Spectrally normalized margin bounds for neural networks

Xintian Han & Saad Lahlou
DS-GA 1005: Inference and Representation
Project write-up

December 19, 2017

Contents

1 Introduction

One of the intriguing property of neural networks is that they are able to fit any dataset, and are in the same time able generalize very well on previously unseen data. Zhang et al. [8] empirically showed that several deep models can fit random labels on common datasets with zero training error. This behavior is particularly surprising when it is analyzed through the spectrum of the classical results of statistical learning theory. Indeed, the generalization bounds obtained in this field imply that models with high complexity (this complexity is usually measured in terms of VC-dimension or Rademacher complexity), will have no guarantee of generalizing well. Several works explored the generalization of deep learning [4] [1] [6] [7] [2] [5]. They all tried to relate the generalization performance of neural network with the complexity of the model.

Our work is primarily based mainly on the results put forward in a recent paper by Peter Barlett, Dylan Foster, and Matus Telgarsky ([1]). This paper develops a new measure of complexity that is more adapted for analyzing the behavior of neural networks, and more importantly, derives a margin bound for neural networks, using that measure of complexity. Section 2 summarizes the theoretical bounds and the proof. Section 3 describes all the experiment we design to verify the spectrally-normalized margin bound and compare different complexity/metric normalized margin bounds. Section 4 shows our conclusions.

2 Theoretical results

2.1 Margin bound

The paper defines a concept of spectral complexity. We are studying a network that takes as input a d-dimensional vector, and aims to perform a k-class classification task. The function learned by the network if $\mathcal{F}_\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^k$, where the outputs are the scores for each class. We denote by $\mathcal{A} = (A_1, A_2, ..., A_L)$ our weight matrices and by $\sigma_1, \sigma_2, ..., \sigma_L$ the non-linearities used in the network. Therefore, we have:

$$\mathcal{F}_\mathcal{A}(x) = \sigma_L(A_L \sigma_{L-1}(A_{L-1} \cdots \sigma_1(A_1 x) \cdots))$$

Moreover, we denote by $\rho_1, \rho_2, \cdots, \rho_L$ the Lipschitz constants associated with our nonlinearities. We define the spectral complexity $R_\mathcal{A}$ as:

$$R_\mathcal{A} = \big(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\big) \cdot \big(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\big)^{3/2}$$

Using this spectral complexity, the paper puts forward the following bound for neural networks:

For $(x, y), (x_1, y_1), \ldots, (x_n, y_n)$ drawn iid from any probability distribution over $\mathbb{R}^d \times \{1, \ldots, k\}$, with probability at least $1 - \delta$ over $((x_i, y_i))_{i=1}^n$, for every margin $\gamma > 0$

$$\Pr(\arg\max_j F_\mathcal{A}(x)_j \neq y) \leq \hat{\mathcal{R}}_\gamma(F_\mathcal{A}) + O(\frac{\|X\|_2 R_\mathcal{A}}{\gamma n} \ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}})$$

Where $\hat{\mathcal{R}}_\gamma(F_\mathcal{A})$ is the empirical margin loss of the network on the training dataset. And $\|X\|_2$ is the l2 norm of the training data matrix.

2.2 Proof of the bound

The idea of the proof presented in [1] is to use a classical margin bound that involves the Rademacher complexity:

$$\Pr[\arg\max_i f(x)_i \neq y] \leq \widehat{\mathcal{R}}_\gamma(f) + 2\mathfrak{R}((\mathcal{F}_\gamma)_{|S})) + 3\sqrt{\frac{\ln(1/\delta)}{2n}}$$

The key idea here is to use covering numbers to obtain an upper bound on the Rademacher complexity $\mathfrak{R}((\mathcal{F}_\gamma)_{|S})$ of our network.

