Annals of Mathematical Sciences and Applications
Volume 7, Number 2, 221–258, 2022

On Breiman's dilemma in neural networks: phase transitions of margin dynamics *

Weizhi Zhu, Yifei Huang, and Yuan Yao

Margin enlargement over training data has been an important strategy since perceptrons in machine learning for the purpose of boosting the robustness of classifiers toward a good generalization ability. Yet Breiman shows a dilemma [5] that a uniform improvement on margin distribution *does not* necessarily reduces generalization errors. In this paper, we revisit Breiman's dilemma in deep neural networks with recently proposed spectrally normalized margins. A novel perspective is provided to explain Breiman's dilemma based on phase transitions in evolution of normalized margin distributions during training, that reflects the trade-off between expressive power of models and complexity of data. When data complexity is comparable to the model expressiveness in the sense that both training and test data share similar phase transitions in normalized margin dynamics, two efficient ways are derived to predict the trend of generalization or test error via classic margin-based generalization bounds with restricted Rademacher complexities. On the other hand, over-expressive models that exhibit uniform improvements on training margins, as a distinct phase transition to test margin dynamics, may lose such a prediction power and fail to prevent the overfitting. Experiments are conducted to show the validity of the proposed method with some basic convolutional networks, AlexNet, VGG-16, and ResNet-18, on several datasets including Cifar10/100 and mini-ImageNet.

1. Introduction

Margin, as a measurement of the robustness allowing some perturbations on classifier without changing its decision on training data, has a long history in

*Part of this work was presented at ICCM 2019 in Beijing. The first two authors make indispensable contributions to this paper in equal importance but different aspects: WZ proves the theorems, conducts some experiments, and writes the paper; YH carries out major experiments; YY designs the project and writes the paper. Correspondence email: yuany@ust.hk

characterizing the performance of classification algorithms in machine learning. As early as [22], it played a central role in the proof on finite-stopping or convergence of perceptron algorithm when training data is separable. Equipped with convex optimization technique, a plethora of large margin classifiers were triggered by support vector machines [8, 27]. For neural networks, [2, 3] showed that the generalization error can be bounded by a margin-sensitive fat-shattering dimension, which is in turn bounded by the ℓ_1-norm of weights, shedding light on possible good generalization ability of over-parameterizd networks with small size weights despite the large VC dimensionality. The same idea was later applied to AdaBoost, an iterative algorithm to combine an ensemble of classifiers proposed by [9], often exhibiting a phenomenon of resistance to overfitting that during the training process the generalization error does not increase even when the training error drops to zero. Toward deciphering such a resistance to overfitting phenomenon, [23] proposed an explanation that the training process keeps on improving a notion of classification margins in boosting, among later improvement [12] and works on establishing consistency of boosting via early stopping regularization [6, 32, 30]. Lately such a resistance to overfitting was again observed in deep neural networks with over-parameterized models [31]. A renaissance of margin theory was brought by [1] with a normalization of network using Lipschitz constants bounded by products of operator spectral norms. It inspires a variety of further investigations [19, 21, 17].

However, the margin theory has a limitation that the improvement of margin distributions does not necessarily guarantee a better generalization performance, which is at least traced back to [5] in his effort to understanding AdaBoost. In this work, Breiman designed an algorithm *arc-gv* such that the margin can be maximized via a prediction game, then he demonstrated an example that one can achieve uniformly larger margin distributions on training data than AdaBoost but suffer a higher generalization error. In the end of this paper, Breiman made the following comments with a dilemma:

"The results above leave us in a quandary. The laboratory results for various arcing algorithms are excellent, but the theory is in disarray. The evidence is that if we try too hard to make the margins larger, then overfitting sets in. ⋯ My sense of it is that we just do not understand enough about what is going on."

In this paper, we are going to revisit Breiman's dilemma in the scenario of deep neural networks. Both success and failure can be witnessed by distinct phase transitions in evolution of normalized margin distributions during training. First of all, let's look at the following illustration example.

Example 1.1 (Breiman's Dilemma with a CNN)**.** *A basic 5-layer convolutional neural network of c channels (see Section 3 for details) is trained with CIFAR-10 dataset whose 10 percent labels are randomly permuted. When* $c = 50$ *with* 92,610 *parameters, Figure 1 (a) shows the training error and generalization (test) error in solid curves. From the generalization error in (a) one can see that overfitting indeed happens after about 10 epochs, despite that training error continuously drops down to zero. One can successfully predict such an overfitting phenomenon from Figure 1 (b), the evolution of normalized margin distributions defined later in this paper. In (b), while small margins are monotonically improved during training, large margins undergoes a phase transition from increase to decrease around 10 epochs such that one can predict the tendency of generalization error in (a) using large margin dynamics. Two particular sections of large margin dynamics are highlighted in (b), one at 9.8 on x-axis that measures the percentage of normalized training margins no more than 9.8 (training margin error) and the other at 0.8 on y-axis that measures the normalized margins at quantile* $q = 0.8$ *(i.e.* $1/\hat{\gamma}_{q,t}$*). Both of them meet the tendency of generalization error in (a) and find good early stopping time to avoid overfitting. However, as we increase the channel number to* $c = 400$ *with about* $5.8M$ *parameters and retrain the model, (c) shows a similar overfitting phenomenon in generalization error; on the other hand, (d) exhibits a monotonic improvement of normalized margin distributions without a phase transition during the training and thus fails to capture the overfitting. This demonstrates the Breiman's dilemma in CNN.*

In this example, normalized margin distributions of a classifier over the data, can be divided into two parts: low/small margins such as some negative margins for misclassified samples vs. high/large margins for high confident correctly classified samples. These parts of margin distributions often exhibit different dynamics during the training process. Small margins for training and test datasets are effectively reduced in training, along with reductions of training and test errors; while large margins may exhibit different dynamics, such as the increase-decrease phase transition in Figure 1 (b), reflecting the trade-off between expressive power of models and complexity of data.

A key insight behind this dilemma, is that one needs a trade-off between the expressive power of models and the complexity of the dataset to endorse margin bounds a prediction power. On one hand, when a model has a limited expressive power relative to the training dataset, in the sense that the small and large portions of training margin distributions *can not* be both monotonically improved during training, the generalization or test error may be

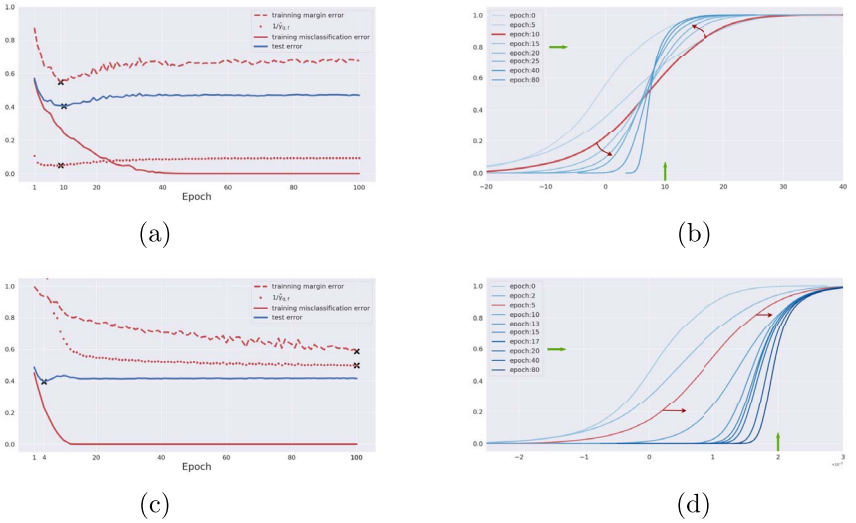


Figure 1: Demonstration of Breiman's Dilemma in Convolutional Neural Networks.

predicted from dynamics of normalized margin distributions. On the other hand, if we push too hard to improve margins by giving model too much degree of freedom such that the training margins are uniformly improved during training process, the predictability may be lost. A trade-off is thus necessary to balance the complexity of model and dataset in addition to margin improvement, otherwise one is doomed to meet Breiman's dilemma when the models arbitrarily increase the expressive power.

The example above shows that the expressive power of models relative to the complexity of dataset, can be observed from the dynamics of normalized margins in training, instead of counting the number of parameters in neural networks. In the sequel, our main contributions are to make these precise by revisiting the Rademacher complexity bounds on network generalization error.

- With the Lipschitz-normalized margins, a linear inequality is established between training margin and test margin in Theorem 1. When both training and test normalized margin distributions undergo similar phase transitions that large margins undergoes an increase-decrease during the training process, one may predict the generalization error based on the training margins as illustrated in Figure 1.
- In a dual direction, one can define a *quantile margin* via the inverse of margin distribution functions, to establish another linear inequality

between the inverse quantile margins and the test margins as shown in Theorem 2. Quantile margin is far easier to tune in practice and enjoys an empirically stronger prediction power exploiting an adaptive selection of margins along model training.

- In all cases, Breiman's dilemma may fail both of the methods above when dynamics of normalized training margins undergo different phase transitions to that of test margins during training, where a uniform improvement of training margins results in overfitting.

Section 2 describes our method to derive the two linear inequalities of generalization bounds above. Extensive experimental results are shown in Section 3 with basic CNNs, AlexNet, VGG, ResNet, and various datasets including CIFAR10, CIFAR100, and mini-Imagenet. Conclusions and future directions are discussed in Section 4. More experimental figures and proofs are collected in Appendices.

2. Methodology

Let $\mathcal{X}$ be the input space (e.g. $\mathcal{X} \subset \mathbb{R}^{C\times W\times H}$ in image classification of size #(channel)-by-#(width)-by-#(height)) and $\mathcal{Y} := \{1, \dots, K\}$ be the space of K classes. Consider a sample set of n observations $S = \{(x_1, y_1), \dots, (x_n, y_n)\colon x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ that are drawn i.i.d. from $P_{X,Y}$. For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, let $\mathbb{P}f = \int_{\mathcal{X}\times\mathcal{Y}} f(X, Y)dP_{X,Y}$ be the population expectation and $\mathbb{P}_n f = (1/n)\sum_{i=1}^n f(x_i)$ be the sample average.

Define $\mathcal{F}$ to be the space of functions $f : \mathcal{X} \to \mathbb{R}^K$ represented by neural networks,

$$(1) \qquad \begin{cases} x_0 &= x, \\ x_i &= \sigma_i(W_i x_{i-1} + b_i), \quad i = 1, \dots, l-1, \\ f(x) &= W_l x_{l-1} + b_l, \end{cases}$$

where l is the depth of the network, W_i is the weight matrix corresponding to a linear operator on x_i and σ_i stands for either element-wise activation function (e.g. ReLU) or pooling operator that are assumed to be Lipschitz bounded with constant L_{σ_i}. For example, in convolutional network, $W_i x_i + b_i = w_i * x_i + b_i$ where $*$ stands for the convolution between input tensor x_l and kernel tensor w_l. We equip $\mathcal{F}$ with the Lipschitz semi-norm, for each f,

$$(2) \qquad \|f\|_{\mathcal{F}} := \sup_{x\neq x'} \frac{\|f(x) - f(x')\|_2}{\|x - x'\|_2} \le L_\sigma \prod_{i=1}^{l} \|W_i\|_\sigma := L_f,$$

where $\|\cdot\|_\sigma$ is the spectral norm and $L_\sigma = \prod_{i=1}^{l} L_{\sigma_i}$. Without loss of generality, we assume $L_\sigma = 1$ for simplicity. Moreover we consider the following family of hypothesis mapping,

$$\mathcal{H} = \{h(x) = [f(x)]_y : \mathcal{X} \to \mathbb{R}, f \in \mathcal{F}, y \in \mathcal{Y}\}, \tag{3}$$

where $[\cdot]_j$ denotes the j^{th} coordinate and we further define the following class induced by Lipschitz semi-norm bound on $\mathcal{F}$,

$$\mathcal{H}_L = \{h(x) = [f(x)]_y : \mathcal{X} \to \mathbb{R}, h(x) = [f(x)]_y \in \mathcal{H} \text{ with } \|f\|_{\mathcal{F}} \leq L, y \in \mathcal{Y}\}. \tag{4}$$

Now, rather than merely looking at whether a prediction $f(x)$ on y is correct or not, we further consider the prediction *margin* defined as $\zeta(f(x), y) = [f(x)]_y - \max_{\{j:j\neq y\}}[f(x)]_j$. With that, we can define the *ramp loss* and *margin error* depending on the confidence of predictions. Given two thresholds $\gamma_2 > \gamma_1 \geq 0$, define the ramp loss to be

$$\ell_{(\gamma_1,\gamma_2)}(\zeta) = \begin{cases} 1 & \zeta < \gamma_1, \\ -\frac{1}{\Delta}(\zeta - \gamma_2) & \gamma_1 \leq \zeta \leq \gamma_2, \\ 0 & \zeta > \gamma_2, \end{cases}$$

where $\Delta := \gamma_2 - \gamma_1$. In particular $\gamma_1 = 0$ and $\gamma_2 = \gamma$, we also write $\ell_\gamma = \ell_{(0,\gamma)}$ for simplicity. Define the margin error to measure if f has margin no more than a threshold γ,

$$e_\gamma(f(x), y) = \begin{cases} 1 & \zeta(f(x), y) \leq \gamma \\ 0 & \zeta(f(x), y) > \gamma \end{cases}. \tag{5}$$

In particular, $e_0(f(x), y)$ is the common mis-classification error and

$$\mathbb{E}[e_0(f(x), y)] = \mathbb{P}[\zeta(f(x), y) < 0].$$

Note that $e_0 \leq \ell_\gamma \leq e_\gamma$, and ℓ_γ is Lipschitz bounded by $1/\gamma$.

