1 Introduction

We aim to adapt the bound of [1] for the case of convolutional networks. Given some assumptions the bound improves over VC bounds. Specifically a good VC bound for fully connected neural networks [4] scales like:

\[ L_0(f_w) \leq \hat{L}_\gamma(f_w) + O(\sqrt{\frac{k^2d^2}{m}}) \]  \( (1) \)

where \( k \) is the number of layers where \( d \) an upper bound on the input and output dimensions across the DNN layers and \( m \) the number of training samples. By contrast the bound obtained by [1] scales like:

\[ L_0(f_w) \leq \hat{L}_\gamma(f_w) + O(\sqrt{\frac{k^2d}{m}}). \]  \( (2) \)

This is a significant improvement over the VC bound.

1.1 Contributions

• We adapt this result to the convolutional setting, by decoupling the spectral norm of convolutional layers from their input and output dimension. We separate two cases. In the first case we analyze the effect of the sparsity of the convolutional layers. In the second case we consider both the sparsity and the weight sharing properties of convolutional layers.

• We also use a different spectral norm formula for the fully connected layers leading to a slightly tighter bound.

2 PAC -Bayes

We state here some results from PAC -Bayes theory. We will consider the binary classification task with an input space \( \mathcal{X} \) and label set \( \mathcal{Y} = \{+1, -1\} \). Let \( D \) be the (unknown) true distribution on \( \mathcal{X} \times \mathcal{Y} \). Let \( \mathcal{H} \) be the hypothesis class of...
functions \( f : \mathcal{X} \Rightarrow \mathcal{Y} \). Let \( \mathcal{P} \) be the space of probability distributions on \( \mathcal{H} \). We consider \( 0,1 \)-valued loss functions \( L : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y}) \Rightarrow \{0,1\} \).

**Definition 2.1.** Let \( Q \in \mathcal{P} \). Define:

\[
\begin{align*}
\text{Risk of } Q & \quad L(Q, \mathcal{D}) = E_{(x,y) \sim \mathcal{D}} E_{h \sim Q}[L(h; (x,y))] \\
\text{Empirical Risk of } Q & \quad L(Q, \mathcal{S}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} E_{h \sim Q}[L(h; (x,y))] \tag{3}
\end{align*}
\]

**Theorem 2.1.** (McAllester) \( \forall \mathcal{D}, \forall \mathcal{H}, \forall P \in \mathcal{P}, \forall \delta > 0 \), we have with probability at least \( 1 - \delta \) over \( S \sim \mathcal{D}^m : \forall Q \in \mathcal{P} \)

\[
KL(L(Q; S)||L(Q; \mathcal{D})) \leq KL(Q||P) + \log\left(\frac{m+1}{\delta}\right) \tag{4}
\]

**Note:** The distribution \( P \) needs to be a proper Bayesian prior, in that it should not be derived from the data. The distribution \( Q \) need not be a proper Bayesian posterior, in that it can be any distribution.

After a few calculations we can obtain the following more convenient form:

**Corollary 2.1.** \( \forall \mathcal{D}, \forall \mathcal{H}, \forall P \in \mathcal{P}, \forall \delta > 0 \), we have with probability at least \( 1 - \delta \) over \( S \sim \mathcal{D}^m : \forall Q \in \mathcal{P} \)

\[
L(Q; S) \leq L(Q; \mathcal{D}) + \sqrt{KL(Q||P) + \log\left(\frac{m+1}{\delta}\right)} \tag{5}
\]

Note that the above relates two probability distributions, as the predictor is a random variable over it’s posterior distribution. However commonly we have access to a specific predictor found by empirical risk minimisation. In this setting it is possible to constrain the posterior appropriately so that we obtain a bound for a single predictor.