More precisely, the authors of the paper use Marvey's sparsification lemma to find an upper bound on the covering number of the outputs of each linear function in the neural network. If we have an input X $\in \mathbb{R}^{n\times d}$(X should be thought of as the training data matrix), such that its norm is bounded, $\|X\|_p \leq b$, with $p \leq 2$ and b a positive real number. Let q be the conjugate exponent of p, and (r,s) be conjugate exponents. Then for all positive reals b, ϵ, we have the following upper bound on the covering number of the outputs XA of the linear mapping associated to the matrix A:

$$\ln\left(\mathcal{N}\left\{XA : A \in \mathbb{R}^{d\times m}, \|A\|_{q,s} \leq a\right\}, \epsilon, \|\cdot\|_2\right) \leq \left\lceil \frac{a^2b^2m^{2/r}}{\epsilon^2} \right\rceil \ln(2dm). \tag{1}$$

For the rest of our analysis, we will use this inequality with the values $q = 2, s = 1, p = 2$. Now that we have this result, one should notice that this covering number bound is only valid for the outputs of one linear layer. Therefore, we should try to combine them in order to obtain a covering bound for the outputs of the whole network, which we will now do using induction.

Let X_i be the output of layer i. Let us suppose that there exists a cover element $\hat{X}_i$ ($\hat{X}_i$ depends on the choices made for the covering elements $\widehat{A}_1, \widehat{A}_2, \cdots, \widehat{A}_{i-1}$ that cover the weight matrices of the network up to layer i). We also know thanks to result (1) that if we fix a desired level of precision ϵ_i there exists a matrix $\widehat{A}_i$ such that $\|A_i\widehat{X}_i - \widehat{A}_i\widehat{X}_i\|_2 \leq \epsilon_i$. The idea is that we now have $\widehat{X}_i$ as a proxy for X_i and $\widehat{A}_i$ as a proxy for the real weight matrix. Using these two proxies, we build $\widehat{X}_{i+1} = \sigma_i(\widehat{A}_i \cdot \widehat{X}_i)$ as proxy for $X_{i+1} = \sigma_i(A_i \cdot X_i)$. Incidentally, if we use the fact that σ_i is ρ_i-Lipschitz we have the following upper bound:

$$\begin{aligned} \|X_{i+1} - \widehat{X}_{i+1}\|_2 &\leq \rho_i\|A_iX_i - \widehat{A}_i\widehat{X}_i\|_2 \\ &\leq \rho_i\|A_iX_i - A_i\widehat{X}_i\|_2 + \|A_i\widehat{X}_i - \widehat{A}_i\widehat{X}_i\|_2 \\ &\leq \rho_i\|A_i\|_\sigma\|X_i - \widehat{X}_i\|_2 + \rho_i\epsilon_i \end{aligned} \tag{2}$$

We can now use this result in order to obtain a covering number bound on the entire network. We start by fixing $s_1, s_2, \cdots, s_L$ positive real numbers, that will be bounds of the spectral norms of the weight matrices, we also define reference matrices $M_1, M_2, \cdots, M_L$ for each layer, and fix $b_1, b_2, \cdots, b_L$ that will be upper bounds on $\|A_i^T - M_i^T\|_{2,1}$. We define $\mathcal{H}_X$, the set of all possible outputs obtained when taking the matrix X as input. We have:

$$\mathcal{H}_X = \left\{F_\mathcal{A}(X^T) \;:\; \mathcal{A} = (A_1, \ldots, A_L),\; \|A_i\|_\sigma \leq s_i,\; \|A_i^\top - M_i^\top\|_{2,1} \leq b_i\right\}$$

We can now use the upper bound on the covering number of the set of outputs of each layer obtained in equation (1), as well as the inequality (2) in order to combine them, and we obtain the following upper bound on the covering number of the set $\mathcal{H}_X$. We have:

$$\ln \mathcal{N}(\mathcal{H}_X, \epsilon, \|\cdot\|_2) \leq \frac{\|X\|_2^2 \ln(2W^2)}{\epsilon^2} \Big(\prod_{j=1}^{L} s_j^2 \rho_j^2\Big)\Big(\sum_{i=1}^{L} \big(\frac{b_i}{s_i}\big)^{2/3}\Big)^3 \tag{3}$$

The last step of the proof is to use this bound on the covering number of the set of all possible outputs to derive an upper bound on the Rademacher complexity $\mathfrak{R}((\mathcal{F}_\gamma)_{|S}).$, which is done using the Dudley entropy upper bound on the Rademacher complexity:

$$\mathfrak{R}(\mathcal{F}_{|S}) \leq \inf_{\alpha > 0} \Big(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_{\alpha}^{\sqrt{n}} \sqrt{\log \mathcal{N}(F_{|S}, \epsilon, \cdot_2)} d\epsilon\Big) \tag{4}$$

If we set $\alpha = 1/n$, we obtain the desired bound.