The central question we try to answer is, *can we find a proper upper bound to predict the tendency of the generalization error along training, such that one can early stop the training near the epoch that* $\mathbb{P}[\zeta(f_t(x), y) < 0]$ *is minimized?* The answer is both a *yes* and a *no*!

We begin with the following lemma, as a margin-based generalization bound with network Rademacher complexity for multi-label classifications, using the uniform law of large numbers [12, 7, 15, 1]

Lemma 2.1. *Given a $\gamma_0 > 0$, then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds for any $f \in \mathcal{F}$ with $\|f\|_{\mathcal{F}} \leq L$,*

$$\mathbb{E}[\ell_{\gamma_0}(f(x),y)] \leq \frac{1}{n}\sum_{i=1}^{n}[\ell_{\gamma_0}(f(x_i),y_i)] + \frac{4K}{\gamma_0}\mathcal{R}_n(\mathcal{H}_L) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{6}$$

where

$$\mathcal{R}_n(\mathcal{H}_L) = \mathbb{E}_{x_i,\varepsilon_i}\sup_{h\in\mathcal{H}_L}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i) \tag{7}$$

is the Rademacher complexity of function class $\mathcal{H}_L$ with respect to n samples, and the expectation is taken over x_i, ε_i, $i = 1, ..., n$.

Unfortunately, direct application of such bound in neural networks with a constant γ_0 will suffer from the so-called *scaling problem*. The following proposition gives an lower bound of Rademacher complexity term.

2.1. A lower bound on the Rademacher complexity

Proposition 1. *Consider the networks with activation functions σ, where we assume σ is Lipschitz continuous and there exists x_0 such that $\sigma'(x_0) \neq 0$ and $\sigma''(x_0)$ exists. Then for any $L > 0$, there holds,*

$$\mathcal{R}_n(\mathcal{H}_L) \geq CL\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2\right] \tag{8}$$

where $C > 0$ is a constant that does not depend on S.

This proposition extends Theorem 3.4 in [1] to general activation functions and multi-class scenario, and the proof is presented in Appendix. It tells us if $L \to \infty$, upper bound (6) becomes trivial since $\mathcal{R}_n(\mathcal{H}_L) \to \infty$. In fact, both [26] and [25] show that the gradient descent method will drive weight estimates in logistic regression and general boosting with exponential loss etc. to max-margin classifier at infinity when the data is linearly separable. In particular, the latter shows the growth rate of weight estimates is $\log(t)$. As for the deep neural network with cross-entropy loss, the input of last layer is usually be viewed as features extracted from original input. Training the last layer with other layers fixed is exactly a logistic regression, and the feature is linearly separable as long as the training error achieves zero. Therefore, without any normalization, the hypothesis space during training has no upper bound on L, and thus the upper bound (6) is useless.

Besides, even for a fixed L, the complexity term $\mathcal{R}_n(\mathcal{H}_L)$ is computationally intractable.

In the following we are going to present two simple generalization error bounds based on normalized margins and restricted Rademacher complexity within certain Lipschitz balls.

2.2. Two simplified bounds with normalized margins and restricted Rademacher complexity

The first remedy is to restrict our attention on $\mathcal{H}_1$ by normalizing f with its Lipschitz semi-norm $\|f\|_{\mathcal{F}}$ or its upper bounds. Note that a normalized network $\tilde{f} = f/C$ has the same mis-classification error as f for all $C > 0$. For the choice of C, it's hard in practice to directly compute the Lipschitz semi-norm of a network, but instead some approximate estimates on the upper bound L_f in (2) are available as discussed in Section 2.3. In the sequel, let $\tilde{f} = f/L_f$ be the normalized network and $\tilde{h} = h/L_f = \zeta(f, y)/L_f = \zeta(\tilde{f}, y) \in \mathcal{H}_1$ be the corresponding normalized hypothesis function. Now a simple idea is to regard $\mathcal{R}_n(\mathcal{H}_1)$ as a constant when the model complexity is not over-expressive against data, then one can predict the tendency of generalization error via training margin error of the normalized network, that avoids the scaling problem and the computation of Rademacher complexity. The following theorem makes this precise.

Theorem 1. *Given γ_1 and γ_2 such that $\gamma_2 > \gamma_1 \geq 0$ and $\Delta := \gamma_2 - \gamma_1 \geq 0$, for any $\delta > 0$, with probability at least $1 - \delta$, along the training epoch $t = 1, \ldots, T$, the following holds for each f_t,*

$$\mathbb{P}[\zeta(\tilde{f}_t(x), y) < \gamma_1] \leq \mathbb{P}_n 1[\zeta(\tilde{f}_t(x), y) < \gamma_2] + \frac{C_{\mathcal{H}}}{\Delta} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{9}$$

where $C_{\mathcal{H}} = 4K\mathcal{R}_n(\mathcal{H}_1)$.

Remark. In particular, when we take $\gamma_1 = 0$ and $\gamma_2 = \gamma > 0$, the bound above becomes,

$$\mathbb{P}[\zeta(f_t(x), y) < 0] \leq \mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma] + \frac{C_{\mathcal{H}}}{\gamma} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{10}$$

Theorem 1 says, we can bound the normalized test margin distribution $\mathbb{P}[\zeta(\tilde{f}_t(x), y) < \gamma_1]$ by the normalized training margin distribution $\mathbb{P}_n[\zeta(\tilde{f}_t(x), y) < \gamma_2]$. Recently [17] investigates for normalized networks, the strong linear relationship between cross-entropy training loss and test loss when the training epochs are large enough. As a contrast, we consider the

whole training process and normalized margins. In particular, we hope to predict the trend of generalization (test) error by choosing $\gamma_1 = 0$ and a proper γ such that the training margin errors $\mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma]$ enjoy a high correlation with test error up to a monotone transform. For this purpose, the following facts are important. First, we do not expect the bound, for example (10), is tight for every choice of $\gamma > 0$, instead we hope there exists some γ such that the training margin error nearly monotonically changes with generalization error. Figure 5 below shows the existence of such γ when models are not too big by exhibiting rank correlations between training margin error at various γ and training/test error. Moreover, Figure 4 below shows that the training margin error at such a good γ successfully recover the tendency of generalization error on CIFAR10 dataset. Second, the normalizing factor is not necessarily to be an upper bound of Lipschitz semi-norm. The key point is to prevent the complexity term of the normalized network going to infinity. Since for any constant $c > 0$, normalization by $\bar{L} = cL$ works in practice where the constant could be absorbed to γ, we could ignore the Lipschitz constant introduced by general activation functions in the hidden layers.

However, as Example 1.1 with Figure 1 shows above, once the training margin distribution is uniformly improved, dynamic of training margin error fails to detect the minimum of generalization error in the early stage. This is because when network structure becomes complex enough, the training margin distribution could be more easily improved. In this case although both $1/\Delta$ and training margins $\mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma]$ are reduced, the restricted Rademacher complexity $\mathcal{R}_n(\mathcal{H}_1)$ in Theorem 1 will blow up such that it is invalid to bound the generalization error using merely the training margins. In this case, the generalization error may overfit while training margins can not show it. This is exactly the same observation in [5] to doubt the margin theory in boosting type algorithms. More detailed discussions will be given in Section 3.2.

The most serious limitation of Theorem 1 lies in we must fix a γ along the complete training process. In fact, the first term and second term in the bound (10) vary in the opposite directions with respect to γ, and thus different f_t may prefer different γ for a trade-off. As in Figure 1 (b) of the example, while choosing γ is to fix an x-coordinate section of margin distributions, its dual is to look for a y-section which leads to different margins for different f_t. This motivates the *quantile margin* in the following theorem. Let $\hat{\gamma}_{q,f}$ be the q^{th} *quantile margin* of the network f with respect to sample S,

$$\hat{\gamma}_{q,f} = \inf\{\gamma : \mathbb{P}_n 1[\zeta(f(x_i), y_i) \leq \gamma] \geq q\}. \tag{11}$$

Theorem 2. *Assume the input space is bounded by $M > 0$, that is $\|x\|_2 \leq M$, $\forall x \in \mathcal{X}$. Given a quantile $q \in [0,1]$, for any $\delta \in (0,1)$ and $\tau > 0$, the following holds with probability at least $1-\delta$ for all f_t satisfying $\hat{\gamma}_{q,\tilde{f}_t} > \tau$,*

$$\mathbb{P}[\zeta(f_t(x), y) < 0] \leq C_q + \frac{C_{\mathcal{H}}}{\hat{\gamma}_{q,\tilde{f}_t}} \tag{12}$$

$$C_q = q + \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{\log\log_2(4(M+l)/\tau)}{n}} \text{ and } C_{\mathcal{H}} = 8K\mathcal{R}_n(\mathcal{H}_1).$$

Remark. We simply denote $\gamma_{q,t}$ for $\gamma_{q,\tilde{f}_t}$ when there is no confusion.

Compared with the bound (10), (12) make the choice of γ varying with f_t and the cost is an additional constant term C_q^2 and the constraint $\hat{\gamma}_{q,t} > \tau$ that typically holds for large enough q in practice. In applications, the stochastic gradient descent method often effectively improves the training margin distributions along the drops of training errors, a small enough τ and large enough q usually meet $\hat{\gamma}_{q,t} > \tau$. Moreover, even with the choice $\tau = \exp(-B)$, constant term $\sqrt{[\log\log_2(4(M+l)/\tau)]/n} = O(\sqrt{\log B/n})$ is still negligible and thus very little cost is paid in the upper bound.

In practice, tuning $q \in [0,1]$ is far easier than tuning $\gamma > 0$ directly and setting a large enough $q \geq 0.9$ usually provides us lots of information about the generalization performance. The quantile margin works effectively when the dynamics of large margin distributions reflects the behavior of generalization error, e.g. Figure 1. In this case, after certain epochs of training, the large margins have to be sacrificed to further improve small margins to reduce the training loss, that typically indicates a possible saturation or overfitting in test error.

2.3. Estimate of normalization factors

In this section we discuss how to estimate the Lipschitz constant bound in (2). Given an operator W associated with a convolutional kernel w, i.e. $Wx = w * x$, there are two ways to estimate its operator norm. We begin with the following proposition, part (A) of which is adapted from the continuous version of Young's convolution inequality in L_p space (see Theorem 3.9.4 in [4]), and part (B) of which is a generalization to multiple channel kernels widely used in convolutional networks nowadays. The proof is presented in Appendix B.5.

