We define the outputs of the ith layer of a neural network by $x^i = A^i \phi(x^{i-1})$, where ϕ is a RelU activation function.

Theorem 4. *A deep net with layers $A^1, A^2, \dots A^d$ and output margin γ on a training set S, the generalization error can be bounded by*

$$\tilde{O}\left(\sqrt{\frac{hd^2 \max_{x\in S}\|x\| \prod_{i=1}^d \|A^i\|_2^2 \sum_{i=1}^d \frac{\|A^i\|_F^2}{\|A^i\|_2^2}}{\gamma^2 m}}\right). \tag{10}$$

Note that $(\sum_{i=1}^d \frac{\|A^i\|_F^2}{\|A^i\|_2^2})$ is sum of stable ranks of the layers, and that $(\prod_{i=1}^d \|A^i\|_2^2)$ is the maximum norm of the vector it can produce if the input is a unit vector. The Lipschitz constant of the full network is at most $\prod_{i=1}^d \|A^i\|_2$.

We prove the theorem in two steps. First, we want to compress a matrix A to a low rank matrix $\hat{A}$. In particular, if $\hat{A}$ has rank r, it can be expressed as a product of two matrices B_1, B_2 of inner dimension r, and so $\hat{A}$ has $2hr$ parameters. We can then round the entries of B_1, B_2 to compress A to a finite set of functions. The second step is to show that if we replace the layers $\{A^i\}$ by $\{\hat{A}_i\}$, then the output of our network doesn't change significantly.

Lemma 5. *For any matrix $A \in \mathbb{R}^{m\times n}$, let $\hat{A}$ be the truncated version of A where singular values that are smaller than $\delta\|A\|_2$ are removed. Then $\|\hat{A} - A\|_2 \leq \delta\|A\|_2$ and $\hat{A}$ has rank at most $\|A\|_F^2/(\delta^2\|A\|_2^2)$.*

Proof. Let r be the rank of $\hat{A}$. By construction, the maximum singular value of $\hat{A}-A$ is at most $\delta\|A\|_2$. Since the remaining singular values are at least $\delta\|A\|_2$, we have $\|A\|_F \geq \|\hat{A}\|_F \geq \sqrt{r}\delta\|A\|_2$. These inequalities come from that fact that if $A = UDV^T$ is the SVD of A with singular values $\{\delta_i\}$, then

$$\|A\|_F^2 = tr[AA^T] = tr[UDV^TVD^TU^T] = tr[DD^TU^TU] = tr[DD^T] = \sum_i \delta_i^2 \tag{11}$$

i.e. the frobenius norm squared of a matrix is just the sum of the squared signular values of that matrix. □

Lemma 6. *Let f_A^i denote the output of the ith layer a d layer neural network with layers $A^1, ..., A^d$. Let $\Delta_i = \|f_{A+B}^i(x) - f_A^i(x)\|_2$. Suppose for all layers i we have $\|B^i\| \leq \frac{1}{d}\|A^i\|$. Then:*

$$\Delta_i \leq \left(1 + \frac{1}{d}\right)^i \left(\prod_{j=1}^i \|A_j\|_2\right) \|x\|_2 \sum_{j=1}^i \frac{\|B_j\|_2}{\|A_j\|_2}.$$

Proof. For the induction case:

$$\begin{aligned}
\Delta_{i+1} &= \| \left(A^{i+1} + B^{i+1}\right) \phi_i(f_{A+B}^i(x)) - A^{i+1}\phi_i(f_A^i(x))\|_2 \\
&= \| \left(A^{i+1} + B^{i+1}\right) \left(\phi_i(f_{A+B}^i(x)) - \phi_i(f_A^i(x))\right) + B^{i+1}\phi_i(f_A^i(x))\|_2 \\
&\leq \left(\|A^{i+1}\|_2 + \|B^{i+1}\|_2\right) \|\phi_i(f_{A+B}^i(x)) - \phi_i(f_A^i(x))\|_2 + \|B^{i+1}\|_2 \|\phi_i(f_A^i(x))\|_2 \\
&\leq \left(\|A^{i+1}\|_2 + \|B^{i+1}\|_2\right) \|f_{A+B}^i(x) - f_A^i(x)\|_2 + \|B^{i+1}\|_2 \|f_A^i(x)\|_2 \\
&= \Delta_i \left(\|A^{i+1}\|_2 + \|B^{i+1}\|_2\right) + \|B^{i+1}\|_2 \|f_A^i(x)\|_2,
\end{aligned}$$

