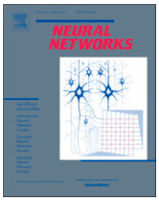
Neural Networks xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



# Improving generalization of deep neural networks by leveraging margin distribution

Shen-Huan Lyu, Lu Wang, Zhi-Hua Zhou [*]

*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China*

ARTICLE INFO

*Article history:*
Received 18 November 2021
Received in revised form 28 January 2022
Accepted 10 March 2022
Available online xxxx

*Keywords:*
Deep neural network
Margin theory
Generalization

ABSTRACT

Recent research has used margin theory to analyze the generalization performance for deep neural networks (DNNs). The existed results are almost based on the spectrally-normalized minimum margin. However, optimizing the minimum margin ignores a mass of information about the entire margin distribution, which is crucial to generalization performance. In this paper, we prove a generalization upper bound dominated by the statistics of the entire margin distribution. Compared with the minimum margin bounds, our bound highlights an important measure for controlling the complexity, which is the ratio of the margin standard deviation to the expected margin. We utilize a convex margin distribution loss function on the deep neural networks to validate our theoretical results by optimizing the margin ratio. Experiments and visualizations confirm the effectiveness of our approach and the correlation between generalization gap and margin ratio.

## 1. Introduction

Deep neural networks (DNNs) are making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years (LeCun, Bengio, & Hinton, 2015), especially in the field of computer vision (Gori, 2022). Recently, many research try to explain the practical success of DNNs via generalization, which is the ability of a classifier to perform well on unseen samples. However, some new empirical evidence has started to question this explanation. Adversarial training samples can cause the model to misclassify seriously by slight feature perturbation (Goodfellow, Shlens, & Szegedy, 2015; Papernot et al., 2017). On the other hand, Zhang, Bengio, Hardt, Recht, and Vinyals (2021) find that the deep neural networks have enough complexity to fit an arbitrarily corrupted data, and a small geometric transformation may cause networks deteriorating in performance (Azulay & Weiss, 2019). This complex and fragile nature of DNNs leads to a key problem, how to use the *data distribution* and *network parameters* to estimate the generalization ability of DNNs. Although several regularization techniques, such as dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), batch normalization (Ioffe & Szegedy, 2015), and weight decay (Krogh & Hertz, 1992), do improve the generalization performance of the over-parameterized deep models, Zhang et al. (2021) show that these regularizers cannot solve this problem either.

Consequently, several recent works (Arora, Ge, Neyshabur, & Zhang, 2018; Bartlett, Foster, & Telgarsky, 2017; Neyshabur, Bhojanapalli, & Srebro, 2018; Neyshabur, Tomioka, & Srebro, 2015) have started to address this question, proving that we can control the capacity of DNNs via different upper bounds based on the minimum margin. However, the generalization bounds based on analyses of model complexity and noise stability only focus on the minimum margin, which is based on the closest distance of the training points to the decision boundary. This notion is brittle and sensitive to outliers due to a lack of the entire margin distribution information. Jiang, Krishnan, Mobahi, and Bengio (2019) propose a measure by looking at the entire distribution of distances, and conduct empirical studies on how well it can predict the generalization gap. However, how the margin distribution information affects the generalization error of the model still needs more specific theoretical analysis, which will lead us to optimize the entire margin distribution appropriately.

The margin distribution has been shown to correspond to generalization properties in the literature on linear models and boosting algorithms, Schapire, Freund, Barlett, and Lee (1997) first introduce it to explain the phenomenon that AdaBoost seems resistant to overfitting problem. Two years later, Breiman (1999) indicates that the minimum margin is crucial for margin theory. Reyzin and Schapire (2006) conjecture that the margin distribution, rather than the minimum margin, plays a key role. The debate has been finally solved by Gao and Zhou (2013) who theoretically proved that the AdaBoost process attempts to maximize the margin mean and minimize margin variance simultaneously; highlighting for the first time that two factors rather than a single

[*] Corresponding author.
*E-mail addresses:* lvsh@lamda.nju.edu.cn (S.-H. Lyu), wangl@lamda.nju.edu.cn (L. Wang), zhouzh@lamda.nju.edu.cn (Z.-H. Zhou).

https://doi.org/10.1016/j.neunet.2022.03.019
0893-6080/© 2022 Elsevier Ltd. All rights reserved.

Please cite this article as: S.-H. Lyu, L. Wang and Z.-H. Zhou, Improving generalization of deep neural networks by leveraging margin distribution. Neural Networks (2022), https://doi.org/10.1016/j.neunet.2022.03.019.

factor are crucial for margin theory. These two factors are the first and second-order statistics describing the margin distribution, while in most cases the higher-order one is less useful. Their result successfully explains why AdaBoost seems resistant to overfitting: even when the training error reaches zero, the margin mean can be increased and/or the margin variance can be decreased further, leading to the improvement of generalization performance; it also discloses that AdaBoost will finally overfit: when the margin mean cannot be increased and margin variance cannot be decreased further. The long march of the theoretical exploration of AdaBoost is summarized in Zhou (2014), and Gao and Zhou (2013)'s result has been theoretically confirmed by Grønlund, Kamma, Larsen, Mathiasen, and Nelson (2019). Inspired by Gao and Zhou (2013)'s finding, powerful learning machines can be built by maximizing the margin mean and minimizing margin variance simultaneously, rather than simply maximizing the minimum margin like in traditional large-margin machines. Zhang and Zhou (2019) propose the optimal margin distribution machine (ODM) for binary classification. In Tan, Tan, Jiang, and Zhou (2020), Zhang, Zhao, and Jin (2020), Zhang and Zhou (2017, 2018a, 2018b), ODM is extended to many forms.

