LESSONS FROM THE RADEMACHER COMPLEXITY FOR DEEP LEARNING

Jure Sokolić[1], **Raja Giryes**[2], **Guillermo Sapiro**[3], **Miguel R. D. Rodrigues**[1]
[1] Department of E&EE, University College London, London, UK
[2] School of EE, Faculty of Engineering, Tel-Aviv University, Tel Aviv, Israel
[3] Department of ECE, Duke University, Durham, North Carolina, USA

ABSTRACT

Understanding the generalization properties of deep learning models is critical for successful applications, especially in the regimes where the number of training samples is limited. We study the generalization properties of deep neural networks via the empirical Rademacher complexity and show that it is easier to control the complexity of convolutional networks compared to general fully connected networks. In particular, we justify the usage of small convolutional kernels in deep networks as they lead to a better generalization error. Moreover, we propose a representation based regularization method that allows to decrease the generalization error by controlling the coherence of the representation. Experiments on the MNIST dataset support these foundations.

1 INTRODUCTION

In the recent years deep neural networks have been used to achieve state-of-the-art results in image recognition, speech recognition and many other fields (LeCun et al., 2015). An important property of any learning method is its generalization error that informs us how well the performance on the training set is aligned with the performance on the testing set. An important measure of the generalization error is the empirical Rademacher complexity (ERC) (Bartlett & Mendelson, 2002), which we consider in this work. The smaller the ERC, the better is the generalization error.

Previous works have bounded the ERC in terms of the network's width and depth (Bartlett & Mendelson, 2002), in terms of the norm of the weight matrices (Neyshabur et al., 2015) or in terms of margin bounds (Sun et al., 2015). Another line of works showed how extensions of the dropout can reduce the ERC in the network (Wan et al., 2013; Huang et al., 2015a). Note that the ERC is not the only measure for generalization error. See Shalev-Shwartz & Ben-David (2014); Xu & Mannor (2012); Giryes et al. (2015); Huang et al. (2015b).

The main contributions of this work are the following:

- First, we compare the ERCs of a fully connected deep network and a convolutional neural network (CNN) with weight norm regularizations. We show that a smaller size of convolutional filters reduces the (upper bound of the) ERC.
- Second, we propose an alternative approach for bounding the ERC of deep networks by controlling the geometry of the representation in the last layer. In particular, we show that we can control the ERC of a deep network by enforcing the representation of the data at the last layer to have a small coherence.

1.1 BACKGROUND

We consider deep neural network for binary classification. The binary classifier is given as $g(\mathbf{x}) = \mathbf{v}^T f(\mathbf{x}) \lessgtr 0$, where $\mathbf{v} \in \mathbb{R}^{M_L}$ represents the normalized linear classifier operating on the output of the deep network with input vector $\mathbf{x} \in \mathbb{R}^N$. The function $f : \mathbb{R}^N \to \mathbb{R}^{M_L}$ represents a deep neural network with L layers. It is computed as

$$f(\mathbf{x}) = f^L(\mathbf{x}) = \left[\mathbf{W}_L^T f^{L-1}(\mathbf{x})\right]_+ , \quad f^i(\mathbf{x}) = \left[\mathbf{W}_i^T f^{i-1}(\mathbf{x})\right]_+ , i = 1, \dots, L , \tag{1}$$

where $f^0(\mathbf{x}) = \mathbf{x}$, $[\cdot]_+ = \max(\cdot, 0)$ represents the element-wise ReLU non-linearity, and $\mathbf{W}_i \in \mathbb{R}^{M_i \times M_{i-1}}, i = 1, \ldots, L$, are the weight matrices. Note that $M_0 = N$.

In CNN the weight matrices are structured. In particular, assuming that signals have only one "spatial" dimension, the weight matrix $\mathbf{W}_i \in \mathbb{R}^{M_i \times M_{i-1}}$ is implicitly defined by the tuple $(\mathbf{K}_i, a_i, s_i)$, where $\mathbf{K}_i \in \mathbb{R}^{a_i \cdot k_{i-1} \times k_i}$ is a kernel matrix and typically $k_i \ll M_i$. k_i represents the number of filters of the kernel $\mathbf{K}_i$ (note that $k_0 = 1$), $a_i \in \mathbb{N}_+$ represents the filter size or the "receptive field" of the filter and $s_i \in \mathbb{N}_+$ represents the "stride" or sub-sampling factor of the convolutional layer.

We will denote the weight matrix associated with the convolutional layer as $\mathbf{W}_i^C = \{\mathbf{K}_i\}_{kl}, (k, l) = (1 + j \cdot s_i, 1 + j, 1 + j \cdot k_i)$, where $j = 0, \ldots, (M_{i-1}/k_{i-1} - a_i)/s_i$ and $\{\cdot\}_{kl}$ denotes the matrix block at row index k and column index l. The output of the i-th layer is then $[(\mathbf{W}_i^C)^T f^{i-1}(\mathbf{x})]_+$.