And so we get

$$\begin{aligned}
\Delta_{i+1} &\leq \Delta_i \left(1+\frac{1}{d}\right) \|A^{i+1}\|_2 + \|B^{i+1}\|_2 \|x\|_2 \prod_{j=1}^{i} \|A^j\|_2 \\
&\leq \left(1+\frac{1}{d}\right)^{i+1} \left(\prod_{j=1}^{i+1} \|A^j\|_2\right) \|x\|_2 \sum_{j=1}^{i} \frac{\|B^j\|_2}{\|A^j\|_2} + \frac{\|B^{i+1}\|_2}{\|A^{i+1}\|_2} \|x\|_2 \prod_{j=1}^{i+1} \|A^i\|_2 \\
&\leq \left(1+\frac{1}{d}\right)^{i+1} \left(\prod_{j=1}^{i+1} \|A^j\|_2\right) \|x\|_2 \sum_{j=1}^{i+1} \frac{\|B^j\|_2}{\|A^j\|_2}
\end{aligned}$$

□

Combining these lemmas, we can prove the theorem:

Proof. For each i, replace A^i by its compression $\hat{A}^i$ with parameter $\delta = \gamma \max_x (e\|x\| d \prod_{i=1}^{d} \|A^i\|_2)^{-1}$. The network $f_{\hat{A}}$ then has error at most

$$\Delta_d \leq e \left(\prod_{j=1}^{d} \|A_j\|_2\right) \|x\|_2 \sum_{j=1}^{d} \frac{\|\hat{A}_j - A_j\|_2}{\|A_j\|_2} \leq e \left(\prod_{j=1}^{d} \|A_j\|_2\right) \|x\|_2 d\delta = \gamma.$$

and the total number of parameters of the network is at most

$$\sum_{i=1}^{d} \frac{\|A\|_F^2}{(\delta^2 \|A\|_2^2)} 2h = 2h\delta^{-2} \sum_{i=1}^{d} \frac{\|A\|_F^2}{(\delta^2 \|A\|_2^2)} = 2e^2 d^2 h \|x\|^2 \prod_{i=1}^{d} \|A^i\|_2^2 \sum_{i=1}^{d} \frac{\|A^i\|_F^2}{\|A^i\|_2^2} / \gamma^2.$$

Claim: We can round the weights in the rank r representation of each $\hat{A}_i$ to the nearest $\|A\|_F / h^2$ while keeping the approximation error sufficiently small. Note that the precise parameters for how we discretize will ordinarily not show up in the $\tilde{O}$ term, because there is only a logarithmic dependence of the model complexity on the number of discrete values each parameter has.

Applying our (γ, S)-compressibility theorem from the previous section, we get that our compressed net has error at most

$$L_0(g_A) \leq \hat{L}_\gamma(f) + \tilde{O}\left(\sqrt{\frac{hd^2 \max_{x \in S} \|x\| \prod_{i=1}^{d} \|A^i\|_2^2 \sum_{i=1}^{d} \frac{\|A^i\|_F^2}{\|A^i\|_2^2}}{\gamma^2 m}}\right) \tag{12}$$

Since f has output margin γ, it has the same error as g_A, and hence the same generalization bound applies.

□

Comment: There is a technical issue with the above proof as presented in the paper, which requires us to be careful when invoking (γ, S)-compressibility. In particular, we are supposed to only choose the set of models we compress f to **before** we see the training set, but the above proof chooses this set based on properties of f after training. In order to deal with this, we should really choose the rank r and precision of descritization we compress our matrices to in advance, and then get a generalization bound based on this choice assuming we can compress f to this set of models we chose in advance.

3 Definition of quantities measuring noise sensitivity

These are quantities measuring the sensitivity to noise of the neural network, and which are used in proving Theorem 7.