Proposition 2. *(A) For convolution operator W with kernel $w \in \mathbb{R}^{\mathrm{Size}}$ where $\mathrm{Size} = (\mathrm{Size}_i)_{i=1}^d$ is the d-dimensional kernel size, there holds*

$$\|w * x\|_2 \le \|w\|_1 \|x\|_2. \tag{13}$$

In other words, $\|W\|_\sigma \le \|w\|_1$.

(B) Consider a multiple channel convolutional kernel $w \in \mathbb{R}^{\mathrm{C_{out}} \times \mathrm{C_{in}} \times \mathrm{Size}}$ with stride S, which maps input signal x of C_{in} channels to the output of C_{out} channels by

$$(Wx)(u, c_{out}) = [w * x](u, c_{out}) := \sum_{v, c_{in}} x(v, c_{in}) w(c_{out}, c_{in}, u - v),$$

where x and w are assumed of zero-padding outside its support. The following upper bounds hold.

1. *Let $\|w\|_{\infty,\infty,1} := \max_{i,j} \|w(j, i, \cdot)\|_1$, then*

$$\|w * x\|_2 \le \sqrt{\|w\|_1 \|w\|_{\infty,\infty,1}} \|x\|_2; \tag{14}$$

2. *Let $D := \prod_i \lceil \mathrm{Size_i}/S \rceil$ where $\lceil t \rceil := \inf_k \{k \in \mathbb{Z} : k \ge t\}$, then*

$$\|w * x\|_2 \le \sqrt{D \|w\|_1 \|w\|_\infty} \|x\|_2. \tag{15}$$

Remark. For stride $S = 1$, the upper bound (14) is tighter than (15), while for a large stride $S \ge 2$, the second bound (15) might become tighter by taking into acount the effect of stride.

In all these cases, the ℓ_1-norm of w dominates the estimates, so in the following we will simply call these bounds ℓ_1-based estimates. Another method is given in [19] based on power iterations [10], as a fast numerical approximation for the spectral norm of the operator matrix. Yet as a shortcoming, the power iteration method is not easy to apply to the ResNets.

We compare the two estimates in Figure 10. It turns out both of them can be used to predict the tendency of generalization error using normalized margins and both of them will fail when the network has large enough expressive power. Although using the ℓ_1-based estimate is very efficient, the power iteration method may be tighter and have a wider range of predictability.

In the remaining of this section, we will particularly discuss the treatment of ResNets. ResNet is usually a composition of the basic blocks shown in Figure 2 with short-cut structure. The following method is used in this paper to estimate upper bounds of spectral norm of such a basic block of ResNet.

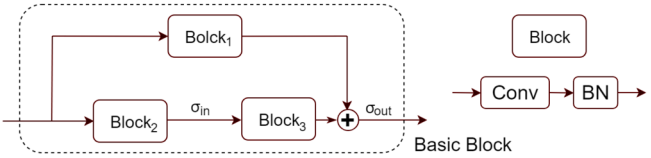


Figure 2: A basic block in ResNets used in this paper. The shortcut consists of one block with convolutional and batch-normalization layers, while the main stream has two blocks. ResNets are constructed as a cascading of several basic blocks of various sizes.

(a) Convolution layer: its operator norm can be bounded either by the ℓ_1-based estimate or by power iteration above.

(b) Batch Normalization (BN): in training process, BN normalizes samples by $x^+ = (x - \mu_B)/\sqrt{\sigma_B^2 + \epsilon}$, where μ_B, σ_B^2 are mean and variance of batch samples, while keeping an online averaging as $\hat{\mu}$ and $\hat{\sigma}^2$. Then BN rescales x^+ by estimated parameters $\hat{\alpha}, \hat{\beta}$ and output $\hat{x} = \hat{\alpha}x^+ + \hat{\beta}$. Therefore the whole rescaling of BN on the kernel tensor w of the convolution layer is $\hat{w} = w\hat{\alpha}/\sqrt{\hat{\sigma}^2 + \epsilon}$ and its corresponding rescaled operator is $\|\hat{W}\|_\sigma = \|W\|_\sigma \hat{\alpha}/\sqrt{\hat{\sigma}^2 + \epsilon}$.

(b) Activation and pooling: their Lipschitz constants L_σ can be known a priori, e.g. $L_\sigma = 1$ for ReLU and hence can be ignored. In general, L_σ can not be ignored if they are in the shortcut as discussed below.

(d) Shortcut: In residue net with basic block in Figure 2, one has to treat the mainstream $(\text{Block}_2, \text{Block}_3)$ and the shortcut Block_1 separately. Since $\|f+g\|_{\mathcal{F}} \le \|f\|_{\mathcal{F}} + \|g\|_{\mathcal{F}}$, in this paper we take the Lipschitz upper bound by $L_{\sigma_{\text{out}}}(\|\hat{W}_1\|_\sigma + L_{\sigma_{\text{in}}}\|\hat{W}_2\|_\sigma\|\hat{W}_3\|_\sigma)$, where $\|\hat{W}_i\|_\sigma$ denotes a spectral norm estimate of BN-rescaled convolutional operator W_i. In particular $L_{\sigma_{\text{out}}}$ can be ignored since all paths are normalized by the same constant, while $L_{\sigma_{\text{in}}}$ can not be ignored due to its asymmetry.

3. Experimental results

We briefly introduce the network and dataset used in the experiments. For the network, our illustration Example 1.1 is based on a simple convolutional neural network whose architecture is shown in Figure 3 (more details in Appendix Figure 11), called *basic CNN(c)* here with c channels that will be specified in different experiments below. Basically, it has five convolutional layers of c channels at each, followed by batch normalization and ReLU, as well as a fully connected layer in the end. Furthermore, we consider various

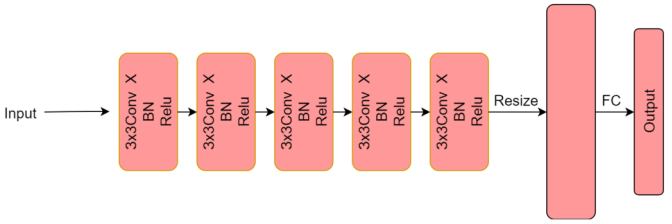


Figure 3: Illustration of the architecture of basic CNN.

popular networks in applications, including AlexNet [14], VGG-16 [24] and ResNet-18 [11]. For the dataset, we consider CIFAR10, CIFAR100 [13] and Mini-ImageNet [28].

The spirit of the following experiments is to show, *when and how, the margin bound could be used to numerically predict the tendency of generalization or test error along the training path?*

3.1. Success: training margin error and quantile margin

In this experiment, we are going to explore when there is a nearly monotone relationship between training margin error and test margin error such that Theorem 1 and Theorem 2 can be applied to predict the tendency of generalization (test) error.

First let's consider training a basic CNN(50) on CIFAR10 dataset with and without random noise. The relations between test error and *training margin error* $e_\gamma(\tilde{f}(x), y)$ with $\gamma = 9.8$, *inverse quantile margin* $1/\hat{\gamma}_{q,t}$ with $q = 0.6$ are shown in Figure 4. In this simple example where the network is small and the dataset is simple, the bounds (9) and (12) show a good prediction power: they stop either near the epoch of sufficient training without noise (Left, original data) or before an overfitting occurs with noise (Right, 10 percents label corrupted).

Why does it work in this case? Here are some detailed explanations on its mechanism. The training margin error ($\mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma]$) and the inverse quantile margin ($1/\hat{\gamma}_{q,t}$) are both closely related to the dynamics of training margin distributions. Figure 1 (b) actually shows that the dynamics of training margin distributions undergo a phase transition: while the low margins have a monotonic increase, the large margins undergo a phase transition from increase to decrease, indicated by the red arrows. Therefore different choices of γ for the linear bounds (9) (a parallel argument holds for q in (12)) will have different effects. In fact, the training margin error with

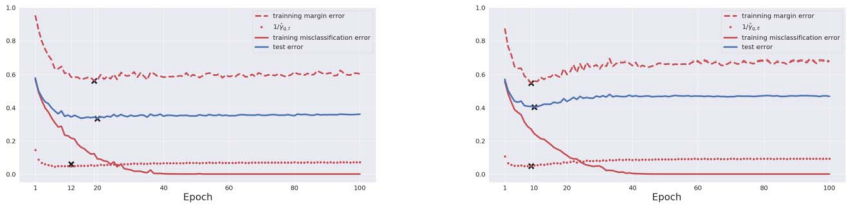


Figure 4: Success examples. Net structure: basic CNN (50). Dataset: Original CIFAR10 (Left) and CIFAR10 with 10 percents label corrupted (Right). In each figure, we show training error (red solid), training margin error $\gamma = 9.8$ (red dash) and inverse quantile margin (red dotted) with $q = 0.6$ and generalization error (blue solid). The marker "x" in each curve indicates the global minimum along epoch $1, \ldots, T$. Both training margin error and inverse quantile margin successfully predict the tendency of generalization error.

a small γ is close to the training error, while that with a large γ is close to test error. Figure 5 shows such a relation using rank correlations (in terms of Spearman-ρ and Kendall-τ[1]) between training margin errors (or inverse quantile margins) and training errors, as well as training margin errors (or inverse quantile margins) and test errors, for each γ (or q, respectively). In these plots one sees that the dynamics of large margins have a similar trend to the test errors, while small margins are close to training errors in rank correlations. Therefore for a good prediction, one can choose a large enough $\gamma = 9.8$ (or $q = 6.8$, respectively) at the peak point of rank correlation curve between training margins and test errors. Under such choices, the epoch when the phase transition above happens is featured with a *cross-over* in dynamics of training margin distributions in Figure 1 (b), and lives near the optima of the training margin error curve.

Although both the training margin error ($\mathbb{P}_n[\zeta(\tilde{f}_t(x_i), y_i) < \gamma]$) and the inverse quantile margin ($1/\hat{\gamma}_{q,t}$) can be used here to successfully predict the trend of test (generalization) error, the latter can be more powerful in our studies. In fact, dynamics of the inverse quantile margins can adaptively select γ_t for each f_t without access to the complexity term. Unlike merely looking at the training margin error with a fixed γ, quantile margin bound (12) in Theorem 2 shows a stronger prediction power than (10) and is even

[1]The Spearman's ρ and Kendall's τ rank correlation coefficients measure how two variables are correlated up to a monotone transform and a larger correlation means a closer tendency.

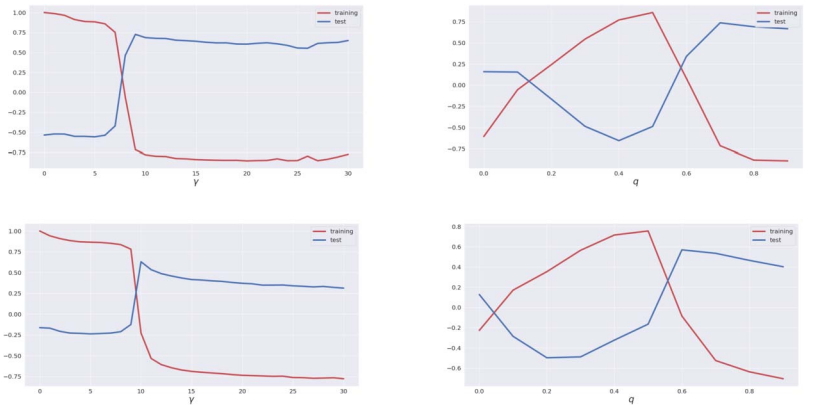


Figure 5: Spearman's ρ and Kendall's τ rank correlations between training (or quantile) margins and training errors, as well as training (or quantile) margins and test errors, at different γ (or q, respectively). Net structure: Basic CNN(50). Dataset: CIFAR10. Top: Spearman's ρ rank correlation. Bottom: Kendall's τ rank correlation. Left: Blue curves show rank correlations between training margin error and test (generalization) error, while Red curves show that between the training margin error and training error, at different γ. Right: Blue curves show rank correlations between inverse quantile margin and test error, and Red curves show that between inverse quantile margin and training error, at different q. Both Spearman's ρ and Kendall's τ show qualitatively the same phenomenon that dynamics of large margins are closely related to the test errors in the sense that they have similar trends marked by large rank correlations. On the other hand, small margins are close to training errors in trend.

able to capture more local optima. In Figure 6, the test error curve has two valleys corresponding to a local optimum and a global optimum, and the quantile margin curve with $q = 0.95$ successfully identifies both. However, if we consider the dynamics of training margin errors, it's rarely possible to recover the two valleys at the same time since their critical thresholds γ_{t_1} and γ_{t_2} are different. Another example of ResNet-18 is given in Figure 12 in Appendix.