3 Experimental evaluation

We first check the performance of AlexNet [3] and one hidden layer MLP (multilayer perceptron) on `cifar10` with original and random labels. Next, we explore the spectrally-normalized margin distribution [1] for AlexNet. Finally, We test three types of complexity/metric [1] [6] [4] normalized margin distribution on one-hidden layer MLP. We do not perform exactly the same experiment of the existing papers. Instead, we use different optimizers and slightly different experiment design. The optimization method in all the experiment is SGD with momentum = 0.9. We use adaptive learning rate. We first set learning rate to be 0.01 and divide it by 10 twice at 150 and 225 epochs. We believe the use of optimizers and adaptive learning rate do not change the interpretation of theoretical results. All the results in our experiment are rescaled for easy comparison.

3.1 Neural network can easily fit random labels

We study the power of neural network with original and random labels. Our experiment design follows [8]. We use dataset `cifar10`. For original labels, we keep the label in the dataset. For random labels, we set the labels in both training and test sets of `cifar10` randomly from $\{1, ..., 10\}$. We try two difference types of Network, one hidden layer MLP and AlexNet. We also explore the influence of data augmentation and weight decay. Our results are shown in Table 1. Both AlexNet and MLP can easily fit random labels. And the test accuracy is of course around 10%, the same as random guessing, since we cannot learn a model to predict random labels. Our results are consistent with [8].

3.2 Generalization case studies via spectrally-normalized margin distributions

We focus on AlexNet. We remove weight decay and normalization in the experiment of this section. Instead of calculate a certain risk of given margin, we visualize the generation performance

model	random crop	weight decay	train accuracy	test accuracy
AlexNet	yes	yes	100.0	78.21
	no	yes	100.0	77.83
	no	no	100.0	68.87
(random labels)	no	no	100.0	9.93
MLP 1×512	no	yes	100.0	53.21
	no	no	100.0	53.73
(random labels)	no	no	99.96	9.89

Table 1: Performance of two types of neural networks with different settings on original and random labels of `cifar`

our network by showing the density estimator of margin distribution and normalized margin distribution. Given n pairs of (x_i, y_i), $i = 1, ..., n$, with x_i as the i th row of the data matrix $X \in \mathbb{R}^{n \times d}$, and a given network $F_{\mathcal{A}} : \mathbb{R}^d \to \mathbb{R}^k$, the margin distribution is the empirical distribution of the following scalar for each data point (x, y):

$$F_{\mathcal{A}}(x)_y - \max_{i \neq y} F_{\mathcal{A}}(x)_i,$$

and the normalized margin distribution is the margin distribution normalized by the spectral complexity multiplied by the norm of the data:

$$\frac{F_{\mathcal{A}}(x)_y - \max_{i \neq y} F_{\mathcal{A}}(x)_i}{R_{\mathcal{A}} \|X\|_2 / n}.$$

The normalized margin distribution is derived from the bound in the main theorem. This margin distribution has a clear explanation. We can compare two margin distributions in this way: Given a fixed point on the horizontal axis, if the cumulative distribution of one density estimator is lower than the other, then it has a corresponding lower margin. The spectral complexity defined before is a correlation of Lipschitz constant. The Lipschitz constant of a given network $F_{\mathcal{A}}$ is:

$$\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma,$$

while the spectral complexity $R_{\mathcal{A}}$ is:

$$(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma) \cdot (\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}})^{3/2}$$

We show the excess risk (test error - train error), Lipschitz constant and spectral complexity across epochs in Figure 1. In plots, though Lipschitz constant has the same trend as excess risk but it does not fit the excess risk completely. And spectrally-complexity fit the excess risk better than the Lipschitz constants. This corresponds to the right hand side of the margin bound.