Let S be the training set.

1. **Layer cushion (μ_i)**: For any layer i, we define the layer cushion μ_i as the largest number such that for any $x \in S$:
$$\mu_i \|A^i\|_F \|\phi(x^{i-1})\| \leq \|A^i \phi(x^{i-1})\|$$

2. **Interlayer cushion ($\mu_{i,j}$)**: For any two layers $i \leq j$, we define interlayer cushion $\mu_{i,j}$ as the largest number such that for any $x \in S$:
$$\mu_{i,j} \|J^{i,j}_{x^i}\|_F \|x^i\| \leq \|J^{i,j}_{x^i} x^i\|$$
Furthermore, we define minimal interlayer cushion $\mu_{i \rightarrow} = \min_{i \leq j \leq d} \mu_{i,j} = \min\{1/\sqrt{h^i}, \min_{i < j \leq d} \mu_{i,j}\}$.

3. **Activation contraction (c)**: The activation contaction c is defined as the smallest number such that for any layer i and any $x \in S$,
$$\|x^i\| \leq c \|\phi(x^i)\|$$

4. **Interlayer smoothness (ρ_δ)**: Interlayer smoothness is defined the smallest number such that with probability $1-\delta$ over noise η for any two layers $i < j$ any $x \in S$:
$$\|M^{i,j}(x^i + \eta) - J^{i,j}_{x^i}(x^i + \eta)\| \leq \frac{\|\eta\| \|x^j\|}{\rho_\delta \|x^i\|}$$

4 Fully connected networks

The intuition behind this result is that if we can find a network $f_{\tilde{A}}$, which does well on the training set, and belongs to a small hypothesis class (independent of the training set), then we can bound its generalization error. The way we ensure it does well on the training set is by ensuring its outputs are not very different from f_A which already does well on the training set (if $\hat{L}_\gamma(f_A)$ is small). The way we ensure it belongs to a small hypothesis class is by constructing $f_{\tilde{A}}$ in a way that it is parametrized by few parameters

Theorem 7. *For any fully connected network f_A with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any margin γ, Algorithm 2 generates weights $\tilde{A}$ for the network $f_{\tilde{A}}$ such that with probability $1-\delta$ over the training set and $f_{\tilde{A}}$, the expected error $L_0(f_{\tilde{A}})$ is bounded by*

$$\hat{L}_\gamma(f_A) + \tilde{O}\left(\sqrt{\frac{c^2 d^2 \max_{x \in S} ||f_A(x)||_2^2 \sum_{i=1}^d \frac{1}{\mu_i^2 \mu_{i\rightarrow}^2}}{\gamma^2 m}}\right)$$

where μ_i, $\mu_{i\rightarrow}$, c and ρ_δ are layer cushion, interlayer cushion, activation contraction and interlayer smoothness.

To prove this, the crucial step is to show that f_A can be "compressed" to a $f_{\tilde{A}}$ with less parameters. This means that the output of $f_{\tilde{A}}$ doesn't differ much from f_A for any input in the training set $x \in S$. This will be shown in Lemma 3 (which relies on Lemma 2). This in turn means that $f_{\tilde{A}}$ can't do much worse than f_A on the training set, in the sense that its margin can't be much smaller than that for f_A, which will allow us to say $\hat{L}_0(f_{\tilde{A}}) \leq \hat{L}_\gamma(f_A)$, shown in Lemma 10. Finally, by using an ϵ-cover of the hypothesis class of $f_{\tilde{A}}$, we can bound the difference between $\hat{L}_0(f_{\tilde{A}})$ and $L_0(f_{\tilde{A}})$ which depends on the number of parameters of $f_{\tilde{A}}$, proving Theorem 4.1.