In a summary, when training and test margin dynamics share similar phase transitions, both theorems we developed can be used to predict test (generalization) error via normalized training margins, even leaving us data-dependent early stopping rule to avoid overfitting when data is noisy. However, below we shall see a different scenario when training and test mar-

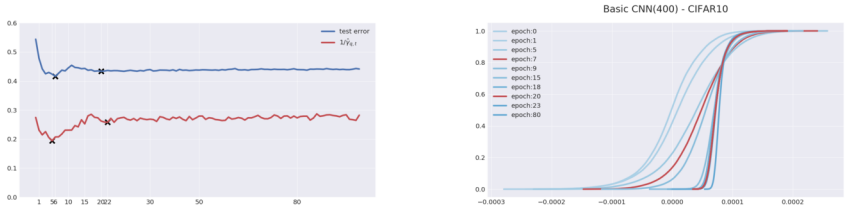


Figure 6: Inverse quantile margin. Net structure: CNN(400). Dataset: CIFAR10 with 10 percents label corrupted. Left: the dynamics of test error (blue) and inverse quantile margin with $q = 0.95$ (red). Two local minima are marked by "x" in each curve. Right: dynamics of training margin distributions, where two distributions in red color correspond to when the two local minima occur. The inverse quantile margin successfully captures two local minima of test error.

gin dynamics are of distinct phase transitions, such a prediction fails as Breiman's dilemma.

3.2. Failure: Breiman's dilemma and phase transitions in margin dynamics

In this section, we show that when the expressive power of models are comparable to data complexity, the dynamics of training margin distributions and that of test margin distributions share similar phase transitions which enables us to predict generalization (test) error utilizing the theorems in this paper. However, when model complexity arbitrarily increase to be over-expressive against the dataset, the training margins can be monotonically improved to undergo different phase transitions to that of test margin dynamics, then the prediction power is lost. This exhibits Breiman's dilemma in neural networks.

We conduct three sets of experiments in the following.

3.2.1. Experiment I: basic CNNs on CIFAR10. In the first experiment shown in Figure 7, we fix the dataset to be CIFAR10 with 10 percent of labels randomly permuted, and gradually increase the channels from basic CNN(50) to CNN(400). For CNN(50) (#(parameters) is 92,610) and CNN(100) (#(parameters) is 365,210), both training margin dyamics and test margin dynamics share a similar phase transition during training: small margins are monotonically improved while large margins are firstly improved then dropped afterwards. The last row in Figure 7 shows the heatmaps as

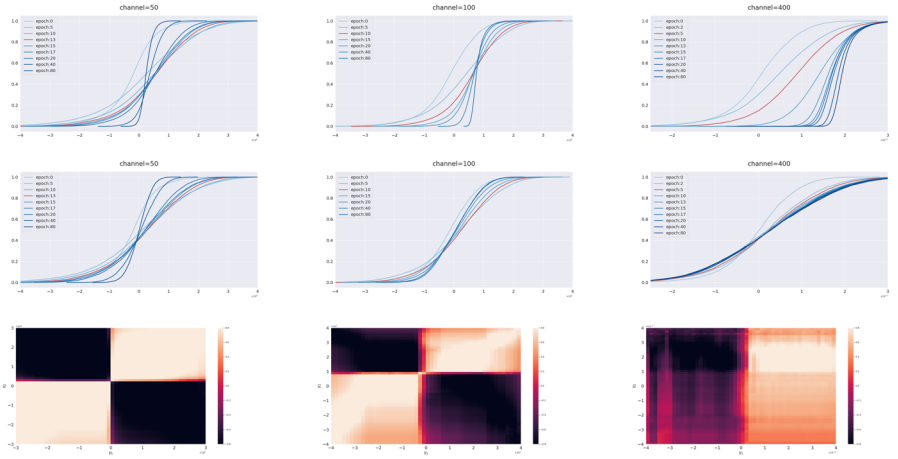


Figure 7: Breiman's Dilemma I: comparisons between dynamics of test margin distributions and training margin distributions. Net structure: Basic CNN(50) (Left), Basic CNN(100) (Middle), Basic CNN(400) (Right). Dataset: CIFAR10 with 10 percent labels corrupted. First row: evolutions of training margin distributions. Second row: evolutions of test margin distributions. Third row: heatmaps are Spearman-ρ rank correlation coefficients between dynamics of training margin error ($\mathbb{P}_n[e_{\gamma_2}(\tilde{f}(x_i), y_i)]$) and dynamics of test margin error ($\mathbb{P}[e_{\gamma_1}(\tilde{f}_t(x), y)]$) drawn on the (γ_1, γ_2) plane. CNN(50) and CNN(100) share similar phase transitions in training and test margin dynamics while CNN(400) does not. When model becomes over-representative to dataset, training margins can be monotonically improved while test margins can not be, losing the predictability.

Spearman-ρ rank correlations between these two dynamics drawn in γ_1-γ_2 plane. The block diagonal structures in the rank correlation heatmaps illustrates such a similarity in phase transitions. To be specific, small (or large) margins in both training margins and test margins share high level rank correlations marked by diagonal blocks in light color, while the difference between small and large margins are marked by off-diagonal blocks in dark color. Particularly at $\gamma_1 = 0$, the test (generalization) error dynamics can be predicted using large training margins, as their rank correlations are high.

However, as the channel number increases to CNN(400) (#(parameters) is 5,780,810), the dynamics of the training margin distributions becomes a monotone improvement without the phase transition above. This phenomenon is not a surprise since with a strong representation power, the whole training margin distribution can be monotonically improved without

sacrificing the large margins. On the other hand, the generalization or test error can not be monotonically improved. The heatmap of rank correlations between training and test margin dynamics thus exhibits such a distinction in phase transitions by changing the block diagonal structure above to double column blocks for CNN(400). In particular, for $\gamma_1 \leq 0$, test margin dynamics have low rank correlations with all training margin dynamics as they are of different phase transitions in evolutions. As a result, one CAN NOT predict test error at $\gamma = 0$ using training margin dynamics.

3.2.2. Experiment II: CNN(400) and ResNet-18 on CIFAR100 and mini-ImageNet. In the second experiment shown in Figure 8, we compare the normalized margin dynamics of training CNN(400) and ResNet-18 on two different datasets, CIFAR100 and Mini-ImageNet. CIFAR100 is more complex than CIFAR10, but less complex than Mini-ImageNet. It shows that: (a) CNN(400) does not have an over-expressive power on CIFAR100, whose normalized training margin dynamics exhibits a phase transition – a sacrifice of large margins to improve small margins during training; (b) ResNet-18 does have an over-expressive power on CIFAR100 by exhibiting a monotone improvement on training margins, but loses such a power in Mini-ImageNet with phase transitions of training margin dynamics.

From this experiment, one can see that simply counting the number of parameters and samples can not tell us if the model and data complexities are over-representative or comparable. Instead, phase transitions of margin dynamics provide us a tool to investigate their relationship. CNN(400) (5.8M parameters) has a too much expressive power for the simplest CIFAR10 dataset such that the training margins can be monotonically improved during training; but CNN(400)'s expressive power seems comparable

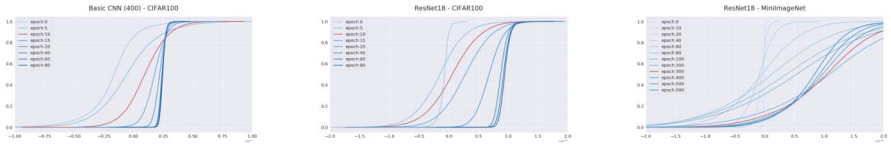


Figure 8: Breiman's Dilemma II. Net structure: Basic CNN(400) (Left), ResNet-18 (Middle, Right). Dataset: CIFAR100 (Left, Middle), Mini-ImageNet (Right) with 10 percent labels corrupted. With a fixed network structure, we further explore how the complexity of dataset influences the margin dynamics. Taking ResNet-18 as an example, margin dynamics on CIFAR100 doesn't have any cross-over (phase transition), but on Mini-Imagenet a cross-over occurs.

to the more complex CIFAR100. Similarly, the more complex model ResNet-18 (11M parameters) has a too much expressive power for CIFAR100, but seems comparable to Mini-ImageNet.

3.2.3. Comparisons of basic CNNs, AlexNet, VGG16, and ResNet-18 in CIFAR10/100 and mini-ImageNet. In this part, we collect comparisons of various networks on CIFAR10/100 and Mini-ImageNet dataset. Figure 9 shows both success and failure cases with different networks and datasets. In particular, the predictability of generalization error based on Theorem 1 and Theorem 2 can be rapidly observed on the third column of Figure 9, the heatmaps of rank correlations between training margin dynamics and test margin dynamics. On one hand, one can use the training margins to predict the test error as shown in the first column of Figure 9, when model complexity is comparable to data complexity such that the training margin dynamics share similar phase transitions with test margin dynamics, indicated by block diagonal structures in rank correlations (e.g. CNN(100) – CIFAR10, AlexNet – CIFAR100, AlexNet – MiniImageNet, VGG16 – MiniImageNet, and ResNet-18 – MiniImageNet). On the other hand, such a prediction fails when models become over-expressive against datasets such that the training margin dynamics undergo different phase transitions to test margin dynamics, indicated by the lost of block diagonal structures in rank correlations (e.g. CNN(400)- CIFAR10, ResNet-18 – CIFAR100, VGG16 – CIFAR100).

As we have shown, phase transitions of margin dynamics play a central role in characterizing the trade-off between model expressive power and data complexity, hence the predictability of generalization error by our theorems. If one tries hard to improve training margins by arbitrarily increasing the model complexity, the training margin distributions can be monotonically enlarged but it may lead to overfitting. This phenomenon is not unfamiliar to us, since Breiman has pointed out that the improvement of training margins is not enough to guarantee a small generalization or test error in the boosting type algorithms [5]. Now again we find the same phenomenon ubiquitous in deep neural networks. In this paper, the inspection of the trade-off between expressive power of models and complexity of data via phase transitions of margin dynamics provides us a new perspective to study the Breiman's dilemma in applications.