In this paper, we study a $d$-layer feed-forward network with ReLU activation functions. Our theoretical result states that the statistics of margin distribution play an important role in the generalization estimation rather than the traditional minimum margin. This result is consistent with the previous results on boosting and linear algorithms (Gao & Zhou, 2013; Schapire et al., 1997; Zhang & Zhou, 2019). It also inspires us to understand the similarities between deep learning and traditional machine learning from the perspective of margin distribution. Specially, we propose a new loss function to optimize the statistics-based measure in the theoretical results. A strong correlation between generalization and our measure is empirically shown by studying a wide range of network structures trained on the MNIST, CIFAR-10 and ImageNet datasets. The detailed contributions of this paper are as follows:

***PAC guarantee***. Our bound shows that we can restrict the capacity of deep nets by the ratio of second- to first-order statistic of margin distribution at the last layer. Compared with the existing results based on minimum margin (Arora et al., 2018; Bartlett et al., 2017; Neyshabur et al., 2018), our bound contains more information on the entire margin distribution to estimate the generalization error. Moreover, the empirical evaluation shows that optimizing the margin ratio can control the model capacity to alleviate the overfitting risk.

***Optimization***. Inspired by our theoretical result, we encourage DNNs to optimize the margin ratio for better generalization performance. Therefore, we propose a new approach called margin distribution Networks (mdNet), which utilizes a convex margin distribution loss function to optimize the first- and second-order statistics of margin. Moreover, we empirically evaluate our loss function on deep neural networks across different image datasets and model structures. Specifically, empirical results demonstrate the effectiveness of the proposed method in learning tasks with limited training data.

The rest of paper is organized as follows. The related work is introduced in Section 2. Some notations are introduced in Section 3. In Section 4, we present a generalization bound leveraging margin distribution rather than minimum margin and demonstrate that the ratio of the margin standard deviation to the expected margin is the key to control the model capacity. Section 5 lists the detailed proofs for our theorems and lemmas. In Section 6, we formulate the convex loss function to optimize the margin ratio. Section 7 reports our experimental studies and empirical observations. Finally, Section 8 concludes with future work.

## 2. Related work

Recently, margin-based deep learning algorithms have developed rapidly. Schroff, Kalenichenko, and Philbin (2015) use the triplet loss to encourage a distance constraint similar to the contrastive loss. Similarly, Chan et al. (2015) enhance the supervision of the learned filters by incorporating the information of class labels in the training data and learn the filters based on the idea of multi-class linear discriminant analysis (LDA) for classification task. Liu, Wen, Yu, and Yang (2016) propose a generalized large-margin softmax loss which explicitly encourages *intra-class compactness* and *inter-class separability* in the learned representation space. It would be interesting to theoretically study feature space transformation which might be a key to understanding mysteries behind the successes of deep neural networks (Zhou, 2021). Since Arora et al. (2018) and Bartlett et al. (2017) associate the generalization of deep neural networks with the minimum margin, a line of work establishes that first-order methods can automatically maximize the minimum margin in the settings of logistic regression (Gunasekar, Lee, Soudry, & Srebro, 2018a), deep linear networks (Gunasekar, Lee, Soudry, & Srebro, 2018b; Ji & Telgarsky, 2019; Li, Ma, & Zhang, 2018; Soudry, Hoffer, Nacson, Gunasekar, & Srebro, 2018), and symmetric matrix factorization (Li et al., 2018). However, Wei, Lee, Liu, and Ma (2018) point that how to extend these results to non-linear neural networks remains unclear. Recently, Wu, Jing, Du, and Chen (2021) propose to understand the model dynamics from the perspective of control theory. Another line of algorithm-dependent analysis of generalization (Chen, Jin, & Yu, 2018; Hardt, Recht, & Singer, 2016; Mou, Wang, Zhai, & Zheng, 2018) uses stability of specific optimization algorithms that satisfy certain generic properties like convexity, smoothness, etc. Specially, Dinh, Pascanu, Bengio, and Bengio (2017), Keskar, Mudigere, Nocedal, Smelyanskiy, and Tang (2017) and Zhu, Wu, Yu, Wu, and Ma (2019) make a connection between the sharpness of the solution obtained using the SGD algorithm and its ability to generalize well. The notion of sharpness corresponds to robustness to adversarial perturbations of parameters. Furthermore, Neyshabur, Bhojanapalli, McAllester, and Srebro (2017) and Neyshabur et al. (2018) draw a connection to the PAC-Bayesian theory for sharpness. The margin distribution measure presented in this paper is closely related to sharpness (Keskar et al., 2017), because we use the statistics of the margin distribution to theoretically describe the value of the allowable perturbation. Compared with the sharpness measure which is difficult to optimize, the margin distribution measure proposed in this paper is easy to calculate, and can be directly optimized through the SGD algorithm by designing a convex loss function. Recently, Jiang et al. (2019) present abundant empirical evidence to validate that the generalization in deep learning can be estimated from the margin statistics. In addition, the relevant theories of domain adaptation (Mansour, Mohri, & Rostamizadeh, 2009; Mansour & Schain, 2014; Zhang, Zhang, & Ye, 2012) are also used to improve the generalization capability of deep learning (Becker, Christoudias, & Fua, 2013; Koniusz, Tas, & Porikli, 2017; Pan, Tsang, Kwok, & Yang, 2009; Rozantsev, Salzmann, & Fua, 2019). Domain generalization cannot see existing training source domains during training. This makes domain generalization more challenging than domain adaptation but more realistic and favorable in practical applications (Dubey, Ramanathan, Pentland, & Mahajan, 2021; Ghifary, Kleijn, Zhang, & Balduzzi, 2015; Matskevych, Wolny, Pape, & Kreshuk, 2022; Wang, Lan, Liu, Ouyang, & Qin, 2021).