We begin with the technical Lemma 8

Lemma 8. *For any $0 < \delta, \varepsilon \leq 1$, et $G = \{(U^i, x^i)\}_{i=1}^m$ be a set of matrix/vector pairs of size m where $U \in \mathbb{R}^{n \times h_1}$ and $x \in \mathbb{R}^{h_2}$, let $\hat{A} \in \mathbb{R}^{h_1 \times h_2}$ be the output of Algorithm 2 with $\eta = \delta/mn$ and $\Delta = \hat{A} - A$. With probability at least $1-\delta$ we have for any $(U, x) \in G$, $\|U \Delta x\| \leq \varepsilon \|A\|_F \|U\|_F \|x\|$.*

Algorithm 2 Matrix-Project (A, ε, η)

```
Require: Layer matrix $A \in \mathbb{R}^{h_1 \times h_2}$, error parameter $\varepsilon$, $\eta$.
Ensure: Returns $\hat{A}$ s.t. $\forall$ fixed vectors $u, v$,
    $$\Pr[\|u^\top \hat{A} v - u^\top A v\| \geq \varepsilon \|A\|_F \|u\| \|v\|] \leq \eta.$$
  Sample $k = \log(1/\eta)/\varepsilon^2$ random matrices $M_1, \ldots, M_k$ with entries i.i.d. $\pm 1$ ("helper string")
  for $k' = 1$ to $k$ do
    Let $Z_{k'} = \langle A, M_{k'} \rangle M_{k'}$.
  end for
  Let $\hat{A} = \frac{1}{k} \sum_{k'=1}^{k} Z_{k'}$
```

Proof. For any fixed vectors u, v, we have (from definition of $\hat{A}$ by Algorithm 2)

$$u^\top \hat{A} v = \frac{1}{k} \sum_{k'=1}^{k} u^\top Z_{k'} v = \frac{1}{k} \sum_{k'=1}^{k} \langle A, M_{k'} \rangle \langle uv^\top, M_{k'} \rangle.$$

This is a sum of independent terms (as the $M_{k'}$ are independent). Furthermore, its expected value is

$$E\left[\frac{1}{k} \sum_{k'=1}^{k} \langle A, M_{k'} \rangle \langle uv^\top, M_{k'} \rangle\right] = \frac{1}{k} \sum_{k'=1}^{k} E\left[\langle A, M_{k'} \rangle \langle uv^\top, M_{k'} \rangle\right] = \langle A, uv^\top \rangle$$

where the last equality is because only the terms involving the same element of $M_{k'}$ contribute to the expectation. We can then use Hoeffding's inequality

$$Pr\left[\left|\frac{1}{k} \sum_{k'=1}^{k} \langle A, M_{k'} \rangle \langle uv^\top, M_{k'} \rangle - \langle A, uv^\top \rangle\right| \geq \varepsilon h^2 \|A\|_F \|uv^\top\|_F\right] \leq 2e^{-k\varepsilon^2/2},$$

This is not exactly what they got. In particular that h^2 which is $\max \|M_{k'}\|_F$. I'm going to ignore this difference in the rest of the section.. as $|\langle A, M_{k'} \rangle \langle uv^\top, M_{k'} \rangle| \leq h^2 \|A\|_F \|uv^\top\|_F$

Therefore for the choice of $k = \log(1/\eta)/\varepsilon^2$ we know

$$\Pr\left[|u^\top \hat{A} v - u^\top A v\| \geq \varepsilon \|A\|_F \|u\| \|v\|\right] \leq \eta.$$

Now for any pair of matrix/vector $(U, x) \in G$, let u_i be the i-th row of U. There are mn such rows, and the inequality of interest holds with probability $\eta = \frac{\delta}{mn}$ for each. Therefore, by union bound we know with probability at least $1 - \delta$ for all u_i we have $|u_i^\top \Delta v\| \leq \varepsilon \|A\|_F \|u_i\| \|v\|$. Since $\|U \Delta x\|^2 = \sum_{i=1}^{n} (u_i^\top \Delta x)^2$ and $\|U\|_F^2 = \sum_{i=1}^{n} \|u_i\|^2$, we immediately get $\|U \Delta x\| \geq \varepsilon \|A\|_F \|U\|_F \|x\|$.

□

Now, we show Lemma 9, which shows that the compressed network doesn't differ much in its outputs with f_A.