3.3. Discussion: effluence of normalization factor estimates

In the end, it's worth to mention that different choices of the normalization factor estimation may affect the range of predictability, but still exhibit

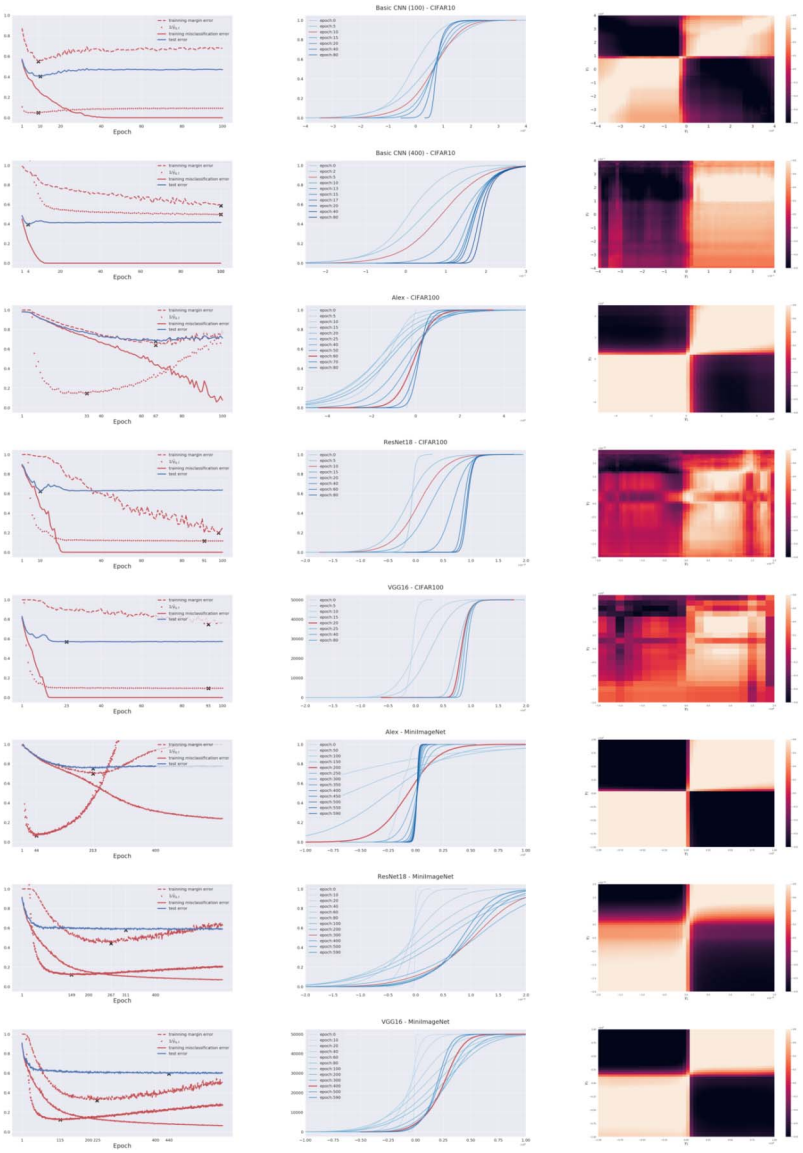


Figure 9: Comparisons of Basic CNNs, AlexNet, VGG16, and ResNet-18 in CIFAR10/100, and Mini-ImageNet. The dataset and network in use are marked in titles of middle pictures in each row. Left: curves of training error, generalization error, training margin error and inverse quantile margin. Middle: dynamics of training margin distributions. Right: heatmaps are Spearman-ρ rank correlation coefficients between dynamics of training margin error ($\mathbb{P}_n[e_{\gamma_2}(\tilde{f}(x_i), y_i)]$) and dynamics of test margin error ($\mathbb{P}[e_{\gamma_1}(\tilde{f}_t(x), y)]$) drawn on the (γ_1, γ_2) plane.

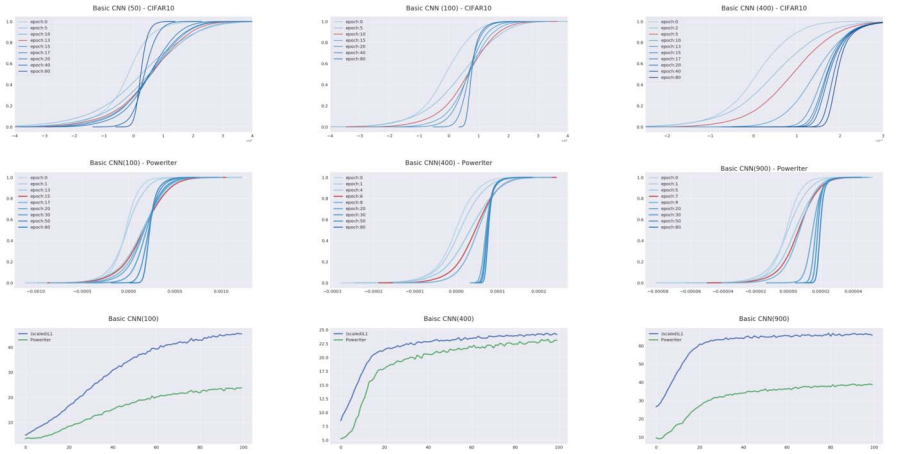


Figure 10: Comparisons on normalization factor estimates by power iteration and the ℓ_1-based estimate. Dataset: CIFAR10 with 10 percents corrupted. Net structure: Basic CNN with channels 50 (Top, Left), 100 (Top, Middle), 400 (Top Right), 200 (Middle, Left), 600 (Middle, Middle), 900 (Middle, Right). In the top row, the spectral norm in L_f is estimated via the ℓ_1-based estimate method and in the middle row, the spectral norm is estimated by power iteration. Bottom pictures show the estimates of L_f by power iterations (in green color) and by the ℓ_1-based estimate method (in blue color), respectively. The curves of L_f estimates are rescaled for visualization since a fixed scaling factor along training doesn't influence the occurrence of cross-overs or phase transitions. Note that the original ℓ_1-based estimates are of order $1e+17, 1e+19, 1e+21$ (100 channels, 400 channels, 900 channels, respectively) and the power iteration estimates are of $1e+3, 1e+3, 1e+3$ (100 channels, 400 channels, 900 channels, respectively). As shown above, a more accurate estimation of spectral norm may extend the range of predictability, but eventually faces the Breiman's dilemma if the model representation power grows too much against the dataset complexity.

Breiman's dilemma. In all experiments above, normalization factor is estimated via the ℓ_1-based estimate in Proposition 2 in Section 2.3. One could also use power iteration [19] to present a more precise estimation on spectral norm. Usually the ℓ_1-based estimates lead to a coarser upper bound than the power iterations, see Figure 10. It is a fact that in training margin dynamics, large margins are typically improved at a slower speed than small margins. Therefore it turns out a more accurate estimation of spectral norm with faster increases in training may bring cross-overs (or phase

transitions) in large training margins and extend the range of predictability. However Breiman's dilemma still persists when the balance between model representation power and dataset complexity is broken as model complexity arbitrarily grows.

4. Conclusion and future directions

In this paper, we show that Breiman's dilemma is ubiquitous in deep learning, in addition to previous studies on Boosting algorithms. We exhibit that Breiman's dilemma is closely related to the trade-off between the expressive power of models and the complexity of data. Large margins on training data do not guarantee a good control on model complexity. Instead, we have shown that phase transitions in dynamics of normalized margin distributions are able to reflect the trade-off between model expressiveness and data complexity. In particular, if high or large training margin distributions undergo increase-decrease phase transitions during training, similar to that of test margins, model expressiveness is comparable to data complexity and generalization bounds based on normalized training margins have the prediction power in capturing possible overfitting. As a result, we have shown two theorems derived from normalized Rademacher complexity bounds that can be used to quantitatively capture a data-driven early stopping rule to prevent overfitting. However, if the training margin distributions, in both high and low parts, undergo a uniform increase during training, model has over expressiveness with respect to the data and margin theory above fails. Such phase transitions in evolution of normalized margin distributions may reflect the degree-of-freedom of models with respect to data, which measures the sensitivity of model prediction over data response. Roughly speaking, a increase-decrease phase transition in high margin distributions together with a decrease in low margin distributions, indicates the degree-of-freedom of models is relatively smaller than the data complexity where one has to sacrifice the high margin predictions to improve the low margin predictions. In contrast, a uniform increase of margins over all sample suggests that the degree-of-freedom of models are larger than the data complexity. A detailed study is left for future on designing data-driven early stopping rule and degree-of-freedom for models by monitoring the margin dynamics.

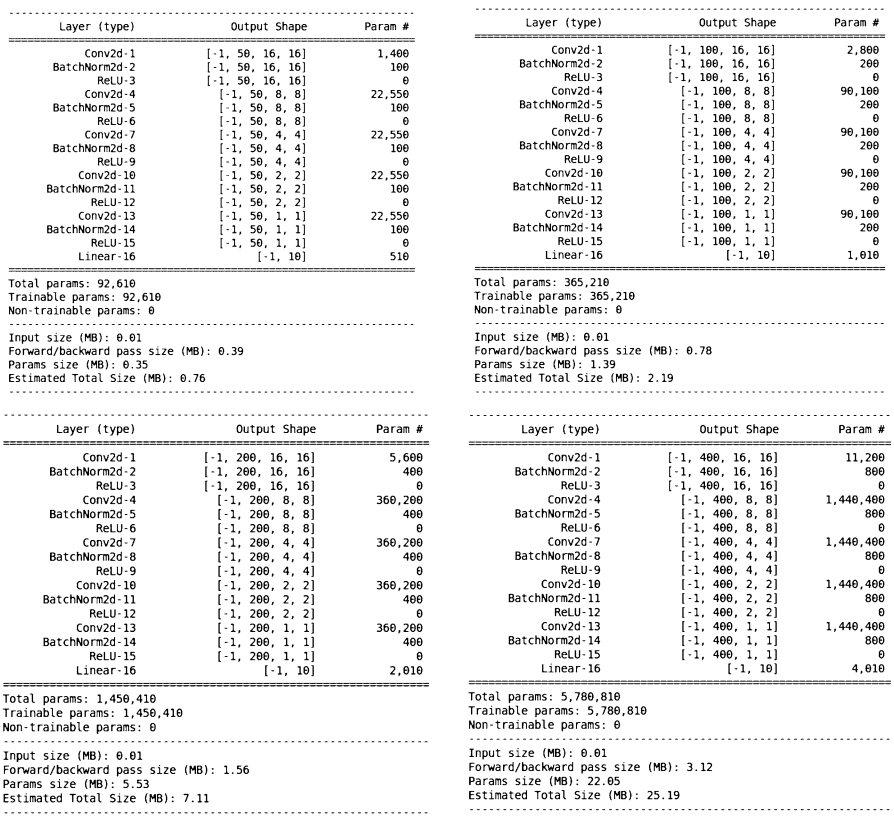
Acknowledgement

We thank Tommy Poggio, Peter Bartlett, and Xiuyuan Cheng for helpful discussions. This work was supported in part by National Natural Science

Foundation of China / Research Grants Council Joint Research Scheme Grant HKUST635/20, Hong Kong Research Grant Council (HKRGC) Grant 16308321, ITF UIM/390, as well as awards from Smale Institute of Mathematics of Computation, Tencent AI Lab, Si Family Foundation, and Microsoft Research-Asia. This research made use of the computing resources of the X-GPU cluster supported by the HKRGC Collaborative Research Fund C6021-19EF.

Appendix A. Appendix: more experimental figures

A.1. Architecture details about basic CNNs



Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 50, 16, 16]	1,400
BatchNorm2d-2	[-1, 50, 16, 16]	100
ReLU-3	[-1, 50, 16, 16]	0
Conv2d-4	[-1, 50, 8, 8]	22,550
BatchNorm2d-5	[-1, 50, 8, 8]	100
ReLU-6	[-1, 50, 8, 8]	0
Conv2d-7	[-1, 50, 4, 4]	22,550
BatchNorm2d-8	[-1, 50, 4, 4]	100
ReLU-9	[-1, 50, 4, 4]	0
Conv2d-10	[-1, 50, 2, 2]	22,550
BatchNorm2d-11	[-1, 50, 2, 2]	100
ReLU-12	[-1, 50, 2, 2]	0
Conv2d-13	[-1, 50, 1, 1]	22,550
BatchNorm2d-14	[-1, 50, 1, 1]	100
ReLU-15	[-1, 50, 1, 1]	0
Linear-16	[-1, 10]	510

Total params: 92,610
Trainable params: 92,610
Non-trainable params: 0

Input size (MB): 0.01
Forward/backward pass size (MB): 0.39
Params size (MB): 0.35
Estimated Total Size (MB): 0.76

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 100, 16, 16]	2,800
BatchNorm2d-2	[-1, 100, 16, 16]	200
ReLU-3	[-1, 100, 16, 16]	0
Conv2d-4	[-1, 100, 8, 8]	90,100
BatchNorm2d-5	[-1, 100, 8, 8]	200
ReLU-6	[-1, 100, 8, 8]	0
Conv2d-7	[-1, 100, 4, 4]	90,100
BatchNorm2d-8	[-1, 100, 4, 4]	200
ReLU-9	[-1, 100, 4, 4]	0
Conv2d-10	[-1, 100, 2, 2]	90,100
BatchNorm2d-11	[-1, 100, 2, 2]	200
ReLU-12	[-1, 100, 2, 2]	0
Conv2d-13	[-1, 100, 1, 1]	90,100
BatchNorm2d-14	[-1, 100, 1, 1]	200
ReLU-15	[-1, 100, 1, 1]	0
Linear-16	[-1, 10]	1,010