### 3 Generalization Error Bound

We now specify the above PAC-Bayes bound for our setting. Specifically let \( f_w \) be any predictor (not necessarily a neural network) learned from the training data and parameterized by \( w \). We consider a posterior \( Q \) over the predictors of the form \( f_{w+u} \), where \( u \) is a random variable whose distribution can depend on the training data. As before we assume a prior distribution \( P \) which should be a proper Bayesian prior and cannot depend on the training data. Then with probability at least \( 1 - \delta \) we get:

\[
E_u[L_0(f_{w+u})] \leq E_u[\hat{L}_0(f_{w+u})] + O\left(\frac{2KL(w+u||P) + \ln \frac{2m}{\delta}}{m-1}\right) \tag{6}
\]

We now restate some useful lemmas from [1].
Lemma 3.1. Let \( f_w(x) : \mathcal{X} \Rightarrow \mathbb{R}^k \) be any predictor (not necessarily a neural network) with parameters \( w \), and \( P \) be any distribution on the parameters that is independent of the training data. Then with probability \( \geq 1 - \delta \) over the training set of size \( m \), for any random perturbation \( u \) s.t. \( \mathbb{P}_u[\max_{x \in \mathcal{X}} |f_{w+u}(x) - f_w(x)|_2 \leq \frac{\gamma}{4}] \geq \frac{1}{2} \), we have:

\[
L_0(f_w) \leq \hat{L}_\gamma(f_w) + O\left(\sqrt{\text{KL}(w||P) + \frac{\ln \frac{m}{\delta}}{m-1}}\right)
\]

where \( \gamma, \delta > 0 \) are constants.

Let’s look at some intuition behind this bound. It links the empirical risk \( \hat{L}_\gamma(f_w) \) of the predictor to the true risk \( L_0(f_w) \), for a specific predictor and not a posterior distribution of predictors. We have also moved to using a margin \( \gamma \) based loss, this is essential step in order to remove the posterior assumption. The perturbation \( u \) quantifies how the true risk would be affected by choosing a bad predictor. The condition \( \mathbb{P}_u[\max_{x \in \mathcal{X}} |f_{w+u}(x) - f_w(x)|_2 \leq \frac{\gamma}{4}] \geq \frac{1}{2} \) can be interpreted as choosing a posterior with small variance, sufficiently concentrated around the current empirical estimate \( w \), so that we can remove the randomness assumption with high confidence.

How small should we choose the variance of \( u \)? The choice is complicated because the KL term in the bound is inversely proportional to the variance of the perturbation. Therefore we need to find the largest possible variance for which our condition holds.

![Figure 1: Overlap between \( x_1 \sim N(0, \sigma^2) \) and \( x_2 \sim N(4, \sigma^2) \)]

The basis of our analysis is the following perturbation bound from [1] on the output of a DNN:

Lemma 3.2. (Perturbation Bound). For any \( B, d > 0 \), let \( f_w : \mathcal{X}_{B,n} \Rightarrow \mathbb{R}^k \) be a \( d \)-layer network with ReLU activations. Then for any \( w \), and \( x \in \mathcal{X}_{B,n} \), and perturbation \( u = \text{vec}(\{U_i\}_{i=1}^d) \) such that \( ||U_i||_2 \leq \frac{1}{d} ||W_i||_2 \), the change in the output of the network can be bounded as follows:

\[
|f_{w+u}(x) - f_w(x)|_2 \leq eB^{d-1} \sum_i ||U_i||_2
\]
where $e$, $B$ and $d_i^{d_i-1}$ are considered as constants after an appropriate normalization of the layer weights.

We note that correctly estimating the spectral norm of the perturbation at each layer is critical to obtaining a tight bound. Specifically if we exploit the structure of the perturbation we can increase significantly the variance of the added perturbation for which our condition holds.

4 Layerwise Perturbations

We need to find the maximum variance for which

$$P_u[\max_{x \in X} |f_{w+u}(x) - f_w(x)|_2 \leq \frac{\gamma}{4}] \geq \frac{1}{2}. \quad (9)$$

We will use two results for the norms of random matrices with independent entries as well as the structure of the convolutional layers.

4.1 Fully Connected Layers

We will use the following bound from [2]:

**Theorem 4.1.** Let $A$ be a $d_2 \times d_1$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$

$$P(||A||_2 \geq \sqrt{d_{i1}} + \sqrt{d_{i2}} + t) \leq e^{-t^2/2} \sigma^2. \quad (9)$$

Noting that $||\sigma U_i||_2 = \sigma||U_i||_2$ we can use this directly to obtain a bound on the spectral norm of a fully connected layer noise $||U_i||_2$.

**Corollary 4.1.1.** For $u \sim N(0, \sigma^2 I)$ and a given fully connected layer $W_i \in \mathbb{R}^{d_2 \times d_1}$ we have for the spectral norm of $U_i$:

$$P(||U_i||_2 - \sigma(\sqrt{d_{i1}} + \sqrt{d_{i2}}) \geq t)) \leq e^{-t^2/2\sigma^2} \quad (10)$$

The result is a bit tighter compared to the method in [1].