Lemma 9. *For any fully connected network f_A with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any error $0 < \varepsilon \leq 1$, Algorithm 2 generates weights $\tilde{A}$ for a network with $\frac{72c^2 d^2 \log(mdh/\delta)}{\varepsilon^2} \cdot \sum_{i=1}^{d} \frac{1}{\mu_i^2 \mu_{i\rightarrow}^2}$ total parameters such that with probability $1 - \delta/2$ over the generated weights $\tilde{A}$, for any $x \in S$:*

$$\|f_A(x) - f_{\tilde{A}}(x)\| \leq \varepsilon \|f_A(x)\|.$$

where μ_i, $\mu_{i\rightarrow}$, c and ρ_δ are layer cushion, interlayer cushion, activation contraction and interlayer smoothness.

Proof. This is proven by induction on the layers i. For any layer $i \geq 0$, let $\hat{x}_i^j$ be the output at layer j if the weights $A^1, \dots, A^i$ in the first i layers are replaced with $\tilde{A}^1, \dots, \tilde{A}^i$. The induction hypothesis is then the following:

Consider any layer $i \geq 0$ and any $0 < \varepsilon \leq 1$. The following is true with probability $1 - \frac{i\delta}{2d}$ over $\tilde{A}^1, \dots, \tilde{A}^i$ for any $j \geq i$:

$$\|\tilde{x}_i^j - x^j\| \leq (i/d)\varepsilon\|x^j\|.$$

For the base case $i = 0$, since we are not perturbing the input, the inequality is trivial. Now assuming that the induction hypothesis is true for $i - 1$, we consider what happens at layer i. Let $\tilde{A}^i$ be the result of Algorithm 2 on A^i with $\varepsilon_i = \frac{\varepsilon\mu_i\mu_{i\to}}{4cd}$ and $\eta = \frac{\delta}{6d^2h^2m}$ (not sure why this choice of η). Now let's analyze the difference in activations between the network with the extra perturbation and the unperturbed network. For any $j \geq i$ we have, using the triangle inequality

$$\|\tilde{x}_i^j - x^j\| = \|(\tilde{x}_i^j - \tilde{x}_{i-1}^j) + (\tilde{x}_{i-1}^j - x^j)\| \leq \|(\tilde{x}_i^j - \tilde{x}_{i-1}^j)\| + \|\tilde{x}_{i-1}^j - x^j\|. \tag{13}$$

The second term can be bounded by $(i-1)\varepsilon\|x^j\|/d$ by the induction hypothesis. Therefore, in order to prove the induction, it is enough to show that the first term is bounded by $\varepsilon/d\|x^j\|$.

First, we define some notation. For any two layer $i \leq j$, denote by $M^{i,j}$ the operator for composition of these layers and $J_x^{i,j}$ be the Jacobian of this operator at input x (a matrix whose p, q is the partial derivative of the pth output coordinate with respect to the q'th input input). Therefore, we have $x^j = M^{i,j}(x^i)$[1]. Furthermore, since the activation functions are ReLU, we have $M^{i,j}(x^i) = J_{x^i}^{i,j} x^i$.

We decompose the first term in Eq. 13 into two error terms one of which corresponds to the error propagation through the network if activation (which ReLU units are 0) were fixed and the other one is the error caused by change in the activations:

$$\begin{aligned}
\|(\tilde{x}_i^j - \tilde{x}_{i-1}^j)\| &= \|M^{i,j}(\tilde{A}^i\phi(\tilde{x}^{i-1})) - M^{i,j}(A^i\phi(\tilde{x}^{i-1}))\| \\
&= \|M^{i,j}(\tilde{A}^i\phi(\tilde{x}^{i-1})) - M^{i,j}(A^i\phi(\tilde{x}^{i-1})) + J_{x^i}^{i,j}(\Delta^i\phi(\tilde{x}^{i-1})) - J_{x^i}^{i,j}(\Delta^i\phi(\tilde{x}^{i-1}))\| \\
&\leq \|J_{x^i}^{i,j}(\Delta^i\phi(\tilde{x}^{i-1}))\| + \|M^{i,j}(\tilde{A}^i\phi(\tilde{x}^{i-1})) - M^{i,j}(A^i\phi(\tilde{x}^{i-1})) - J_{x^i}^{i,j}(\Delta^i\phi(\tilde{x}^{i-1}))\|
\end{aligned}$$