We will consider the ERC as a proxy to the generalization error of deep networks. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are independent samples drawn from a distribution P defined on $\mathbb{R}^N$, and let $\mathcal{G}$ be a class of functions that map $\mathbb{R}^N$ to $\mathbb{R}$. The ERC of $\mathcal{G}$ is

$$\hat{R}_m(\mathcal{G}) = \mathbb{E}_{\xi_i \in \{\pm 1\}} \left[\frac{1}{m} \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{m} \xi_i g(\mathbf{x}_i) \right| \right], \tag{2}$$

where $\xi_i \in \{\pm 1\}$, $i = 1, \ldots, m$ are independent Rademacher distributed random variables. See Bartlett & Mendelson (2002) for a detailed description of the relation between the ERC and the generalization error.

2 Weight Norm Based Regularization

In this section we first review the recent result by Neyshabur et al. (2015) that bounds the ERC for deep networks with the norm of the weight matrices. We only report the most relevant result for our discussion. The class of binary classification networks with L layers is given by

$$\mathcal{G}_{L,W_L} = \left\{ g : \|\mathbf{v}\|_2 = 1, \prod_{i=1}^{L} \|\mathbf{W}_i\|_F \leq W_L \right\}. \tag{3}$$

Theorem 1 (Theorem 1 in (Neyshabur et al., 2015))**.** *The ERC of a function class $\mathcal{G}_{L,W_L}$ as defined in* (3) *can be upper bounded by*

$$\hat{R}_m(\mathcal{G}_{L,W_L}) \leq \frac{1}{\sqrt{m}} 2^{L+\frac{1}{2}} W_L \max_i \|\mathbf{x}_i\|_2 . \tag{4}$$

We provide a tighter version of this result for CNNs, where $\mathbf{W}_i^C$ is defined via $(\mathbf{K}_i, a_i, s_i)$:

$$\mathcal{G}_{L,K_L}^C = \left\{ g : \|\mathbf{v}\|_2 = 1, \prod_{i=1}^{L} \|\mathbf{K}_i\|_F \leq K_L \right\}. \tag{5}$$

Theorem 2. *The ERC of a function class $\mathcal{G}_{L,K_L}^C$ as defined in* (5) *can be upper bounded by*

$$\hat{R}_m(\mathcal{G}_{L,K_L}^C) \leq \frac{1}{\sqrt{m}} 2^{L+\frac{1}{2}} K_L \prod_{i=1}^{L} \sqrt{a_i} \sqrt{\frac{M_L}{k_L}} \max_i \|\mathbf{x}_i\|_2 . \tag{6}$$

The proof will be presented in an extended version of this paper. Next, we provide a few remarks and comparisons between the ERCs of fully connected and convolutional networks:

- We first compare the factors $\prod_{i=1}^{L} \|\mathbf{W}_i\|_F \leq W_L$ in (4) and $\prod_{i=1}^{L} \sqrt{a_i} \|\mathbf{K}_i\|_F \leq K_L \prod_{i=1}^{L} \sqrt{a_i}$ in (6). The direct application of Theorem 1 to a CNN problem provides a looser bound compared to our proposed Theorem 2 because in a CNN, $\|\mathbf{W}_i\|_F = \sqrt{\frac{M_{i-1}/k_{i-1} - a_i}{s_i}} \|\mathbf{K}_i\|_{2,2}$. For example, if we consider a typical setup with $a_i = 3$ and $s_i = 1$, then $\|\mathbf{W}_i\|_F = \sqrt{M_{i-1}/k_{i-1} - 3} \|\mathbf{K}_i\|_F \gg \sqrt{3} \|\mathbf{K}_i\|_F$ as $M_{i-1}/k_{i-1} - 3 \gg 3$.

- Assuming that K_L is fixed, the factors $\sqrt{a_i}$ in equation (6) imply that smaller filter sizes reduce the ERC and lead to better generalization. This theoretical result is aligned with the designs used in the recent state-of-the-art CNNs, where filters of size 3×3 are used. See Simonyan & Zisserman (2014) or He et al. (2015).
- The bounds in Theorems 1 and 2 exhibit the same convergence rate $\mathcal{O}(1/\sqrt{m})$ and depend in the same way on the training set. The bound in Theorem 2 contains an additional term $\sqrt{M_L/k_L}$, where M_L is the dimension of the representation in the last layer and k_L is the number of filters in the last layer. Note that $K_L \prod_{i=1}^{L} \sqrt{a_i}\sqrt{M_L/k_L} \ll W_L$ for a typical CNN. Note also that M_L can be decreased by using larger sub-sampling factors s_i in the convolutional layers.
- Finally, we note that both bounds include the factor $2^{L+\frac{1}{2}}$, which implies that ERC of a deep network grows exponentially with the depth even when W_L and K_L are bounded. For small values of L this is not a concern, however, recent practical results imply that networks with $L > 100$ can be trained successfully (He et al. (2015)). Therefore, there seems to be a gap between theory and practice, at least for very deep networks.