4.2 Convolutional Layers

4.2.1 Without Weight Sharing

For the case of convolutional layers we adapt the following theorem from [3]:

**Theorem 4.2.** Let $A$ be a $d_2 \times d_1$ random rectangular matrix with $A_{ij} = \xi_{ij}\psi_{ij}$ where $\{\xi_{ij} : 1 \leq i \leq d_2, 1 \leq j \leq d_1\}$ are independent $N(0,1)$ random variables and $\{\psi_{ij} : 1 \leq i \leq d_2, 1 \leq j \leq d_1\}$ are scalars. Then:

$$P(||A||_2 \geq (1+\epsilon)(\sigma_1 + \sigma_2 + \frac{5}{\sqrt{\log(1+\epsilon)}\log(d_2 \land d_1)} + t)) \leq e^{-t^2/2\sigma^2} \quad (11)$$

for any $0 \leq \epsilon \leq 1/2$ and $t \geq 0$ with:
\[
\sigma_1 := \max_i \sqrt{\sum_j \psi_{ij}^2} \quad \sigma_2 := \max_i \sqrt{\sum_j \psi_{ij}^2} \quad \sigma_* := \max_{ij} |\psi_{ij}| \quad (12)
\]

Note that the theorem invoked for the convolutional layers and the theorem invoked for the fully connected layers coincide in the case of dense matrices.

**Corollary 4.2.1.** For \( u \sim \mathcal{N}(0, \sigma^2 I) \) and a given convolutional layer \( W_i \in \mathbb{R}^{d_2 \times d_1} \) with \( a_i \) input channels, \( b_i \) output channels, and convolutional filters \( f_{ij} \in \mathbb{R}^{q \times q} \) the spectral norm of \( U_i \) is bounded as:

\[
P(\|U_i\|_2 - \sigma (1 + \epsilon) (q\sqrt{a_i} + q\sqrt{b_i} + 5\sqrt{\frac{\log(d_{i1})}{\log(1 + \epsilon)}}) \geq t) \leq e^{-t^2/2\sigma^2(1+\epsilon)^2} \quad (13)
\]

**Proof.** We will consider standard normal noise first. A convolutional layer is characterised by its output channels. For each output channel each input channel is convolved with an independent filter, the output of the output channel is then the sum of the results of these convolutions. Below we show a typical example of such a convolutional architecture:

![Figure 2: A Convolutional DNN architecture](image)

given the above the structure of the layer weights is a concatenation of block toeplitz matrices. We plot for the case of one dimensional signals the implied structure.
we need to evaluate $\sigma_1 := \max_i \sqrt{\sum_j \psi_{ij}^2}$, $\sigma_2 := \max_j \sqrt{\sum_i \psi_{ij}^2}$ and $\sigma_* := \max_{ij} |\psi_{ij}|$ for a matrix this matrix. Below we plot what these sums represent:
For $\sigma_1$ we can find an upper bound, by considering that the sum for a given filter and a given pixel location represents the maximum number of overlaps for all 2d shifts. For the case of 2d this is $(q)^2$ equal to the support of the filters and we also need to consider all input channels. We then get

$$\sigma_1 := \max_i \sqrt{\sum_j \psi_{ij}^2} \leq \sqrt{\sum_i \sum_{q^2} \psi_{ij}^2} = \sqrt{\sum_i \sum_{q^2} \psi_{ij}^2} = \sqrt{a_i q^2} = q \sqrt{a_i}$$  \hspace{1cm} (14)$$

For $\sigma_2$ we need to consider that each column in the matrix represents a concatenation of convolutional filters $f \in \mathbb{R}^{3 \times 3}$. Then it is straightforward to derive that:

$$\sigma_1 := \max_i \sqrt{\sum_j \psi_{ij}^2} \leq \sqrt{\sum_i \sum_{q^2} \psi_{ij}^2} = \sqrt{\sum_i \sum_{q^2} \psi_{ij}^2} = \sqrt{b_i q^2} = q \sqrt{b_i}$$  \hspace{1cm} (15)$$

Furthermore it is trivial to show that $\sigma_\star = 1$. The theorem results extends trivially to $\sigma > 0$ by considering that $\|\sigma U_i\|_2 = \sigma \|U_i\|_2$.