Total params: 365,210
Trainable params: 365,210
Non-trainable params: 0

Input size (MB): 0.01
Forward/backward pass size (MB): 0.78
Params size (MB): 1.39
Estimated Total Size (MB): 2.19

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 200, 16, 16]	5,600
BatchNorm2d-2	[-1, 200, 16, 16]	400
ReLU-3	[-1, 200, 16, 16]	0
Conv2d-4	[-1, 200, 8, 8]	360,200
BatchNorm2d-5	[-1, 200, 8, 8]	400
ReLU-6	[-1, 200, 8, 8]	0
Conv2d-7	[-1, 200, 4, 4]	360,200
BatchNorm2d-8	[-1, 200, 4, 4]	400
ReLU-9	[-1, 200, 4, 4]	0
Conv2d-10	[-1, 200, 2, 2]	360,200
BatchNorm2d-11	[-1, 200, 2, 2]	400
ReLU-12	[-1, 200, 2, 2]	0
Conv2d-13	[-1, 200, 1, 1]	360,200
BatchNorm2d-14	[-1, 200, 1, 1]	400
ReLU-15	[-1, 200, 1, 1]	0
Linear-16	[-1, 10]	2,010

Total params: 1,450,410
Trainable params: 1,450,410
Non-trainable params: 0

Input size (MB): 0.01
Forward/backward pass size (MB): 1.56
Params size (MB): 5.53
Estimated Total Size (MB): 7.11

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 400, 16, 16]	11,200
BatchNorm2d-2	[-1, 400, 16, 16]	800
ReLU-3	[-1, 400, 16, 16]	0
Conv2d-4	[-1, 400, 8, 8]	1,440,400
BatchNorm2d-5	[-1, 400, 8, 8]	800
ReLU-6	[-1, 400, 8, 8]	0
Conv2d-7	[-1, 400, 4, 4]	1,440,400
BatchNorm2d-8	[-1, 400, 4, 4]	800
ReLU-9	[-1, 400, 4, 4]	0
Conv2d-10	[-1, 400, 2, 2]	1,440,400
BatchNorm2d-11	[-1, 400, 2, 2]	800
ReLU-12	[-1, 400, 2, 2]	0
Conv2d-13	[-1, 400, 1, 1]	1,440,400
BatchNorm2d-14	[-1, 400, 1, 1]	800
ReLU-15	[-1, 400, 1, 1]	0
Linear-16	[-1, 10]	4,010

Total params: 5,780,810
Trainable params: 5,780,810
Non-trainable params: 0

Input size (MB): 0.01
Forward/backward pass size (MB): 3.12
Params size (MB): 22.05
Estimated Total Size (MB): 25.19

Figure 11: Detailed information about CNN(50), CNN(100), CNN(200), and CNN(400).

A.2. Two local minimums in ResNet-18

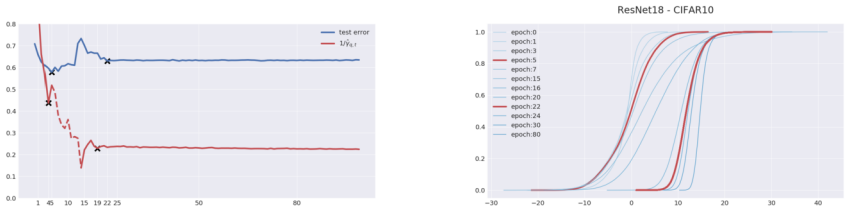


Figure 12: Inverse quantile margin captures local optima, though may fail in predicting their relative order when model complexity is over-representative. Network: ResNet-18. Data: CIFAR10 with 10 percents label corrupted. Normalization factor, spectral complexity estimated by power iteration. Left: the dynamics of test error and inverse quantile margin with $q = 0.95$. Overfitting occurs and two local minimums are marked with "x" in each dynamic. The dash line highlights the epochs when the training margins are monotonically improved. Right: dynamics of training margin distribution. Two distributions corresponding to local minima of test error are highlighted in red color. Since after the first (better) local minimum, the training margin distribution is uniformly improved in the second (worse) local minimum, that leads to the inverse quantile margin showing the second local minimum of smaller value. Yet the true order of the two local minima of test error is opposite. However, the inverse quantile margin still captures the optima locally, where the training margin distributions have cross-overs (phase-transitions) near local minima of test error.

Appendix B. Appendix: proofs

B.1. Auxiliary lemmas

Lemma B.1. *For any $\delta \in (0,1)$ and bounded-value functions $\mathcal{F}_B := \{f : \mathcal{X} \to \mathbb{R} : \|f\|_\infty \leq B\}$, the following holds with probability at least $1-\delta$,*

$$\sup_{f \in \mathcal{F}_B} \mathbb{E}_n f(x) - \mathbb{E} f(x) \leq 2\mathcal{R}_n(\mathcal{F}_B) + B\sqrt{\frac{\log(1/\delta)}{2n}} \tag{16}$$

where

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \tag{17}$$

is the Rademacher Complexity of function class $\mathcal{F}$.

For completeness, we include its proof that also needs the following well-known McDiarmid's inequality (see, e.g. [29]).

Lemma B.2 (McDiarmid's Bounded Difference Inequality)**.** *For B_i-bounded difference functions $h : \mathcal{X} \to \mathbb{R}$ s.t. $|h(x_i, x_{-i}) - h(x_i', x_{-i})| \le B_i$,*

$$\mathbb{P}\left\{\mathbb{E}_n h - \mathbb{E}_x h(x) \ge \varepsilon\right\} \le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n B_i^2}\right),$$

Proof of Lemma B.1. It suffices to show that for $\bar{f} = f(x) - \mathbb{E}f(x)$,

$$\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} = \sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} - \mathbb{E}\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} + \mathbb{E}\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} \tag{18}$$

where with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} - \mathbb{E}\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} \le B\sqrt{\frac{\log 1/\delta}{2n}} \tag{19}$$

by McDiarmid's bounded difference inequality, and

$$\mathbb{E}\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} \le 2\mathcal{R}_n(\mathcal{F}) \tag{20}$$

using Rademacher complexity.

To see (19), we are going to show that $\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f}$ is a bounded difference function. Consider $g(x_1^n) = \mathbb{E}_n \bar{f} = \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}_x f(x)$. Assume that the i-th argument x_i changes to x_i', then for every g,

$$\begin{aligned} g(x_i, x_{-i}) - \sup_g g(x_i', x_{-i}) &\le g(x_i, x_{-i}) - g(x_i', x_{-i}) \\ &\le \frac{1}{n}[f(x_i) - f(x_i')] \\ &\le \frac{B}{n}. \end{aligned}$$

Hence $\sup_g g(x_i, x_{-i}) - \sup_g g(x_i', x_{-i}) \le B/n$, which implies that $\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f}$ is a B/n- bounded difference function. Then (19) follows from the McDiarmid's inequality (Lemma B.2) using $B_i = B/n$ and $\delta = \exp(-2n\varepsilon^2/B^2)$.

As to (20),

$$\mathbb{E}\sup_{f\in\mathcal{F}_B} \mathbb{E}_n \bar{f} = \mathbb{E}_{x_1^n} \sup_{f\in\mathcal{F}_B} \mathbb{E}_{y_1^n} [\mathbb{E}_n f(x_1^n) - \mathbb{E}_n f(y_1^n)]$$

$$
\begin{aligned}
&\leq \mathbb{E}_{x_1^n, y_1^n} \sup_{f \in \mathcal{F}_B} \left[\mathbb{E}_n f(x_1^n) - \mathbb{E}_n f(y_1^n)\right] \\
&= \mathbb{E}_{x_1^n, y_1^n} \sup_{f \in \mathcal{F}_B} \mathbb{E}_{\varepsilon_1^n} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(x_i) - f(y_i)\right), \quad \varepsilon_i \in \{\pm 1\} \sim \mathcal{B}(n, 1/2) \\
&\leq \mathbb{E}_{x_1^n, y_1^n, \varepsilon_1^n} \sup_{f \in \mathcal{F}_B} \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i f(x_i) - \varepsilon_i f(y_i)) \\
&\leq 2\mathbb{E}_{x_1^n, \varepsilon_1^n} \sup_{f \in \mathcal{F}_B} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) = 2\mathcal{R}(\mathcal{F}_B)
\end{aligned}
$$

that ends the proof. □

We also need the following contraction inequality of Rademacher Complexity [16, 18].

Lemma B.3 (Rademacher Contraction Inequality)**.** *For any Lipschitz function:* $\phi : \mathbb{R} \to \mathbb{R}$ *such that* $|\phi(x) - \phi(y)| \leq L|x - y|$,

$$\mathcal{R}(\phi \circ \mathcal{F}) \leq L\mathcal{R}(\mathcal{F}).$$

[16] has an additional factor 2 in the contraction inequality which is dropped in [18]. Its current form is stated in [20] as Talagrand's Lemma (Lemma 4.2).

The last lemma gives the Rademacher complexity of the hypothesis space of maximum over functions in different hypothesis spaces [16].

Lemma B.4. *Let* $\mathcal{F}_1, \ldots, \mathcal{F}_m$ *be* m *hypothesis space and define*

$$\mathcal{M} = \{\max\{f_1(x), \ldots, f_m(x)\} : \mathcal{X} \to \mathbb{R}, f_i \in \mathcal{F}_i, i = 1, \ldots, m\}.$$

Then,

$$\mathcal{R}_n(\mathcal{M}) \leq \sum_{i=1}^{m} \mathcal{R}_n(\mathcal{F}_i).$$

B.2. Proof of Proposition 1

Proof of Proposition 1. The key idea is to approximate the linear function restricted in the Lipschitz ball by the neural network, where the local linearity of activation functions plays an important role. Therefore, we can show a subset of $\mathcal{H}_L$ whose Rademacher complexity is larger than that of the (restricted) linear function.

We consider the Taylor expansion of $\sigma(x)$ around x_0, $\sigma(x) = \sigma(x_0) + \sigma'(x_0)(x - x_0) + o(x - x_0)$, and thus,

$$\sup_{x\in[x_0-\delta,x_0+\delta]} \frac{|\sigma(x) - (\sigma(x_0) + \sigma'(x_0)(x - x_0))|}{\delta} \to 0 \text{ as } \delta \to 0^+, \tag{21}$$

and there exists a $\delta_0 > 0$, $\forall\ 0 < \delta \le \delta_0$,

$$\frac{1}{\sigma'(x_0)}(\sigma(x) - \sigma(x_0)) + x_0 \in [x_0 - \delta, x_0 + \delta] \text{ if } x \in [x_0 - \delta/2, x_0 + \delta/2]. \tag{22}$$

Without loss of generality, we assume $x_0 = 0, \sigma(0) = 0$ and $\sigma'(0) = 1$ since we can always do a linear transformation before and after each activation function and the additional Lipschitz can be bounded by a constant. We further assume the Lipschitz constant $L_\sigma = 1$ for simplicity.

Let $\mathcal{T}(r) := \{\langle w_0, x\rangle : \|w_0\|_2 \le r\}$ be the class of linear function with Lipschitz semi-norm less than r and we show that given a $M > 0$, for each $t \in \mathcal{T}(L)$, there exists $f \in \mathcal{F}$ with $\|f\|_{\mathcal{F}} \le L$ and $y_0 \in \{1, \dots, K\}$ such that $h(x) = [f(x)]_{y_0}$ satisfying $|h(x) - \langle w_0, x\rangle| \to 0,\ \forall\ \|x\|_2 \le M$.