where $\Delta^i = \tilde{A}^i - A^i$. To bound the first term, we can apply Lemma 8 with the set $G = \{(J_{x^i}^{i,j}, x^i) | x \in S, j \geq i\}$ which has size at most dm (at most d $J_{x^i}^{i,j}$ for each of the m x_i). We will also need to define several quantities that measure how much error propagates through the network. The term can be bounded as follows:

$$\begin{aligned}
&\|J_{x^i}^{i,j}\Delta^i\phi(\tilde{x}^{i-1})\| \\
&\leq (\varepsilon\mu_i\mu_{i\to}/6cd)\|J_{x^i}^{i,j}\|_F\|A^i\|_F\|\phi(\tilde{x}^{i-1})\| && \text{Lemma 8} \\
&\leq (\varepsilon\mu_i\mu_{i\to}/6cd)\|J_{x^i}^{i,j}\|\|A^i\|_F\|\tilde{x}^{i-1}\| && \text{Lipschitzness of the activation function} \\
&\leq (\varepsilon\mu_i\mu_{i\to}/3cd)\|J_{x^i}^{i,j}\|_F\|A^i\|_F\|x^{i-1}\| && \text{Induction hypothesis} \\
&\leq (\varepsilon\mu_i\mu_{i\to}/3d)\|J_{x^i}^{i,j}\|_F\|A^i\|_F\|\phi(x^{i-1})\| && \text{Activation Contraction} \\
&\leq (\varepsilon\mu_{i\to}/3d)\|J_{x^i}^{i,j}\|_F\|A^i\phi(x^{i-1})\| && \text{Layer Cushion} \\
&= (\varepsilon\mu_{i\to}/3d)\|J_{x^i}^{i,j}\|_F\|x^i\| && x^i = A^i\phi(x^{i-1}) \\
&\leq (\varepsilon/3d)\|x^j\| && \text{Interlayer Cushion}
\end{aligned}$$

where this holds with probability $1 - \frac{\delta}{6dh}$, I think.., because of the first step. However, we want it to hold with probability $1 - \frac{\delta}{2d}$ for them to add up correctly?. The second term can be bounded as:

$$\begin{aligned}
&\|M^{i,j}(\tilde{A}^i\phi(\tilde{x}^{i-1})) - M^{i,j}(A^i\phi(\tilde{x}^{i-1})) - J_{x^i}^{i,j}(\Delta^i\phi(\tilde{x}^{i-1}))\| \\
&= \|(M^{i,j} - J_{x^i}^{i,j})(\tilde{A}^i\phi(\tilde{x}^{i-1})) - (M^{i,j} - J_{x^i}^{i,j})(A^i\phi(\tilde{x}^{i-1}))\| \\
&\leq \|(M^{i,j} - J_{x^i}^{i,j})(\tilde{A}^i\phi(\tilde{x}^{i-1}))\| + \|(M^{i,j} - J_{x^i}^{i,j})(A^i\phi(\tilde{x}^{i-1}))\|.
\end{aligned}$$

[1] Remember that x^i are the preactivations feeding into layer i

Both terms can be bounded using interlayer smoothness condition of the network. First, notice that $A^i\phi(\tilde{x}^{i-1}) = \tilde{x}^i_{i-1}$. Therefore, using by induction hypothesis $\|A^i\phi(\tilde{x}^{i-1}) - x^i\| = \|\tilde{x}^i_{i-1} - x^i\| \leq (i-1)\varepsilon\|x^i\|/d \leq \varepsilon\|x^i\|$ (as $i-1<d$).