4.2.2 With Weight Sharing

We now turn our attention to the case of weight sharing. We start from the theorem employed in [1]. This originates from [5].

**Theorem 4.3.** Consider a finite sequence $\{B_k\}$ of fixed matrices with dimension $d_1 \times d_2$, and let $\{\xi_k\}$ be a finite sequence of independent standard normal or independent Rademacher random variables. Define the variance parameter

$$\sigma^2 := \max \{\|\sum_k B_k B_k^*\|, \|\sum_k B_k^* B_k\|\}$$  \hspace{1cm} (16)$$

Then, for all $t \geq 0$

$$\mathbb{P}(\|\sum_k \xi_k B_k\| \geq t) \leq (d_1 + d_2)e^{-t^2/2\sigma^2}$$  \hspace{1cm} (17)$$
We get taking into account weight sharing:

**Corollary 4.3.1.** For $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 I)$ and a given convolutional layer $\mathbf{W}_i \in \mathbb{R}^{d_2 \times d_1}$ with $a_i$ input channels, $b_i$ output channels, and convolutional filters $f_{ij} \in \mathbb{R}^{q \times q}$ the spectral norm of $\mathbf{U}_i$ is bounded as:

$$
P(||\mathbf{U}_i||_2 \geq t) \leq \left( d_1 + d_2 \right) e^{-t^2/2\sigma^2 (q^2 b)}
$$

(18)

**Proof.** We will show a small example. Given as before matrix with $a = 3$, $b = 2$ and $q^2 = 1$ we get:

$$
\sum_k \xi_k \mathbf{B}_k = \begin{bmatrix} \xi_1 A & \xi_2 B & \xi_3 C \\ \xi_4 D & \xi_5 E & \xi_6 F \end{bmatrix}
$$

(19)

Where each $A, B, C, D, E, F$ represents all the shifts of a perturbation for a given convolutional filter with support $q^2 = 1$. Then:

$$
\sum_k \mathbf{B}_k \mathbf{B}_k^* = \begin{bmatrix} A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix}
$$

(20)

with $K_1 = AA^T + BB^T + CC^T$, $K_2 = DD^T + EE^T + FF^T$ and:

$$
AA^T = BB^T = CC^T = DD^T = EE^T = FF^T = I \sigma^2
$$

(21)

Then:

$$
\sum_k \mathbf{B}_k \mathbf{B}_k^* = \begin{bmatrix} I \sigma^2 & 0 \\ 0 & I \sigma^2 \end{bmatrix}
$$

(22)

This is the result for the trivial support $q = 1$ we need to sum over the support for $q > 1$:

$$
\sum_{q^2} \sum_k \mathbf{B}_k \mathbf{B}_k^* = \begin{bmatrix} I \sigma^2 q^2 & 0 \\ 0 & I \sigma^2 q^2 \end{bmatrix}
$$

(23)

we get an equivalent result for $\sum_k \mathbf{B}_k^* \mathbf{B}_k$ and the corollary is obtained by noting that $b \gg a$. 

\qed
Note that this increases the spectral norm of the layer pertubation. This in turn lowers the maximum allowed variance, making our bound looser. How looser does our bound become? We answer this in the experiment section, with a result on a VGG16-like architecture. This result is counterintuitive, in particular weight sharing should reduce the generalization error. We conjecture that the difference results from the formulas used to derive the bound on the spectral norm of the pertubation.

5 Pertubation Bound

Let $K = \{1, ..., K_1\}$ be the set of all the convolutional layers and $T = \{1, ..., T_1\}$ be the set of all dense layers. From Lemma 3.2 we need to find the concentration inequality for the sum $\sum_i \|U_i\|_2$.

**Lemma 5.1.** For $\sum_i \|U_i\|_2$ we have:

$$\mathbb{P}_{u \sim \mathcal{N}(0, \sigma^2 I)}(\sum_i \|U_i\|_2 - \sigma C_1 \geq t) \leq e^{-t^2/2 \sum_i \sigma}$$

(24)

**Proof.** We note that $\|U_i\|_2$ are subgaussian random variables with $\sigma_i = \sigma$. Due to the linearity of the expected value we can obtain:

$$\mathbb{E} \sum_i \|U_i\|_2 = \sum_i \mathbb{E}[\|U_i\|_2] = \sigma C_1$$

(25)

where:

$$C_1 = \sum_{i \in K} \left[ q \sqrt{a_i} + q \sqrt{b_i} + 5 \sqrt{\frac{\log(d_{i1})}{\log(1 + \epsilon)}} \right] + \sum_{j \in T} \left[ \sqrt{d_{j1}} + \sqrt{d_{j2}} \right]$$

(26)

We can then apply Hoeffding's inequality to bound the concentration of $\sum_i \|U_i\|_2$ around the mean $\sigma C_1$ and we get the result trivially.