To see this, define $t(x) = \langle w_0, x\rangle$ with $\|w_0\|_2 \le L$, which satisfies $t \in \mathcal{T}(L)$. Next we construct a particular l-layer network $f_{(w_0,\delta,M)} : x \to x_l$ as follows

$$\begin{aligned}
x^0 &= x,\\
x^{0.5} &= W_1 x^0 + b_1 = ((\langle w_0, x^0\rangle \frac{\delta}{ML}), 0, \dots, 0),\\
x^1 &= \sigma(x^{0.5}) = (\sigma(\langle w_0, x^0\rangle \frac{\delta}{ML}), 0, \dots, 0),\\
x^{i-0.5} &= W_i x^{i-1} + b_i = ([x^{i-1}]_1, 0, \dots, 0),\\
x^i &= \sigma(x^{i-0.5}) = (\sigma[x^{i-1}]_1, 0, \dots, 0) \quad i = 2, \dots, l-1,\\
f_{w_0,\delta,M}(x) &= W_l x^{l-1} + b_l = (\frac{MLx^{l-1}}{\delta}, 0, \dots, 0).
\end{aligned}$$

With such a construction $f_{w_0,\delta,M}(x)$, define $h(x) = [f_{w_0,\delta,M}(x)]_1$. Then $h \in \mathcal{H}_L$ since $\|f\|_{\mathcal{F}} \le \Pi_{i=1}^{l}\|W_i\|_\sigma \le \|w_0\|\frac{\delta}{ML}\frac{ML}{\delta} \le L$, and

$$\begin{aligned}
|\langle w_0, x\rangle - [f_{w_0,\delta,M}(x)]_1| &\le \frac{ML}{\delta}|\sigma^{l-1}(\tilde{x}) - \tilde{x}|,\\
&\le ML\sum_{i=1}^{l-1}\frac{|\sigma^i(\tilde{x}) - \sigma^{i-1}(\tilde{x})|}{\delta} \xrightarrow{\delta\to 0^+} 0,
\end{aligned} \tag{23}$$

where $\tilde{x} = \langle w_0, x^0 \rangle \frac{\delta}{ML}$ and σ^k stands for the composite of k σ functions. The second inequality is implied from (21) and (22) since $\tilde{x} \in [-\delta, \delta]$. Moreover, given $M > 0$ and $\delta > 0$, we define a subclass $\mathcal{H}_L^{\delta,M} \subset \mathcal{H}_L$ by,

$$\mathcal{H}_L^{\delta,M} = \{h(x) : h(x) = [f_{w,\delta,M}(x)]_1 \text{ with } \|w\|_2 \leq L\}$$

We firstly consider the empirical Rademacher complexity for a given sample set S of size n. Let $M_S = \sup_{x \in S} \|x\|_2$ and for any given $\delta > 0$,

$$\begin{aligned}
\mathcal{R}_S(\mathcal{H}_L) &\geq \mathcal{R}_S(\mathcal{H}_L^{\delta,M_S}), \\
&= \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_L^{\delta,M_S}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(x_i), \\
&= \mathbb{E}_\epsilon \sup_{\|w\|_2 \leq L} \frac{1}{n} \sum_{i=1}^n \epsilon_i [f_{w,\delta,M_S}(x_i)]_1, \\
&= \mathbb{E}_\epsilon \sup_{\|w\|_2 \leq L} \frac{1}{n} \sum_{i=1}^n \epsilon_i \left(\langle w, x_i \rangle - (\langle w, x_i \rangle - [f_{w,\delta,M_S}(x_i)]_1) \right), \\
&\geq \mathbb{E}_\epsilon \left[\sup_{\|w\|_2 \leq L} \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle w, x_i \rangle \right] + \dots \\
&\qquad - \mathbb{E}_\epsilon \left[\sup_{\|w\|_2 \leq L} \frac{1}{n} \sum_{i=1}^n \epsilon_i (\langle w, x_i \rangle - [f_{w,\delta,M_S}(x_i)]_1) \right], \\
&\geq \mathbb{E}_\epsilon \left[\sup_{\|w\|_2 \leq L} \frac{1}{n} \sum_{i=1}^n \epsilon_i \langle w, x_i \rangle \right] - \sup_i \sup_{\|w\|_2 \leq L} |\langle w, x_i \rangle - [f_{w,\delta,M_S}(x_i)]_1|, \\
&= L \mathbb{E}_\epsilon \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i \right\|_2 - \sup_i \sup_{\|w\|_2 \leq L} |\langle w, x_i \rangle - [f_{w,\delta,M_S}(x_i)]_1|, \qquad (24) \\
&\geq CL \sqrt{\frac{1}{n} \sum_{i=1}^n \|x_i\|_2} - \sup_i \sup_{\|w\|_2 \leq L} |\langle w, x_i \rangle - [f_{w,\delta,M_S}(x_i)]_1|, \qquad (25)
\end{aligned}$$

where (24) is implied from the Cauchy-Schwarz inequality and (25) is due to the Khintchine inequality.
From (23), we can choose proper $\delta_{M_S} > 0$ such that,

$$\sup_i \sup_{\|w\|_2 \leq L} |\langle w, x_i \rangle - [f_{w,\delta_{M_S},M_S}(x_i)]_1| \leq \frac{CL}{2} \mathbb{E}_S \sqrt{\frac{1}{n} \sum_{i=1}^n \|x_i\|_2},$$

and the right hand side is independent with S. Then by taking expectation over S in upper bound (25),

$$\mathcal{R}_n(\mathcal{H}_L) \geq CL\mathbb{E}_S\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2},$$

where we absorb a factor $1/2$ into constant C without changing the notation. □

B.3. Proof of Theorem 1

Proof of Theorem 1. Given $\theta > 0$, we firstly introduce a useful lower bound of $\zeta(f(x), y)$,

$$\begin{aligned}
\zeta^{\theta}(f(x), y) :=& [f(x)]_y - \max_{y'}([f(x)]_{y'} - \theta 1[y = y']),\\
=& \min\left\{[f(x)]_y - \max_{y'\neq y}([f(x)]_{y'} - \theta 1[y = y']), \theta 1[y = y']\right\},\\
\leq& [f(x)]_y - \max_{y'\neq y}([f(x)]_{y'} - \theta 1[y = y']),\\
=& [f(x)]_y - \max_{y'\neq y}[f(x)]_{y'} = \zeta(f(x), y).
\end{aligned}$$

Therefore $\zeta^{\theta}(f(x), y) = \min(\zeta(f(x), y), \theta)$, that implies following equality for all $\theta \geq \gamma_2$

$$\ell_{(\gamma_1,\gamma_2)}(\zeta^{\theta}(f(x), y) = \ell_{(\gamma_1,\gamma_2)}\zeta(f(x), y).$$

Now define $\mathcal{G}_L$ and $\mathcal{G}_L^{\theta}$ as follows,

$$\begin{aligned}
\mathcal{G}_L =& \{g(x, y) = \zeta(f(x), y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, f \in \mathcal{F} \text{ with } \|f\|_{\mathcal{F}} \leq L\},\\
\mathcal{G}_L^{\theta} :=& \{g(x, y) = \zeta^{\theta}(f(x), y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, f \in \mathcal{F} \text{ with } \|f\|_{\mathcal{F}} \leq L\},
\end{aligned}$$

and we can shift our attention from $\mathcal{G}_L$ to $\mathcal{G}_L^{\theta}$ which is the key to achieve a $O(K)$ factor in Theorem 1 rather than $O(K^2)$.

To see this, let $\tilde{f} := f/L_f$ be the normalized network and thus $\zeta^{2\gamma_2}(\tilde{f}(x), y) \in \mathcal{G}_1^{2\gamma_2}$. Then for any $\gamma_2 > \gamma_1 \geq 0$,

$$\begin{aligned}
P[\zeta(\tilde{f}(x), y) < \gamma_1] \leq& \mathbb{E}[\ell_{(\gamma_1,\gamma_2)}(\zeta(\tilde{f}(x), y))],\\
=& \mathbb{E}[\ell_{(\gamma_1,\gamma_2)}(\zeta^{2\gamma_2}(\tilde{f}(x), y))],
\end{aligned}$$

$$\leq \mathbb{P}_n \ell_{(\gamma_1,\gamma_2)}(\tilde{f}(x), y) + 2\mathcal{R}_n(l_{(\gamma_1,\gamma_2)} \circ \mathcal{G}_1^{2\gamma_2}) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

$$\leq \mathbb{P}_n \ell_{(\gamma_1,\gamma_2)}(\tilde{f}(x), y) + \frac{2}{\Delta}\mathcal{R}_n(\mathcal{G}_1^{2\gamma_2}) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{26}$$

where the first inequality is implied from $1[\zeta < \gamma_1] \leq \ell_{(\gamma_1,\gamma_2)}(\zeta)$, the second inequality is a direct consequence of Lemma B.1, the third inequality results from Rademacher Contraction Inequality (Lemma B.3).

Now we will do a detailed analysis on $\mathcal{R}_n(\mathcal{G}_1^{2\gamma_2})$,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{G}_1^{2\gamma_2}) &= \frac{1}{n}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i\left([f(x_i)]_{y_i} - (\max_{y'}[f(x)]_{y'} - 2\gamma_2 1[y_i = y'])\right)\right], \\
&\leq \underbrace{\frac{1}{n}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_{y_i}\right]}_{A_1} + \dots \\
&\quad + \underbrace{\frac{1}{n}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i \max_{y'}\left([f(x_i)]_{y'} - 2\gamma_2 1[y_i = y']\right)\right]}_{A_2},
\end{aligned} \tag{27}$$

For the first term A_1 in (27),

$$\begin{aligned}
A_1 &= \frac{1}{n}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i\sum_{y\in\mathcal{Y}}[f(x_i)]_y 1[y = y_i]\right], \\
&\leq \frac{1}{n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_y 1[y = y_i]\right], \\
&= \frac{1}{n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_y\left(\frac{2\cdot 1[y = y_i] - 1}{2} + \frac{1}{2}\right)\right], \\
&\leq \frac{1}{2n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_y\left(2\cdot 1[y = y_i] - 1\right)\right] + \dots \\
&\qquad + \frac{1}{2n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_y\right], \\
&\leq \frac{1}{2n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i'[f(x_i)]_y\right] + \frac{1}{2n}\sum_{y\in\mathcal{Y}}\mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\leq 1}\sum_{i=1}^{n}\epsilon_i[f(x_i)]_y\right],
\end{aligned}$$

$$\text{where } \epsilon_i' := \epsilon_i(2 \cdot 1[y = y_i] - 1) \overset{d}{=} \epsilon_i \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1,$$
$$= \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\le 1} \sum_{i=1}^{n} \epsilon_i [f(x_i)]_y\right],$$
$$\le \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{h\in\mathcal{H}_1} \sum_{i=1}^{n} \epsilon_i h(x_i)\right],$$
$$= K\mathcal{R}_n(\mathcal{H}_1).$$

For the second term A_2 in (27),

$$A_2 \le \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\le 1} \sum_{i=1}^{n} \epsilon_i \left([f(x_i)]_y - 2\gamma_2 1[y_i = y]\right)\right],$$
$$= \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\le 1} \sum_{i=1}^{n} \epsilon_i [f(x_i)]_y\right] - \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sum_{i=1}^{n} \epsilon_i 2\gamma_2 1[y_i = y]\right],$$
$$= \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{\|f\|_{\mathcal{F}}\le 1} \sum_{i=1}^{n} \epsilon_i [f(x_i)]_y\right],$$
$$\le \frac{1}{n}\sum_{y\in\mathcal{Y}} \mathbb{E}_{S,\epsilon}\left[\sup_{h\in\mathcal{H}_1} \sum_{i=1}^{n} \epsilon_i h(x_i)\right],$$
$$= K\mathcal{R}_n(\mathcal{H}_1),$$

where the first inequality is followed by Lemma B.4. Note that $\zeta^{2\gamma_2}$ allows us to take maximum over $y \in \mathcal{Y}$ rather than $y \in \mathcal{Y}/\{y_i\}$, where in the second case, we have to take summation over two indices, that is y and y_i, to get a margin function on x, and this will result in a factor $O(K^2)$. We finish the proof by combining the upper bound on A_1 and A_2 into (26),

$$P[\zeta(\tilde{f}(x), y) < \gamma_1] \le \mathbb{P}_n \ell_{\gamma_1,\gamma_2}(\tilde{f}(x), y) + \frac{4K}{\Delta}\mathcal{R}_n(\mathcal{H}_1) + \sqrt{\frac{\log(1/\delta)}{2n}},$$
$$\le \mathbb{P}_n \ell_{\gamma_2}(\tilde{f}(x), y) + \frac{4K}{\Delta}\mathcal{R}_n(\mathcal{H}_1) + \sqrt{\frac{\log(1/\delta)}{2n}},$$

where the second inequality is implied from $\ell_{(\gamma_1,\gamma_2)}(\zeta) \le 1[\zeta < \gamma_2]$. □

Remark. The key idea, that constructing ζ^θ to use summation over one index results in y in a factor $O(K)$, follows the proof of Theorem 2 in [15]. How-

ever, typical result toward multi-class margin bound has the factor $O(K^2)$ instead [7, 20].