Now by interlayer smoothness property, $\|(M^{i,j} - J^{i,j}_{x^i})(A^i\phi(\tilde{x}^{i-1}))\| \leq \frac{\|\tilde{x}^i_{i-1}-x^i\|\|x^j\|}{\|x^i\|\rho_\delta} \leq \frac{\varepsilon\|x^i\|\|x^j\|}{\|x^i\|\rho_\delta} \leq (\varepsilon/3d)\|x^j\|$. (as $\rho_\delta \geq 3d$, by assumption) On the other hand, we also know $\tilde{A}^i\phi(\tilde{x}^{i-1}) = A^i\phi(\tilde{x}^{i-1}) + \Delta^i\phi(\tilde{x}^{i-1})$, therefore $\|\tilde{A}^i\phi(\tilde{x}^{i-1}) - x^i\| \leq \|A^i\phi(\tilde{x}^{i-1}) - x^i\| + \|\Delta^i\phi(\tilde{x}^{i-1})\| \leq ((i-1)\varepsilon/d + \varepsilon/3d)\|x^i\| \leq \varepsilon\|x^i\|$, so again we have $\|(M^{i,j} - J^{i,j}_{x^i})(\tilde{A}^i\phi(\tilde{x}^{i-1}))\| \leq (\varepsilon/3d)\|x^j\|$.

Putting everything together completes the induction, which for the last layer gives us the desired result $\|f_A(x) - f_{\tilde{A}}(x)\| \leq \varepsilon\|f_A(x)\|$, with probability at least $1-\delta/2$.

□

We will now demonstrate that because of the perturbed network producing similar outputs to the original network, its margin of error can't be much smaller. That means that if the original had a certain error with margin γ, the perturbed one has at most that same error, albeit with a smaller margin (but how much smaller it has to be is bounded). In particular, we focus on the zero margin loss for the perturbed network in Lemma 10.

Lemma 10. *For any fully connected network f_A with $\rho_\delta \geq 3d$, any probability $0 < \delta \leq 1$ and any margin $\gamma > 0$, f_A can be compressed (with respect to a random string) to another fully connected network $f_{\tilde{A}}$ such that for any $x \in S$, $\hat{L}_0(f_{\tilde{A}}) \leq \hat{L}_\gamma(f_A)$ and the number of parameters in $f_{\tilde{A}}$ is at most:*

$$\tilde{O}\left(\frac{c^2d^2\max_{x\in S}\|f_A(x)\|_2^2}{\gamma^2}\sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\rightarrow}^2}\right)$$

where μ_i, $\mu_{i\rightarrow}$, c and ρ_δ are layer cushion, interlayer cushion, activation contraction and interlayer smoothness.

Proof. (of Lemma 10) If $\gamma^2 > 2\max_{x\in S}\|f_A(x)\|_2^2$, for any pair (x,y) in the training set we have $|f_A(x)[y] - \max_{i\neq y} f_A(x)[j]|^2 \leq 2\max_{x\in S}\|f_A(x)\|_2^2 \leq \gamma^2$ which means the output margin cannot be greater than γ and therefore $\hat{L}_\gamma(f_A) = 1$ which proves the statement.

If $\gamma^2 \leq 2\max_{x\in S}\|f_A(x)\|_2^2$, by setting $\varepsilon^2 = \gamma^2/2\max_{x\in S}\|f_A(x)\|_2^2$ in Lemma 9, we know that for any $x \in S$, $\|f_A(x) - f_{\tilde{A}}(x)\|_2 \leq \gamma/\sqrt{2}$.

For any (x,y), the margin of the original network $|f_A(x)[y] - \max_{i\neq y} f_A(x)[j]|$ can be reduced by at most $|f_A - f_{\tilde{A}}| \leq \gamma$. Therefore, if for any (x,y), the margin loss on the right hand side is zero $|f_A(x)[y] - \max_{i\neq y} f_A(x)[j]| > \gamma$, and $|f_{\tilde{A}}(x)[y] - \max_{i\neq y} f_{\tilde{A}}(x)[j]| > 0$, and so the classification loss on the left hand size is zero. Therefore whenever f_A classifies well with margin γ, $f_{\tilde{A}}$ classifies well with margin 0, which implies the inequality.

□

Finally, we prove Theorem 7, by bounding the difference between $\hat{L}_0(f_{\tilde{A}})$ and $L_0(f_{\tilde{A}})$ (generalization gap)

Proof. (of Theorem 7) We show the generalization by bounding the covering number of the network with weights $\tilde{A}$. In order to get a covering number, we need to find out the required accuracy for each parameter in the second network to cover the original network. We start by bounding the norm of the weights $\tilde{A}^i$.