3 Representation Based Regularization

In this section we propose an alternative approach for bounding the ERC that focuses on the coherence of the representation at the output of the last layer and is independent of the number of layers:

Theorem 3. *Assume that the last layer of the network obeys:* $d_{\min}|\mathbf{x}_i^T\mathbf{x}_j| \leq f(\mathbf{x}_i)^T f(\mathbf{x}_j) \leq d_{\max}|\mathbf{x}_i^T\mathbf{x}_j|$ *for any* $\mathbf{x}_i, \mathbf{x}_j$*, where* $d_{\max}, d_{\min} > 0$ *are constant. Then*

$$\hat{R}_m = \frac{1}{m}\mathbb{E}_{\xi_i \in \{\pm 1\}}\left[\sum_i d_{\max}\|\mathbf{x}_i\|_2^2 + \sum_{i \neq j} \mathbb{1}_{\xi_i = \xi_j} d_{\max}|\mathbf{x}_i^T\mathbf{x}_j| - \sum_{i \neq j} \mathbb{1}_{\xi_i \neq \xi_j} d_{\min}|\mathbf{x}_i^T\mathbf{x}_j|\right] . \quad (7)$$

Therefore, we can conclude that keeping $d_{\max}$ small can reduce the ERC and lead to better generalization. We have tested this idea empirically by adding a regularization term $\lambda \cdot \sum_{i \neq j} f(\mathbf{x}_i)^T f(\mathbf{x}_j)$, where $\lambda > 0$ is a parameter, to the standard cross-entropy loss when training a CNN on the MNIST dataset. Figure 1 shows that this term improves the generalization error.

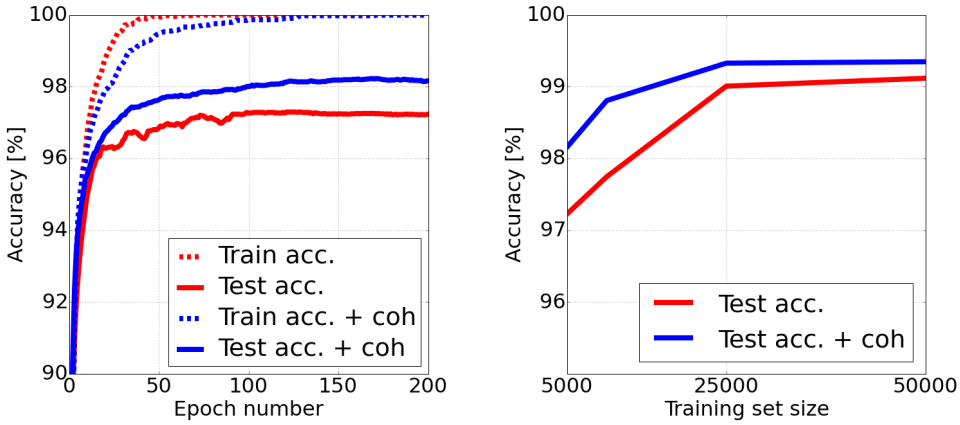

Figure 1: We compare training and testing accuracy of CNNs trained without (red) and with (blue) the proposed regularization term. The network architecture: Conv$(32 \times 5 \times 5)$, Pool(2×2), Conv$(32 \times 5 \times 5)$, Pool(2×2), Full(256), Softmax(10). Left plot shows the behaviour of training and testing accuracies when training with 5000 samples as a function of the number of training epochs. Right plot show the testing accuracies as a function of the number of training samples. We see that the proposed regularization term significantly reduces the generalization error, especially when the number of samples is small.

REFERENCES

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

Raja Giryes, Guillermo Sapiro, and Alex M. Bronstein. Deep Neural Networks with Random Gaussian Weights: A Universal Classification Strategy? *arXiv Prepr. arXiv1504.08291*, Apr 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv Prepr. arXiv1512.03385*, Dec 2015.

Jiaji Huang, Qiang Qiu, Robert Calderbank, and Guillermo Sapiro. GraphConnect: A Regularization Framework for Neural Networks. *arXiv Prepr. arXiv1512.06757*, Dec 2015.

Jiaji Huang, Qiang Qiu, Guillermo Sapiro, and Robert Calderbank. Discriminative Robust Transformation Learning. *Adv. Neural Inf. Process. Syst.*, pp. 1333–1341, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep Learning. *Nature*, 521(7553):436–444, May 2015.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. *arXiv Prepr. arXiv1503.00036*, Feb 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv Prepr. arXiv1409.1556*, Sep 2014.

Shizhao Sun, Wei Chen, Liwei Wang, and Tie-Yan Liu. Large Margin Deep Neural Networks: Theory and Algorithms. *arXiv Prepr. arXiv1506.05232*, Jun 2015.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann LeCun, and Rob Fergus. Regularization of Neural Networks using DropConnect. *Proc. 30th Int. Conf. Mach. Learn.*, pp. 1058–1066, 2013.

Huan Xu and Shie Mannor. Robustness and Generalization. *Mach. Learn.*, 86(3):391–423, 2012.