## 3. Notations

Consider the multi-class task with feature domain $\mathcal{X}$ and label domain $\mathcal{Y}$. Let $\mathcal{D}$ be an unknown (underlying) distribution over $\mathcal{X} \times \mathcal{Y}$. A training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ and a validation set $S' = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{m'}, y_{m'})\}$ are drawn identically and independently according to $\mathcal{D}$. We denote a labeled sample as $(\boldsymbol{x}, y) \in \mathcal{D}$.

Let $f_{\boldsymbol{w}}: \mathcal{X}_{B,n} \to \mathcal{Y}'$ be the function represented by a $d$-layer feed-forward network with parameters

$$\boldsymbol{w} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_d\}$$

and output domain $\mathcal{Y}' = \mathbb{R}^k$. The entire network can be formulated as

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{W}_d \phi(\boldsymbol{W}_{d-1}\phi(\ldots\phi(\boldsymbol{W}_1\boldsymbol{x}))),$$

where $\phi$ is the ReLU activation function and let $\rho$ be an upper bound on the number of output units in each layer.

We can define the fully connected networks (FNNs) recursively:

$$\boldsymbol{x}^1 = \boldsymbol{W}_1\boldsymbol{x} \quad \text{and} \quad \boldsymbol{x}^i = \boldsymbol{W}_i\phi(\boldsymbol{x}^{i-1}),$$

where $\boldsymbol{x}^i$ denotes the output of the $i$th layer.

The predicted label is denoted by

$$h(\boldsymbol{x}) = \arg\max_j f_{\boldsymbol{w},j}(\boldsymbol{x}) \in \mathcal{H},$$

where $h: \mathcal{X} \to \mathcal{Y}$ is a map from the feature domain to the label domain and $f_{\boldsymbol{w},j}$ is the $j$th element of the score vector.

In the multi-class setting (Mohri, Rostamizadeh, & Talwalkar, 2018, Chapter 9.2), the label associated to point $\boldsymbol{x}$ is the one resulting in the largest score $h(\boldsymbol{x}) = \arg\max_i f_{\boldsymbol{w},i}(\boldsymbol{x})$. This naturally leads to the following definition of the margin $\gamma_h(\boldsymbol{x}, y)$ of the function $h$ at a labeled example $(\boldsymbol{x}, y)$:

$$\gamma_h(\boldsymbol{x}, y) = f_{\boldsymbol{w},y}(\boldsymbol{x}) - \max_{j \neq y} f_{\boldsymbol{w},j}(\boldsymbol{x}). \tag{1}$$

Thus, $h$ misclassifies $(\boldsymbol{x}, y)$ iff $\gamma_h(\boldsymbol{x}, y) < 0$.

## 4. Margin distribution rather than minimum margin

In Section 4.1, we list error-resilience assumptions that will be used. In Section 4.2, we introduce the existed results based on the minimum margin. In Section 4.3, we present our main results based on the entire margin distribution.

### 4.1. Error-resilience assumptions

Here we formalize the error-resilience properties for deep neural networks. Arora et al. (2018) show that if we inject a scaled Gaussian noise to the input of deep nets, as it propagates up, the noise has rapidly decreasing effect on higher layers. This fact implies *compressibility* of deep nets, i.e., low rank of parameters' matrix. The empirical version of noise-sensitivity parameters is first proposed by Arora et al. (2018). It inspires us to bound the perturbation caused by Gaussian noise with the validation-based version of noise-sensitivity parameters below.

**Assumption 1** (*Layer Cushion*)**.** The layer cushion of layer $i$ is defined to be largest number $\mu_i$ such that for any validation data $\boldsymbol{x} \in S'$:

$$\mu_i \|\boldsymbol{W}_i\|_F \|\phi(\boldsymbol{x}^{i-1})\|_2 \leq \|\boldsymbol{x}^i\|_2. \tag{2}$$

**Assumption 2** (*Interlayer Cushion*)**.** For any two layers $i < j$, we define the interlayer cushion $\mu_{i,j}$, as the largest number such that for any validation data $\boldsymbol{x} \in S'$:

$$\mu_{i,j} \|J^{i,j}_{\boldsymbol{x}^i}\|_F \|\phi(\boldsymbol{x}^{i-1})\|_2 \leq \|\boldsymbol{x}^j\|_2. \tag{3}$$

Furthermore, for any layer $i$ we define the minimal interlayer cushion as $\mu_{i\to} = \min_{i \leq j \leq L} \mu_{i,j} = \min\{\frac{1}{\sqrt{\rho}}, \min_{i \leq j \leq L} \mu_{i,j}\}$. For any two layer $i < j$, denote by $M^{i,j}$ the operator for composition of these layers and $J^{i,j}_{\boldsymbol{x}}$ be the Jacobian matrix (the partial derivative) of this operator at input $\boldsymbol{x}$. Therefore, we have $\boldsymbol{x}^j = M^{i,j}(\boldsymbol{x}^i)$. Furthermore, since the activation functions are ReLU (hence piece-wise linear), we have $M^{i,j}(\boldsymbol{x}^i) = J^{i,j}_{\boldsymbol{x}^i}\boldsymbol{x}^i$.