We are now ready to find the value for the **standard deviation** parameter $\sigma$:

**Lemma 5.2.** (Pertubation Bound). For any $B, d > 0$, let $f_w : \mathcal{X}_{B,n} \rightarrow \mathbb{R}^k$ be a $d$-layer network with ReLU activations. Then for any $w$, and $x \in \mathcal{X}_{B,n}$, and a perturbation for $u \sim \mathcal{N}(0, \sigma^2 I)$, for any $\gamma > 0$ with $\sigma = \frac{\gamma}{42BD^{d-1}(C_1 + \sqrt{2k \ln(2)})}$ we have:

$$\mathbb{P}_u[\max_{x \in \mathcal{X}} |f_{w+u}(x) - f_w(x)|_2 \leq \frac{\gamma}{4}] \geq \frac{1}{2}$$

(27)

where $e$, $B$ and $d^{d-1}$ are considered as constants after an appropriate normalization of the layer weights.
Proof. We start by using Lemma 5.1:

$$\mathbb{P}_{u \sim N(0, \sigma^2 I)} \left( \sum_i ||U_i||_2 - \sigma C_1 \geq t \right) \leq e^{-t^2/2 \sum_i \sigma} \quad (28)$$

Next we assume that $$e^{-t^2/2 \sum_i \sigma} \leq \frac{1}{2} \Rightarrow t \geq \sigma \sqrt{2 \ln(2)}$$ we then get trivially:

$$\mathbb{P}_{u \sim N(0, \sigma^2 I)} \left( \sum_i ||U_i||_2 \leq \sigma (C_1 + \sqrt{2 \ln(2)}) \right) \geq \frac{1}{2} \quad (29)$$

We then apply this result to lemma 3.2 with probability $$\geq \frac{1}{2}$$ we have:

$$|f_{w+u}(x) - f_w(x)|_2 \leq eB\beta^{d-1} \sum_i ||U_i||_2 \leq eB\beta^{d-1} \sigma (C_1 + \sqrt{2 \ln(2)}) \leq \frac{\gamma}{4} \quad (30)$$

where the final inequality results by choosing the maximum allowed standard deviation $$\sigma = \frac{\gamma}{eB\beta^{d-1}(C_1 + \sqrt{2 \ln(2)})}$$.

We our now ready to state our main result:

**Theorem 5.3.** (Generalization Bound). For any $$B,d,h > 0$$, let $$f_w : \mathcal{X}_{B,n} \Rightarrow \mathbb{R}^k$$ be a d-layer network with ReLU activations. Then for any $$\gamma, \delta > 0$$, with probability $$\geq 1 - \delta$$ over the training set of size $$m$$, for any $$w$$ and a pertubation for $$u \sim N(0, \sigma^2 I)$$ for any $$\gamma > 0$$ with $$\sigma = \frac{\gamma}{eB\beta^{d-1}(C_1 + \sqrt{2 \ln(2)})}$$ we have:

$$L_0(f_w) \leq \hat{L}_\gamma(f_w) + \mathcal{O}\left(\sqrt{\frac{C_1^2}{m}}\right) \quad (31)$$

Proof. We now calculate the KL-term in Lemma 4.1 with the chosen distributions for $$P$$ and $$u$$, for the value of $$\sigma = \frac{\gamma}{eB\beta^{d-1}(C_1 + \sqrt{2 \ln(2)})}$$.

$$KL(w + u||P) \leq \frac{|w|^2}{2\sigma^2} \leq \mathcal{O}(B^2 C_1 \prod_{i=1}^k \frac{||W_i||^2}{\gamma^2} \sum_{i=1}^k \frac{||W_i||^F}{||W_i||_2^F}) \quad (32)$$

where we have used the fact that both $$P$$ and $$u$$ follow multivariate Gaussian distributions. Theorem 4.3 results from substituting the value of $$KL$$ into Equation 17.