B.4. Proof of Theorem 2

Proof of Theorem 2. Firstly, we show after normalization, the normalize margin has an upper bound,

$$\begin{aligned}\|f(x)\|_2 &= \|\sigma_l(W_l x_{l-1} + b_l)\|_2,\\ &\le L_{\sigma_l}\|W_l x_{l-1} + b_l\|_2,\\ &\le (L_{\sigma_l}\|\bar{W}_l\|_\sigma)(\|x_{l-1}\|_2 + 1)\\ &\dots\\ &\le \Pi_{i=1}^{l}(L_{\sigma_i}\|\bar{W}_i\|_\sigma)\|x\|_2 + \Sigma_{i=1}^{l}(\Pi_{j=i}^{l}(L_{\sigma_i}\|\bar{W}_i\|_\sigma)),\end{aligned}$$

where $x_i = \sigma_i(W_i x_{i-1} + b_i)$ with $x_0 = x$, $\bar{W}_i = (W_i, b_i)$ and L_{σ_i} is the Lipschitz constant of activation function σ_i with $\sigma_i(0) = 0, i = 1, \dots, l$. In the sequel as we consider the explosion of network Lipschitz over depths typically met in applications, we assume without loss of generality that $\|\bar{W}_i\|_\sigma \ge 1$ (otherwise we take the unit ball bound). Then, for normalized network $\tilde{f} = f/L_f$ with $L_f = \Pi_{i=1}^{l}(L_{\sigma_i}\|\bar{W}_i\|_\sigma)$ and $\|x\|_2 \le M$,

$$\|\tilde{f}(x)\|_2 \le M + l.$$

Therefore $\zeta(\tilde{f}(x), y) \le 2\|\tilde{f}(x)\|_2 = 2(M+l) =: M_1$, and the quantile margin is also bounded $\hat{\gamma}_{q,t} \le M_1$ for all $q \in (0,1), t = 1, \dots, T$.

The remaining proof follows the idea from [12, 20]. For any $\epsilon > 0$, we take a sequence of ϵ_k and $\gamma_k, k = 1, 2, \dots$ by $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{n}}$ and $\gamma_k = M_1 2^{-k}$. Let A_k be the event $\mathbb{P}[\zeta(\tilde{f}_t(x), y) < 0] > \mathbb{P}_n[\zeta(\tilde{f}(x), y) < \gamma_k] + \frac{4K}{\gamma_k}\mathcal{R}(\mathcal{H}_1) + \epsilon_k$. Then by Theorem 1,

$$\mathbb{P}(A_k) \le \exp(-2n\epsilon_k^2),$$

where the probability is taken over samples $\{x_1, \dots x_n\}$. We further consider the probability for none of A_k occurs,

$$\begin{aligned}\mathbb{P}(\exists A_k) &\le \Sigma_{k=1}^{\infty} P(A_k),\\ &\le \Sigma_{k=1}^{\infty}\frac{1}{k^2}\exp(-2n\epsilon^2),\\ &\le 2\exp(-2n\epsilon^2).\end{aligned}$$

Hence, fix a $q \in [0,1]$, for any $t = 1, \ldots, T$, if $\hat{\gamma}_{q,t} > 0$, there exists a $\hat{k}_t \geq 1$ (denoted as $\hat{k}$ for simplicity) such that,

$$\gamma_{\hat{k}+1} \leq \hat{\gamma}_{q,t} < \gamma_{\hat{k}}. \tag{28}$$

Therefore,

$$\begin{aligned}
A_{\hat{k}+1} &\supseteq \mathbb{P}[\zeta(\tilde{f}_t(x), y) < 0] > \mathbb{P}_n[\zeta(\tilde{f}_t(x), y) < \hat{\gamma}_{q,t}] + \frac{4K}{\gamma_{\hat{k}+1}}\mathcal{R}(\mathcal{H}_1) + \epsilon_{\hat{k}+1}, \\
&\supseteq \mathbb{P}[\zeta(\tilde{f}_t(x), y) < 0] > \mathbb{P}_n[\zeta(\tilde{f}_t(x), y) < \hat{\gamma}_{q,t}] + \frac{8K}{\hat{\gamma}_{q,t}}\mathcal{R}(\mathcal{H}_1) + \epsilon_{\hat{k}+1}, \\
&= \mathbb{P}[\zeta(\tilde{f}_t(x), y) < 0] > \mathbb{P}_n[\zeta(\tilde{f}_t(x), y) > \hat{\gamma}_{q,t}] + \frac{8K}{\hat{\gamma}_{q,t}}\mathcal{R}(\mathcal{H}_1) + \ldots \\
&\qquad + \epsilon + \sqrt{\frac{\log(\hat{k}+1)}{n}}, \\
&\supseteq \mathbb{P}[\zeta(\tilde{f}_t(x), y) < 0] > \mathbb{P}_n[\zeta(\tilde{f}_t(x), y) > \hat{\gamma}_{q,t}] + \frac{8K}{\hat{\gamma}_{q,t}}\mathcal{R}(\mathcal{H}_1) + \ldots \\
&\qquad + \epsilon + \sqrt{\frac{\log\log_2(2M_1/\hat{\gamma}_{q,t})}{n}}.
\end{aligned}$$

The first inequality is implied from $\mathbb{P}_n[\zeta(\tilde{f}_t(x), y) < \hat{\gamma}_{q,t}] > \mathbb{P}_n[\zeta(\tilde{f}_t(x), y) < \gamma_{\hat{k}+1}]$, since $\gamma_{\hat{k}+1} \leq \hat{\gamma}_{q,t}$. The second inequality is implied from $\hat{\gamma}_{q,t} < 2\gamma_{\hat{k}+1}$ and thus, $1/\gamma_{\hat{k}+1} < 2/\hat{\gamma}_{q,t}$. The third equality is the direct definition of $\epsilon_{\hat{k}}$. The last inequality is implied from $\hat{k} + 1 = \log_2(M_1/\gamma_{\hat{k}+1})$ and again, $1/\gamma_{\hat{k}+1} < 2/\hat{\gamma}_{q,t}$. The conclusion is proved immediately by letting $\epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$. □

B.5. Proof of Proposition 2

Proof of Lemma 2. (A)

$$\begin{aligned}
\|w * x\|_2^2 &= \sum_u \left(\sum_v x(v)w(u-v)\right)^2 \\
&= \sum_u \left(\sum_v (x(v)\sqrt{|w(u-v)|} \cdot \sqrt{|w(u-v)|}\right)^2 \\
&\leq \sum_u \left\{\left(\sum_v x(v)^2|w(u-v)|\right)\left(\sum_v |w(u-v)|\right)\right\},
\end{aligned}$$

$$= \|w\|_1^2 \|x\|_2^2,$$

where the second last step is due to the Cauchy-Schwartz inequality.

(B) Similarly,

$$\begin{aligned}
\|w * x\|_2^2 &= \sum_{u,j \le \mathrm{C_{out}}} \left(\sum_{v,i \le \mathrm{C_{in}}} x(v,i) w(j,i,u-v) \right)^2 \\
&= \sum_{u,j} \left(\sum_{v,i} (x(v,i) \sqrt{|w(j,i,u-v)|} \cdot \sqrt{|w(j,i,u-v)|} \right)^2 \\
&\overset{(a)}{\le} \sum_{u,j} \left\{ \left(\sum_{v,i} x(v,i)^2 |w(j,i,u-v)| \right) \left(\sum_{v,i} |w(j,i,u-v)| \right) \right\}, \\
&\overset{(b)}{=} \sum_{j} \|w(j,\cdot,\cdot)\|_1 \left(\sum_{u,v,i} x(v,i)^2 |w(j,i,u-v)| \right), \\
&\overset{(c)}{\le} \sum_{j} \|w(j,\cdot,\cdot)\|_1 \left(\sum_{v,i} x(v,i)^2 \|w(j,i,\cdot)\|_1 \right), \\
&\overset{(d)}{\le} \sum_{j} \|w(j,\cdot,\cdot)\|_1 (\max_i \|w(j,i,\cdot)\|_1) \left(\sum_{v,i} x(v,i)^2 \right), \\
&\overset{(e)}{\le} (\max_{i,j} \|w(j,i,\cdot)\|_1) \|w\|_1 \left(\sum_{v,i} x(v,i)^2 \right),
\end{aligned}$$

where the inequality (a) is due to the Cauchy-Schwartz inequality, step (b) and (c) are due to $\sum_u |w(j,i,u-v)| \le \sum_v |w(j,i,u-v)| = \|w(j,i,\cdot)\|_1$ where equality holds if the stride is 1.

In particular, for a convolution kernel w of large stride $S \ge 1$, $\sum_u |w(j, i, u-v)| \le D\|w(j,i,\cdot)\|_\infty \le \max_{i,j} D\|w(j,i,\cdot)\|_\infty$. Hence step (e) becomes

$$D\|w\|_\infty \|w\|_1 \|x\|_2^2,$$

which gives the stride-sensitive bound. □

References

[1] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *The 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA*. 2017.

[2] Peter L. Bartlett. For valid generalization the size of the weights is more important than the size of the network. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 134–140. MIT Press, 1997. MR1607706

[3] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. MR1607706

[4] Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007. MR2267655

[5] Leo Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[6] Peter Bühlmann and Bin Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2002. MR1995709

[7] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *International Conference on Machine Learning*, pages 46–54, 2013.

[8] Corinna Cortes and Vladimir N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119, 1997. MR1473055

[10] Gene H Golub and Henk A Van der Vorst. Eigenvalue computation in the 20th century. In *Numerical analysis: historical developments in the 20th century*, pages 209–239. Elsevier, 2001. MR1798518

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[12] Vladimir Koltchinskii, Dmitry Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002. MR1892654

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

[15] Vitaly Kuznetsov, Mehryar Mohri, and U Syed. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

[16] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag Berlin Heidelberg, 1991. MR1102015

[17] Qianli Liao, Brando Miranda, Andrzej Banburski, Jack Hidary, and Tomaso Poggio. A surprising linear relationship predicts test performance in deep networks. *MIT CBMM memo*, No. 91, 2018.

[18] Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003. MR2075999

[19] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *The 6th International Conference on Learning Representations (ICLR)*, 2018.

[20] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012. MR3057769

[21] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *The 6th International Conference on Learning Representations (ICLR)*, 2018.

[22] A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, 1962. MR0175722

[23] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998. MR1673273

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *The 6th International Conference on Learning Representations (ICLR)*, 2018.

[26] Matus Telgarsky. Margins, shrinkage, and boosting. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[27] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998. MR1641250

[28] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[29] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. MR3967104

[30] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. MR2327601

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

[32] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005. MR2166555

Weizhi Zhu
The Hong Kong University of Science and Technology
Hong Kong SAR
China
E-mail address: wzhuai@ust.hk

Yifei Huang
The Hong Kong University of Science and Technology
Hong Kong SAR
China
E-mail address: yhuangcc@ust.hk

Yuan Yao
Department of Mathematics
The Hong Kong University of Science and Technology
Hong Kong SAR
China
E-mail address: yuany@ust.hk

Received July 14, 2022