**Assumption 3** (*Interlayer Smoothness*)**.** For any two layers $i < j$, we define the interlayer smoothness $\rho_\delta$ as the smallest number such that with probability $1 - \delta$ over noise $\eta$ for any validation data $\boldsymbol{x} \in S'$:

$$\|M^{i,j}(\boldsymbol{x}^i + \eta) - J^{i,j}_{\boldsymbol{x}^i}(\boldsymbol{x}^i + \eta)\| \leq \frac{\|\eta\| \|\boldsymbol{x}^j\|}{\rho_\delta \|\boldsymbol{x}^i\|} \tag{4}$$

For a single layer, $\rho_\delta$ captures the ratio of input/weight alignment to noise/weight alignment. Arora et al. (2018) show that the interlayer smoothness is indeed good: $1/\rho_\delta$ is a small constant.

The next two conditions qualify a common appearance: if the input in the activation and margin calculations is well-distributed and the calculations do not correlate with the magnitude of the input, then one would expect that, the effect of applying activation at any layer and margin at last layer is to decrease the norm of the vector by at most some small constant factor, i.e., $c$ and $\alpha$.

**Assumption 4** (*Activation Contraction*)**.** The activation contraction $c$ is defined as the smallest number such that for any layer $i$ and any validation data $\boldsymbol{x} \in S'$:

$$c\|\phi(\boldsymbol{x}^i)\|_2 \geq \|\boldsymbol{x}^i\|_2. \tag{5}$$

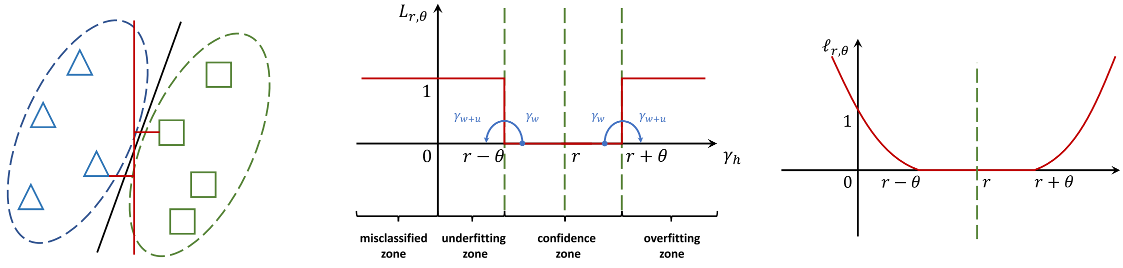
**Assumption 5** (*Margin Contraction*)**.** The margin contraction $\alpha$ is defined as the smallest number such that for any validation data $\boldsymbol{x} \in S'$:

$$\alpha\|\gamma_h(\boldsymbol{x}, y)\|_2 \geq \|\boldsymbol{x}^d\|_2. \tag{6}$$

In this paper, we only use the noise-sensitivity parameters in Assumptions 1–5 as descriptions of error-resilience properties, from which the margin distribution term of our bound is derived. Therefore, we just need estimate these parameters based on validation data to show the magnitude of our bound rather than optimizing these parameters in the training process like Arora et al. (2018) did.

### 4.2. Existed results

In the deep learning theory community, great efforts have been made to explain why over-parameterized deep neural networks can success, which is contrary to the classical VC dimension analysis (Bartlett, Maiorov, & Meir, 1998; Harvey, Liaw, & Mehrabian, 2017). Bartlett et al. (2017) and Neyshabur et al. (2018) made an important stride by showing minimum margin based bounds for multi-layer neural networks. These bounds do not depend directly on the number of parameters of the network but depends on the normalized minimum margin. Theorem 1 provides a unified description of these bounds. The only difference between them lies in the value of constants and the type of norms.



(a) The minimum margin based classifier (red line) and the margin distribution based classifier (black line). The blue triangle and green square represent the two classes of instances, while the dotted ellipses represent their underlying distribution.

(b) The $(r,\theta)$-margin distribution loss function (red line). The green dotted lines represent the confidence area of margins, which has zero loss. Lemma 4 shows that the perturbation caused by $\boldsymbol{u}$ (blue arrow) is related to the generalization error.

(c) The convex margin distribution loss function (red line). This convex function is used as an alternative function of $(r,\theta)$-margin distribution loss function, so that the deep neural networks can optimize the margin distribution through SGD algorithm.