$$L_0(f_w) \leq \hat{L}_\gamma(f_w) + \mathcal{O}\left(\sqrt{\frac{B^2C_1^2 \prod_{i=1}^k \frac{||W_i||^2}{\gamma^2} \sum_{i=1}^k \frac{||W_i||^F}{||W_i||_2^F} + \ln \frac{km}{\delta}}}{\gamma^2 m}\right) \quad (33)$$

Note that the derivation is a bit more involved requiring a union bound in respect to different values of $$\beta$$ but we omit those for clarity given that the calculations are identical to those in [1].
6 Experiments

We plot here the constant $C_i$ for different layers of the VGG-16 architecture. We see that for convolutional layers the original approach gives too pessimistic estimates, orders of magnitude higher than our approach.

We take also a closer look at the training sample estimate for a VGG16-like architecture. We assume the same structure as the VGG16 architecture but make a strong assumption that $||W_i||_2 \approx 1$, $\forall i \in \{1,\ldots,K\}$. Based on the above framework our bound scales like

$$L_0(f_w) \leq \hat{L}_\gamma(f_w) + O(\sqrt{\frac{10^6}{m}}) \quad (34)$$

The VGG16 architecture was trained for the ILSVRC-2012 visual recognition which has $1.2 \times 10^6$ training samples making our estimate realistic. By comparison the VGG16-like architecture contains $\approx 1.4 \times 10^8$ parameters, our analysis here suggests that despite using much fewer training samples than the total number of trainable parameters a VGG16-like architecture is still able to avoid overfitting and have small generalization error.

In the figure we plot also the result for weight sharing. We see that the effect is large but manageable. In particular even with union bounding we still get:

$$L_0(f_w) \leq \hat{L}_\gamma(f_w) + O(\sqrt{\frac{10^6}{m}}) \quad (35)$$

11
7 Conclusion

We have presented an adaptation of the PAC-Bayesian bound \([1]\) for convolutional neural networks. In our analysis we have taken into account the sparsity of the convolutional layers of a DNN and also the effect of weight sharing. We present some empirical results that show that this bound leads to realistic estimates for the sample complexity of a VGG16-like architecture.

8 APPENDIX

Why the bound in Corollary 4.1.1. is slightly better than the one in \([1]\). For \([1]\) we have:

\[
P(\|U_i\|_2 \geq t) \leq 2he^{-t^2/2h\sigma^2} \tag{36}
\]

then \(2e^{-t^2/2h\sigma^2} \leq \frac{1}{2} \Rightarrow t \geq \sqrt{2h\ln(4h)}\). And we get:

\[
P(\|U_i\|_2 \leq \sigma(\sqrt{2h\ln(4h)})) \geq \frac{1}{2} \tag{37}
\]

For our bound we have:

\[
P(\|U_i\|_2 - \sigma(2\sqrt{h}) \geq t) \leq e^{-t^2/2\sigma^2} \tag{38}
\]

then \(e^{-t^2/2\sigma^2} \leq \frac{1}{2} \Rightarrow t \geq \sqrt{\ln(2)}\). And we get:

\[
P(\|U_i\|_2 \leq \sigma(2\sqrt{h} + \sqrt{\ln(2)})) \geq \frac{1}{2} \tag{39}
\]

We plot below the upper bound for our method and the method of \([1]\) for a range of values for \(h\).

![Figure 7: Comparison for the case of \(d_1 = d_2 = h\) of the \(\|U\|_2\).](image)
References

[1] Neyshabur, Behnam, et al. "A pac-bayesian approach to spectrally-normalized margin bounds for neural networks." arXiv preprint arXiv:1707.09564 (2017).

[2] Vershynin, Roman. "Introduction to the non-asymptotic analysis of random matrices." arXiv preprint arXiv:1011.3027 (2010).

[3] Bandeira, Afonso S., and Ramon van Handel. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries." The Annals of Probability 44.4 (2016): 2479-2506.

[4] Harvey, Nick, Chris Liaw, and Abbas Mehrabian. "Nearly-tight VC-dimension bounds for piecewise linear neural networks." arXiv preprint arXiv:1703.02930 (2017).

[5] Tropp, Joel A. "User-friendly tail bounds for sums of random matrices." Foundations of computational mathematics 12.4 (2012): 389-434.