Then we compare the margin-distributions and spectrally-normalized margin distributions for AlexNet on `cifar10` with original and random labels in Figure 2. For normalized margin distributions, we can easily find that random labels is a harder problem.

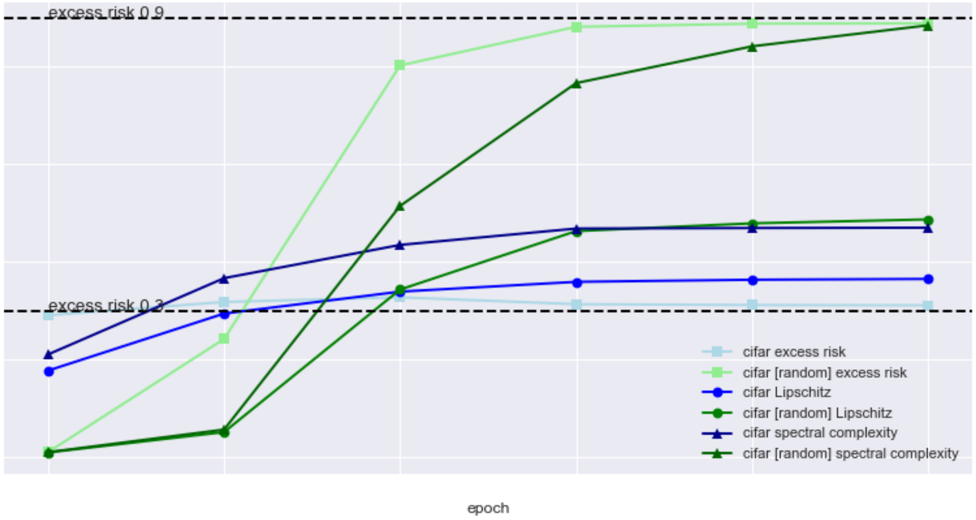

Figure 1: An analysis of AlexNet trained with SGD on `cifar10`, with both original and random labels. Blue curves show the results of original labels while Green curves show the results of random labels. Square-marked curves, circle-marked curves, and triangle marked curves track excess risk, Lipschitz constant, and spectral complexity respectively.

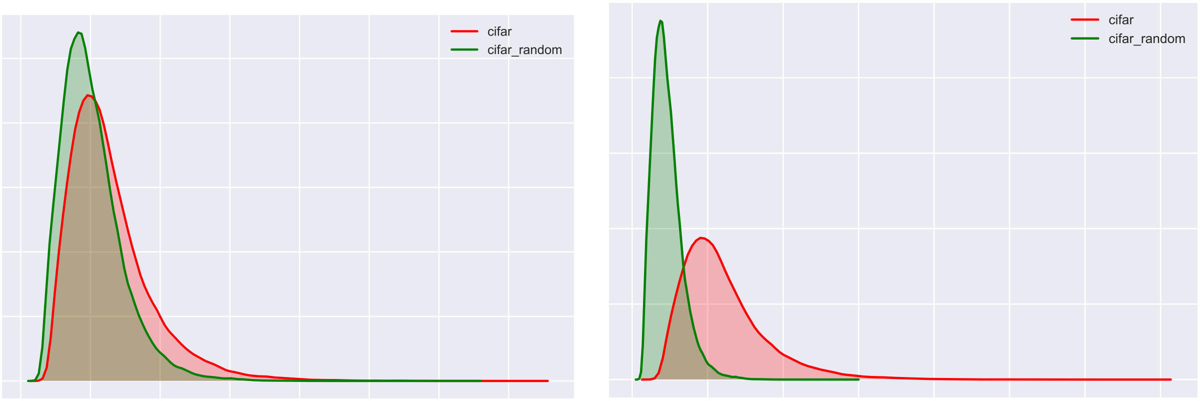

Figure 2: The left panel shows margin distributions at the end of training AlexNet on `cifar`, with and without random labels. The right panel shows normalized ones. Simple margin distributions can tell nothing. But from normalized ones, we can see random labels correspond to a harder problem.

Comparison of datasets. We compare normalized margin distributions of different datasets `cifar10` `cifar10` with random labels, and `cifar100` in Figure 3. `cifar100` is even a harder problem than `cifar10` random labels which explains the poor test accuracy for `cifar10`.