**Fig. 1.** Illustration of the margin distribution analysis and loss functions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Theorem 1** (Bartlett et al., 2017; Neyshabur et al., 2018)**.** *For any $d, \rho > 0$ and $\|\boldsymbol{x}\|_2 \leq B$, let $f_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^k$ be a d-layer feed-forward network with ReLU activation. Then, for any $\delta > 0$, with probability $\geq 1-\delta$ over a training set of size m, for any $\boldsymbol{w} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \dots, \boldsymbol{W}_d\}$, we have:*

$$L_0\left(f_{\boldsymbol{w}}\right) \leq \widehat{L}_\gamma\left(f_{\boldsymbol{w}}\right) + \mathcal{O}\left(\sqrt{\frac{B^2 d^2 \rho \ln(dh) \Pi_{i=1}^{d} \|\boldsymbol{W}_i\|_2^2 \sum_{i=1}^{d} \frac{\|\boldsymbol{W}_i\|_F^2}{\|\boldsymbol{W}_i\|_2^2} + \ln \frac{dm}{\delta}}{\gamma^2 m}}\right), \tag{7}$$

*where $L_0(\cdot)$ is the 0–1 loss, $\widehat{L}_\gamma(\cdot) = \mathrm{Pr}_S\left[\gamma_h(\boldsymbol{x}, y) \leq \gamma\right]$ is the empirical estimation of $\gamma$-margin loss and $\mathcal{O}(\cdot)$ describes the limiting behavior of a function.*

Based on this margin theory view, Arora et al. (2018) provide an improved bound by considering the compressibility of deep nets as follows:

**Theorem 2** (Arora et al., 2018)**.** *For any $d > 0$, let $f_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^k$ be a d-layer feed-forward network with ReLU activation. Then, for any $\delta > 0$, with probability $\geq 1-\delta$ over a training set of size m, for any $\boldsymbol{w} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \dots, \boldsymbol{W}_d\}$, we have:*

$$L_0\left(f_{\boldsymbol{w}}\right) \leq \widehat{L}_\gamma\left(f_{\boldsymbol{w}}\right) + \mathcal{O}\left(\sqrt{\frac{c^2 d^2 \max_{\boldsymbol{x}\in S} \|f_{\boldsymbol{w}}(\boldsymbol{x})\|_2^2 \sum_{i=1}^{d} \frac{1}{\mu_i^2 \mu_{i\to}^2}}{\gamma^2 m}}\right) \tag{8}$$

*where $\mu_i, \mu_{i\to}, c, \alpha$ are layer cushion, interlayer cushion, activation contraction and interlayer smoothness defined in Assumptions 1, 2, 4 and 5 respectively*

These existed results follow the traditional margin theory, so they only focus on the minimum margin $\gamma$. Because they lack the description of the entire margin distribution, they can only take the minimum margin $\gamma$ as the optimization target to improve the generalization performance. These methods ignore the information of the entire margin distribution. In the next subsection, we expect to prove a bound related to the entire margin distribution, so as to inspire us to directly optimize margin distribution for improving the generalization performance for DNNs.

### 4.3. Main results

We begin with an intuitive comparison of the minimum margin based classifier and the margin distribution based classifier. Fig. 1(a) shows that maximizing the minimum margin will make the classifier easy to be misled by a small number of samples, thus ignoring the distribution information of samples, while the margin distribution based classifier considers the mean and variance of samples and generalizes better.

In order to utilize the mean and variance information into the theoretical analysis, we design a new margin loss, which uses $r$ to adjust the mean of margin and $\theta$ to adjust the variance of margin. For any parameter $r > \theta > 0$, we can define a $(r,\theta)$-margin distribution loss function (see Fig. 1(b)), which penalizes $h$ with a cost of 1 when it predicts $\boldsymbol{x}$ with a margin smaller than $r-\theta$, but also penalizes $h$ when it predicts $\boldsymbol{x}$ with a margin larger than $r+\theta$. The margin distribution bound is presented in terms of this loss function, which is formally defined as follows.

**Definition 1** (*Expected Margin Distribution Loss Function*)**.** For any $r > \theta > 0$, the $(r,\theta)$-margin loss is the function $L_{r,\theta}(\cdot)$ defined for all $h \in \mathcal{H}$ as:

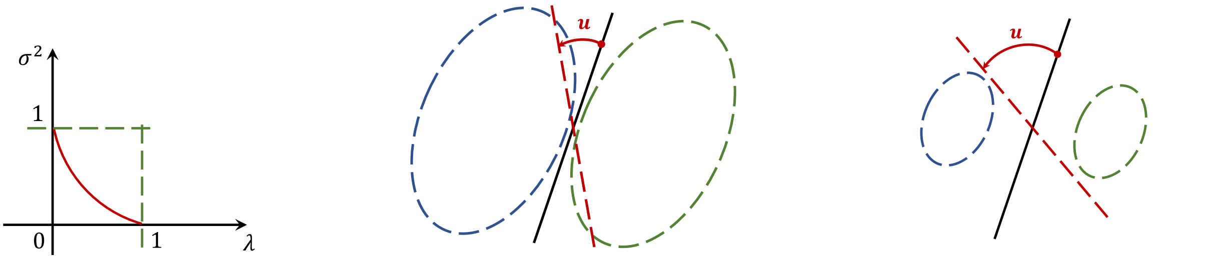
$$L_{r,\theta}(h) = \Pr_{\mathcal{D}}\left[\gamma_h(\boldsymbol{x}, y) \leq r-\theta\right] + \Pr_{\mathcal{D}}\left[\gamma_h(\boldsymbol{x}, y) > r+\theta\right]. \tag{9}$$