Because of positive homogeneity of ReLU activations, we can assume without loss of generality that the network is balanced, i.e for any $i \neq j$, $\|A^i\|_F = \|A^j\|_F = \beta$ (otherwise, one could rebalance the network before approximation and cushion is invariant to this rebalancing). Therefore, for any $x \in S$ we have:

$$\|A^i\| \leq \frac{\|A^i\phi(x^{i-1})\|}{\mu_i\|\phi(x^{i-1})\|} = \frac{\|x^i\|}{\mu_i\|\phi(x^{i-1})\|} \leq \frac{\|x^i\|c}{\mu_i\|x^{i-1}\|}$$

which we can apply layer-wise to get:

$$\beta^d = \prod_{i=1}^{d}\|A^i\| \leq \frac{c\|x^1\|}{\|x\|\mu_1}\prod_{i=2}^{d}\|A^i\| \leq \frac{c^2\|x^2\|}{\|x\|\mu_1\mu_2}\prod_{i=2}^{d}\|A^i\| \leq \frac{c^d\|f_A(x)\|}{\|x\|\prod_{i=1}^{d}\mu_i}$$

By Lemma 9, $\|\tilde{A^i}\|_F \leq \beta(1+1/d)$ (**I can't see why this is. can show using definition with sum that $\tilde{A^i} \leq \beta h^2$..**). We know that $\tilde{A^i} = \frac{1}{k}\sum_{k'=1}^{k}\langle A^i, M_{k'}\rangle M_{k'}$ where $\langle A^i, M_{k'}\rangle$ are the parameters. Therefore, if $\hat{A}^i$ correspond to the weights after approximating each parameter in $\tilde{A^i}$ with accuracy ν, we have: $\|\hat{A}^i - \tilde{A^i}\|_F \leq \sqrt{k}h\nu \leq \sqrt{q}h\nu$ where q is the total number of parameters. Now by Lemma 11, we get:

$$|\ell_\gamma(f_{\tilde{A}}(x), y) - \ell_\gamma(f_{\hat{A}}(x), y)| \leq \frac{2e}{\gamma}\|x\|\left(\prod_{i=1}^{d}\|\tilde{A^i}\|\right)\sum_{i=1}^{d}\frac{\|\tilde{A^i} - \hat{A}^i\|}{\|\tilde{A^i}\|} < \frac{e^2}{\gamma}\|x\|\beta^{d-1}\sum_{i=1}^{d}\|\tilde{A^i} - \hat{A}^i\|_F$$
$$\leq \frac{e^2c^d\|f_A(x)\|\sum_{i=1}^{d}\|\tilde{A^i} - \hat{A}^i\|_F}{\gamma\beta\prod_{i=1}^{d}\mu_i} \leq \frac{qh\nu}{\beta}$$

where the last inequality is because by Lemma 10, $\frac{e^2d\|f_A(x)\|}{\gamma\prod_{i=1}^{d}\mu_i} < \sqrt{q}$ (because $\|f_A(x)\| \leq \max_{x\in S}\|f_A(x)\|$), and c^d was ignored because c=1 for ReLU, I think see Lipsicthzness assumption above (**I think**) Since the absolute value of each parameter in layer i is at most βh, the logarithm of number of choices for each parameter in order to get ε-cover is $\log(qh^2/\varepsilon) \leq 2\log(qh/\varepsilon)$ which results in the covering number $2q\log(kh/\varepsilon)$. Bounding the Rademacher complexity by Dudley entropy integral (See here and here) completes the proof. □

Lemma 11. *Let f_A be a d-layer network with weights $A = \{A^1, \ldots, A^d\}$. Then for any input x, weights A and $\hat{A}$, if for any layer i, $\|A^i - \hat{A}^i\| \leq \frac{1}{d}\|A^i\|$, then we have:*

$$\|f_A(x) - f_{\hat{A}}(x)\|_2 \leq e\|x\|\left(\prod_{i=1}^{d}\|A^i\|\right)\sum_{i=1}^{d}\frac{\|A^i - \hat{A}^i\|}{\|A^i\|}$$