Comparison of regularization. We compare different ℓ_2 regularization levels in Figure 4. 10^{-4} regularization level corresponds to a easier problem while small regularization levels 10^{-5}

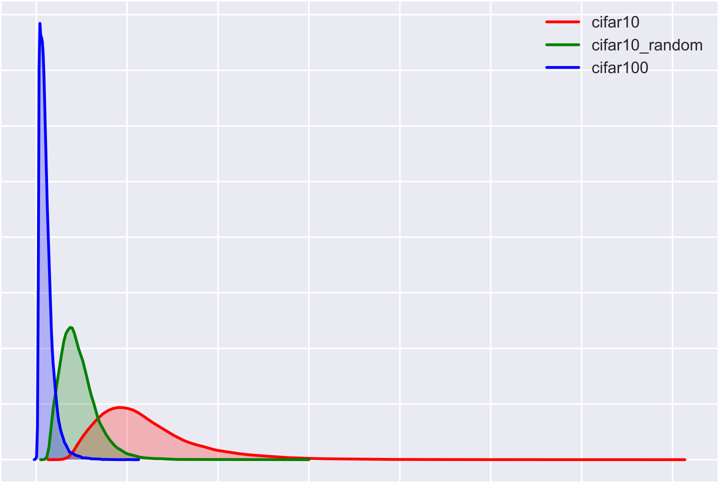

Figure 3: Normalized margin distributions at the end of training AlexNet on `cifar10`, `cifar10` random labels, and `cifar100`. `cifar100` is even a harder problem than `cifar10` random labels.

and 10^{-6} do not make a difference. This is consistent with the fact using $5 * 10^{-4}$ regularization level has a better performance in practice.

Comparison across epochs. We show the margin distributions and normalized margin distributions across epochs in Figure 5. The margin distributions does not have a certain pattern but the normalized margin distributions seem to converge at the end which is consistent with our finding in the analysis.

Comparison of different corrupted probabilities. We compare different corrupted probabilities (the percentage of corrupted labels) of `cifar10` in Figure 6. The unnormalized margins are mixed up. We cannot tell which one is harder from the unnormalized margin distributions. But for normalized margin distributions, the larger corrupted probability leads to a harder problem except corrupt = 0.2.

3.3 Three different types of complexity/metric

We restate the three different type of complexity here. We call the corresponding complexity in the bound of [6] modified spectral complexity. Only the bound with spectral complexity can handle convolutional neural network with max-pooling. Only the bounds with spectral complexity and modified spectral complexity can handle multiple classes. But we also show the margin distributions normalized by Fisher-Rao metric for `cifar10` here. It turns out the three different normalized margin distributions are similar. They all can tell random labels is a harder problem than original labels.

- spectral complexity [1]

$$R_{\mathcal{A}} := \Big(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\Big) \cdot \Big(\sum_{i=1}^{L} \frac{\|A_i^\top - M_i^\top\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}}\Big)^{3/2}$$

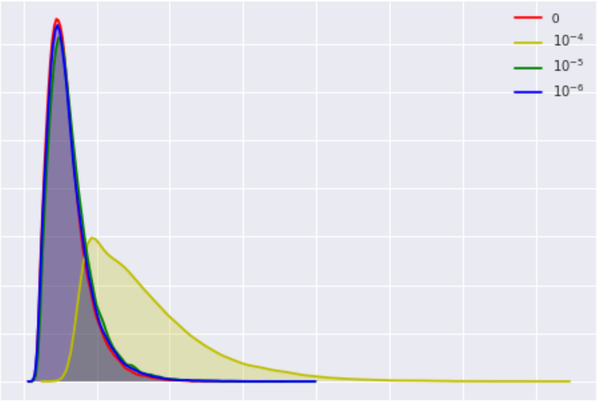

Figure 4: Normalized margin distributions at the end of training AlexNet on `cifar10`, with different regularization levels. Small regularizations do not make a difference. Slightly larger regularization $10^{-}4$ has a smaller margin bounds.