Intuitively, our $(r,\theta)$-margin distribution loss function looks for a classifier $h$ which forces as many data points as possible into the *zero-loss band* $(r-\theta \leq \gamma_h(\boldsymbol{x}, y) < r+\theta)$. Therefore, we let $r \simeq \mathbb{E}_{\mathcal{D}}[\gamma_h(\boldsymbol{x}, y)]$, $\theta^2 \simeq \mathrm{Var}_{\mathcal{D}}[\gamma_h(\boldsymbol{x}, y)]$, which implies that the expected margin is larger than the standard deviation. Actually, $\theta$ just need to be a second-order statistic, so we can re-scale $\theta^2 = a \cdot \mathrm{Var}_{\mathcal{D}}[\gamma_h(\boldsymbol{x}, y)]$ to satisfy $r > \theta$. In this way, the $(r,\theta)$-margin distribution loss is a surrogate loss function. In particular, for $r = \theta$ and $\theta \to \infty$, the zero-loss band is the positive area $(\gamma_h(\boldsymbol{x}, y) > 0)$ and $L_{r,\theta}$ corresponds to the 0–1 loss $L_0$. Let $\widehat{L_{r,\theta}}(f_{\boldsymbol{w}})$ be the empirical estimate of the expected margin distribution loss. So we also denote the expected risk and the empirical risk as $L_0(f_{\boldsymbol{w}})$ and $\widehat{L_0}(f_{\boldsymbol{w}})$, which are bounded between 0 and 1.

We begin with bounding the change of output caused by the noise on the classifier $\boldsymbol{u}$ with the noise-sensitivity parameters and the statistics of margin distribution:

**Lemma 3.** *Let $f_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^k$ be a d-layer network. For any $d > 0$, and $vec(\{\boldsymbol{U}_i\}_{i=1}^{d}) = (\boldsymbol{U}_1, \boldsymbol{U}_2, \dots, \boldsymbol{U}_d)$ is a vector of perturbation parameters with $\boldsymbol{U}_i = \boldsymbol{\beta}_i \|\boldsymbol{W}_i\|_F$, and $\boldsymbol{\beta} = vec(\{\boldsymbol{\beta}_i\}_{i=1}^{d}) = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_d)$ is a vector of random vectors with $\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\top] = \sigma^2 \boldsymbol{I}$, the change of the output of the network can be bounded with a fixed probability ($\delta = 1/2$):*

$$\max_{\boldsymbol{x}\in\mathcal{X}} |f_{\boldsymbol{w}+\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2^2 \leq \mathcal{O}\left(\sum_{i=1}^{d} \frac{d\alpha^2 c^2 \sigma^2 (r+\theta)^2}{\mu_i^2 \mu_{i\to}^2}\right). \tag{10}$$



(a) $\sigma^2 \propto \left(\frac{1-\lambda}{1+\lambda}\right)^2$.

(b) Fixed $r$, larger $\theta$, larger $\lambda$, smaller $\sigma^2$.

(c) Fixed $r$, smaller $\theta$, smaller $\lambda$, larger $\sigma^2$.

**Fig. 2.** Illustration of the relationship between margin distribution and allowable perturbation.

The result shows that the perturbation caused by $\boldsymbol{u}$ increases with the variance $\sigma^2$ is related to the outermost edge of margin distribution $r+\theta$, that is, the right green dotted line in Fig. 1(b). The next Lemma shows that we can bound the generalization gap through a Kullback–Leibler divergence term, if we can guarantee that the perturbation caused by $\boldsymbol{u}$ is smaller than $\frac{r-\theta}{8}$ with a constant probability. Therefore, the allowable value of $\sigma^2$ under the $(r,\theta)$-margin distribution assumption is consistent with the intuitive understanding (see Fig. 2), i.e., $\sigma^2 \propto \frac{r-\theta}{r+\theta} = \frac{1-\lambda}{1+\lambda}$, where $\lambda = \theta/r \in (0,1)$. When the margin distribution is more compact (smaller $\lambda$), the larger noise $\sigma^2$ can be allowed, that is, it is not easy to cause misclassification. When the margin distribution is more loose (larger $\lambda$), even a small noise have misclassification risk.

**Lemma 4.** *Let $f_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^k$ be any predictor with parameters $\boldsymbol{w}$, and P be any distribution on the parameters that is independent of the training data. Then, for any $r > \theta > 0$, $\delta > 0$, with probability at least $1-\delta$ over the training set of size m, for any $\boldsymbol{w}$, and any random perturbation $\boldsymbol{u}$ s.t. $\Pr_{\boldsymbol{u}}\left[\max_{\boldsymbol{x}\in\mathcal{X}} |f_{\boldsymbol{w}+\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2 < \frac{r-\theta}{8}\right] \geq \frac{1}{2}$, we have:*

$$L_0(f_{\boldsymbol{w}}) \leq \widehat{L}_{r,\theta}(f_{\boldsymbol{w}}) + 4\sqrt{\frac{D_{\mathrm{KL}}(\boldsymbol{w}+\boldsymbol{u}||P) + \ln\frac{6m}{\delta}}{m-1}}. \qquad (11)$$

The detailed proof is presented in Section 5.2. This Lemma improves the result of Neyshabur et al. (2018, Lemma 1), especially using two parameters $\theta, r$ to describe the entire margin distribution instead of using one parameter $\gamma$ to describe the minimum margin. Based on this result, we can derive the following generalization bound, with proof deferred to Section 5.3 showing that the Kullback–Leibler divergence term is inversely proportional to $\sigma^2$.