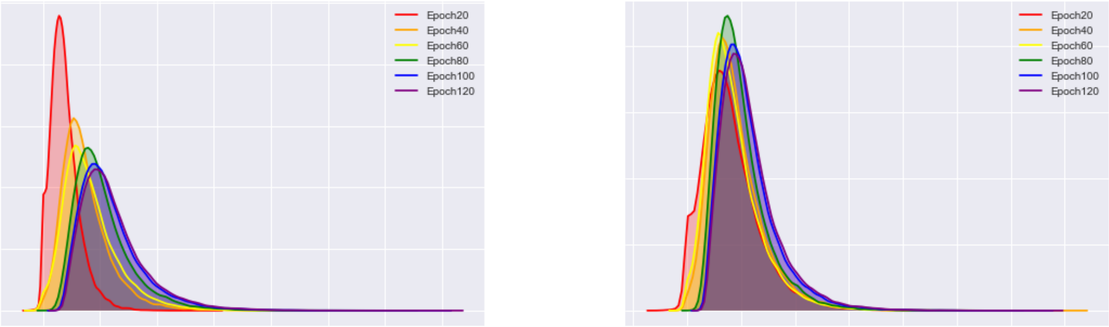

Figure 5: The left panel shows margin distributions of training AlexNet on `cifar` across epochs. The right panel shows normalized ones.

- modified spectral complexity [6]

$$R'_{\mathcal{A}} := \Big(\prod_{i=1}^{L} \rho_i \|A_i\|_\sigma\Big) \cdot L\Big(\sum_{i=1}^{L} \frac{(\sqrt{W}\|A_i^\top - M_i^\top\|_2)^2}{\|A_i\|_\sigma^2}\Big)^{1/2}$$

- Empirical Fisher-Rao metric [4]

$$\|\mathcal{A}\|_{fr,emp}^2 = (L+1)^2 \frac{1}{m} \sum_{i=1}^{m} [\langle softmax(F_{\mathcal{A}}(x_i)), F_{\mathcal{A}}(x_i)\rangle - F_{\mathcal{A}}(x_i)_{y_i}]^2$$

The first two are computed on the training data. The third one is computed from a sample of size m on the test data. We compare the margin distributions normalized by three types of complex-

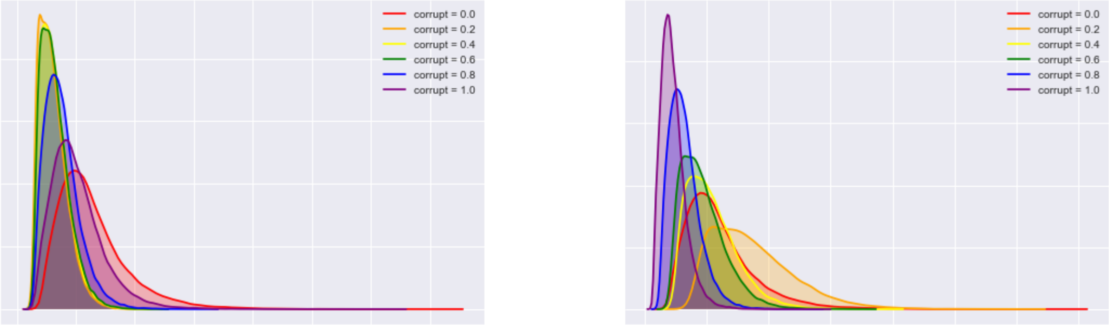

Figure 6: The left panel shows margin distributions of training AlexNet on `cifar` with different corrupted probabilities. The right panel shows normalized ones.

ity/metric at the end of training MLP on `cifar10` with both original and random labels in Figure 7. The MLP we use has only one hidden layer with 512 units.

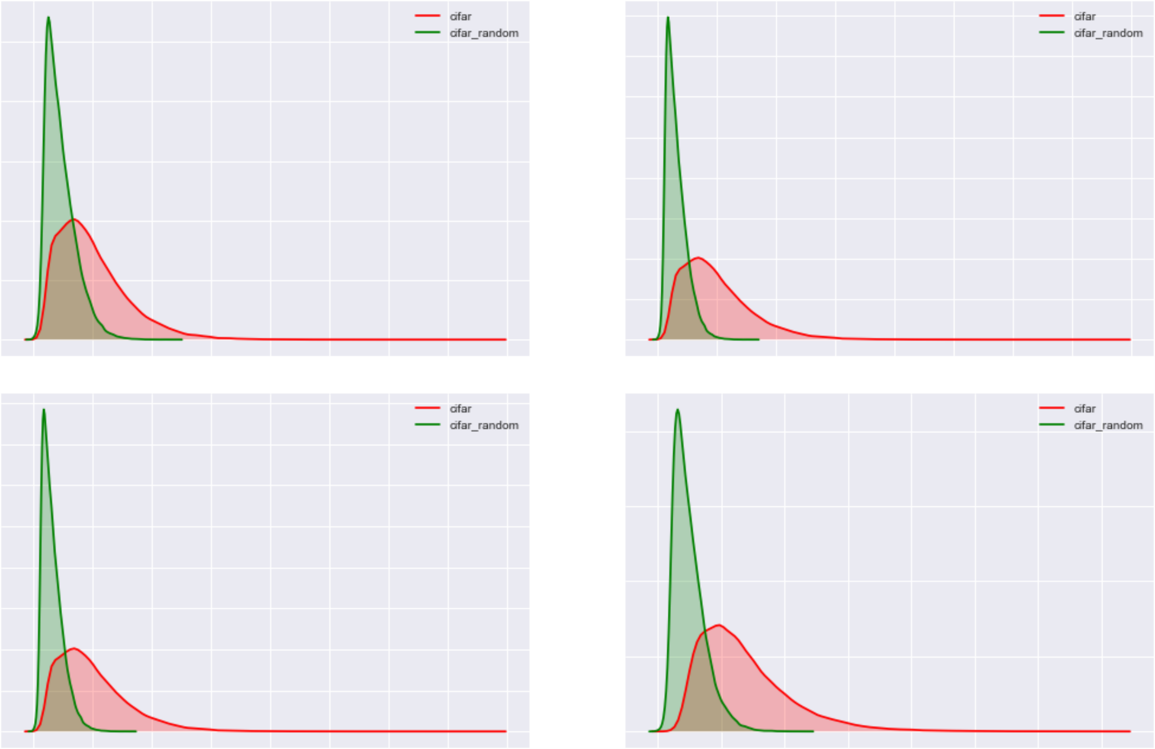

Figure 7: Margin distributions normalized by three types of complexity/metric at the end of training MLP on `cifar10` with both original and random labels. The top left figure shows the unnormalized margin distributions. The top right shows the margin distributions normalized by spectral complexity. The bottom left shows the margin normalized by modified spectral complexity. The bottom right shows the margin normalized by Fisher-Rao metric.

4 Conclusions

We summarized the theoretical results and the proof of the interesting paper [1]. We did numerical experiment to find the interpretation of the bounds in the paper for AlexNet on a commonly used dataset `cifar10`. We used margin distribution normalized by the spectral complexity in the paper to compare different datasets, different corrupted probabilities, different regularization levels and different training epochs. We also compare margin distributions normalized by spectral complexity and other two metric [6] [4] on MLP. Our results can be summarized as follows:
For AlexNet,

- Spectrally complexity is more consistent with the excess risk than the Lipschitz constant.
- Random labels `cifar10` is harder than original labels in the sense of spectrally-normalized margin distributions.
- `cifar100` is even harder than `cifar10` with random labels in the sense of spectrally-normalized margin distributions.
- Small levels of regularization makes almost no difference for spectrally-normalized margin distributions which corresponds to no difference in excess risk.
- Larger corrupted probability leads to a harder problem in the sense of spectrally-normalized margin distributions.
- Spectrally-normalized margin distributions across epochs seem to converge while the spectrally-complexity and margin distributions both do not converge.

For MLP, we test the different complexity/metric normalized distributions in [4], [6] [1]. We find that three types of complexity/metric have similar interpretation of the hardness of a problem.

References

[1] Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

[2] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[4] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

[5] Charles H Martin and Michael W Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

[6] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[7] Behnam Neyshabur, Srinadh Bhojanapalli, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5943–5952, 2017.

[8] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.