**Theorem 5.** *For any $d, \rho > 0$, let $f_{\boldsymbol{w}} : \mathcal{X} \to \mathbb{R}^k$ be a d-layer feed-forward network with ReLU activation. Then, for any $\delta > 0$, with probability $\geq 1-\delta$ over a training set of size m, for any $\boldsymbol{w}$, we have:*

$$L_0(f_{\boldsymbol{w}}) \leq \widehat{L}_{r,\theta}(f_{\boldsymbol{w}}) + \mathcal{O}\left(\sqrt{\frac{\left(\frac{1+\lambda}{1-\lambda}\right)^2\left(\sum_{i=1}^{d}\frac{d\alpha^2c^2}{\mu_i^2\mu_{i\to}^2}\right) + \ln\frac{6m}{\delta}}{m}}\right). \qquad (12)$$

*where the margin ratio is defined by $\lambda = \theta/r$.*

We prove an upper bound on generalization gap related to the margin ratio term, where $\lambda$ is a parameter denoting the ratio of the margin standard deviation $\theta$ to the expected margin $r$ over the underlying distribution, and the error-resilience term relies on the noise sensitivity (Arora et al., 2018) quantified by $\mu_i, \mu_{i\to}, c, \alpha$ (see Assumptions 1, 2, 4 and 5). Theorem 5 states that the entire margin distribution has much leverage in generalization performance rather than the minimum margin. Specifically, restricting a smaller $\lambda$ (larger $r$ and smaller $\theta$) can effectively control the capacity of models, so as to reduce the risk of overfitting. It inspires us that optimizing margin distribution can get better generalization performance than the traditional minimum margin maximization algorithm.

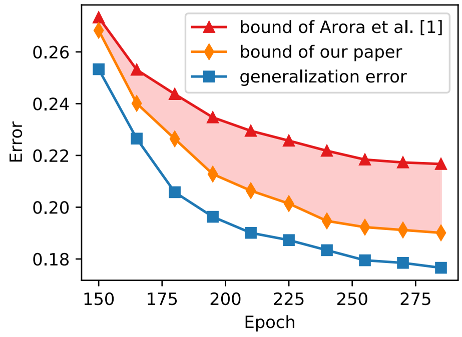


**Fig. 3.** Comparing our bound and Arora et al. (2018) to empirical generalization error during training. All bounds are rescaled to be within the same range as the generalization error together.

**Discussion.** The main difference between Arora et al. (2018) and our paper: Arora et al. (2018) proved that the generalization performance of deep neural network is related to the *sparsity* of its parameters, focusing on how to compress the parameters of the trained model. Our paper studies the relationship between DNN generalization performance and *margin distribution* under the condition that DNN parameters is sparsity, focusing on optimizing margin distribution during training. Fig. 3 shows that the improvement of our bound relative to Arora et al. (2018) (the shaded part in the figure) is because the margin ratio will gradually decrease during the training process. The main difference between Jiang et al. (2019) and our paper: Jiang et al. (2019) conjectured that the generalization performance of DNN may be related to the interval distribution. The correlation between generalization and $R^2$ (Glantz & Slinker, 2001) is calculated experimentally, and no theoretical proof is given. Our paper proves theoretically that the generalization performance of DNNs can be bound by the margin ratio and gives the improved algorithm.

## 5. Proofs

In this section, we provide the detailed proofs for the main theorem and lemmas. First, we present a useful lemma as follows:

**Lemma 6.** *Let $\mathcal{Q}$ be a probability distribution over the reals. For any random variable $v, v_1, v_2, \ldots, v_{m'} \sim \mathcal{Q}$ identically and independently (i.i.d.), we have*

$$\Pr_{v\sim\mathcal{Q}}\left[v \geq \max_{v_1,\ldots,v_{m'}\sim\mathcal{Q}}\{v_1, v_2, \ldots, v_{m'}\}\right] = \frac{1}{m'+1}. \tag{13}$$

**Proof of Lemma 6.** Let the Cumulative Distribution Function (CDF) and Probability Density Function (PDF) of random variable $v$ be $F(x)$ and $f(x)$, and we denote the maximum of a set of $m'$ random variables by $v^{(m')} = \max_{v_1,\ldots,v_{m'}\sim\mathcal{Q}}\{v_1, v_2, \ldots, v_{m'}\}$. Then

$$\Pr_{v_1,\ldots,v_{m'}\sim\mathcal{Q}}\left[v^{(m')} \leq x\right] \tag{14}$$

$$= \Pr_{v_1,\ldots,v_{m'}\sim\mathcal{Q}}[(v_1 \leq x) \wedge \cdots \wedge (v_{m'} \leq x)] \tag{15}$$

$$= \Pr_{\mathcal{Q}}[v_1 \leq x] \times \ldots \times \Pr_{\mathcal{Q}}[v_{m'} \leq x] = F^{m'}(x). \tag{16}$$

In other word, the CDF and PDF of the minimum $v^{(m')}$ are $F^{m'}(x)$ and $m'F^{m'-1}(x)f(x)$. Then we can use the minimum value of the sample's set to bound the random variable $v$ with a probability $\frac{m'}{m'+1}$, which converges to 1 with a rate $\mathcal{O}(1/m')$:

$$\Pr_{v,v_1,\ldots,v_{m'}\sim\mathcal{Q}}\left[v \leq v^{(m')}\right] \tag{17}$$

$$= \Pr_{v,v_1,\ldots,v_{m'}\sim\mathcal{Q}}\left[v - v^{(m')} \leq 0\right] \tag{18}$$

$$= \iint_{x\leq y} f(x)F^{m'-1}(y)f(y)\,\mathrm{d}x\,\mathrm{d}y \tag{19}$$

$$= \int_{-\infty}^{+\infty} m'F^{m'-1}(y)f(y)\,\mathrm{d}y \int_{-\infty}^{y} f(x)\,\mathrm{d}x \tag{20}$$

$$= \int_{-\infty}^{+\infty} m'F^{m'}(y)f(y)\,\mathrm{d}y \tag{21}$$

$$= m'F^{m'+1}(y)\mid_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} m'^2F^{m'}(y)f(y)\,\mathrm{d}y. \tag{22}$$

According to Eqs. (21) and (22), we have

$$\Pr_{v,v_1,\ldots,v_{m'}\sim\mathcal{Q}}\left[v \leq v^{(m')}\right] = \frac{m'}{m'+1}. \quad \square \tag{23}$$

### 5.1. Proof of Lemma 3

We begin with a lemma as follows:

**Lemma 7.** *For any layer i, the point-wise* compressibility *of the layer-wise parameters can hold with a probability $1 - \frac{1}{m'+1}$ over $\boldsymbol{x} \in \mathcal{D}$ as follows:*

$$\mu_i\|\boldsymbol{W}_i\|_F\|\phi(\boldsymbol{x}^{i-1})\|_2 \leq \|\boldsymbol{x}^i\|_2, \tag{24}$$

$$\mu_{i,j}\|J_{\boldsymbol{x}^i}^{i,j}\|_F\|\phi(\boldsymbol{x}^{i-1})\|_2 \leq \|\boldsymbol{x}^j\|_2, \tag{25}$$

$$\|M^{i,j}(\boldsymbol{x}^i + \eta) - J_{\boldsymbol{x}^i}^{i,j}(\boldsymbol{x}^i + \eta)\| \leq \frac{\|\eta\|\|\boldsymbol{x}^j\|}{\rho_\delta\|\boldsymbol{x}^i\|} \tag{26}$$

$$c\|\phi(\boldsymbol{x}^i)\|_2 \geq \|\boldsymbol{x}^i\|_2, \tag{27}$$

$$\alpha\|\gamma_h(\boldsymbol{x}, y)\|_2 \geq \|\boldsymbol{x}^d\|_2, \tag{28}$$

*where $m'$ is the size of the validation set.*

**Proof of Lemma 7.** According the independence between $S$ and $S'$, we can regard the noise-sensitivity parameters $\frac{1}{\mu_i}$, $\frac{1}{\mu_{i,j}}$, $c$ and $\alpha$ as random variables over reals relying on the randomness of variable $\boldsymbol{x} \in S'$. Then, the cushion parameters defined in Assumptions 1–5 can be interpreted as choosing the maximum of multiple independent samples. We first prove Lemma 6 on the tail of a random variable $v \sim \mathcal{Q}$ by choosing the maximum of multiple independent samples of the random variable. Specifically, using the following simple lemma based on the distribution of the maximum, we can guarantee the point-wise *compressibility* of the learned parameters over the underlying data distribution $\mathcal{D}$ with a high probability by calculating the maximum of the empirical dataset, i.e., $\frac{1}{\mu_i}$, $\frac{1}{\mu_{i,j}}$, $c$ and $\alpha$. □

**Proof of Lemma 3.** First, we need to bound the perturbation of linear operator caused by injecting a scaled Gaussian noise $\boldsymbol{U} = \beta\|\boldsymbol{W}\|_F$, $\mathbb{E}[\beta\beta^\top] = \sigma\boldsymbol{I}$. For any fixed vectors $\boldsymbol{a}, \boldsymbol{b}$, we have

$$\mathbb{E}_\beta\|\boldsymbol{a}^\top(\boldsymbol{W} + \boldsymbol{U})\boldsymbol{b} - \boldsymbol{a}^\top\boldsymbol{W}\boldsymbol{b}\|_2 = \mathbb{E}_\beta\|\boldsymbol{b}\|_2\|\boldsymbol{a}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{a}\|_2$$

$$= \mathbb{E}_\beta\|\boldsymbol{W}\|_F\|\boldsymbol{b}\|_2\|\boldsymbol{a}^\top\beta\beta^\top\boldsymbol{a}\|_2 \tag{29}$$

$$= \sigma\|\boldsymbol{W}\|_F\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2. \tag{30}$$

According the Markov inequality, we have

$$\Pr_\beta\left[\|\boldsymbol{a}^\top(\boldsymbol{W} + \boldsymbol{U})\boldsymbol{b} - \boldsymbol{a}^\top\boldsymbol{W}\boldsymbol{b}\|_2 \geq \sigma\sqrt{d}/\sqrt{\delta}\|\boldsymbol{W}\|_F\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2\right]$$

$$\leq \frac{\sigma^2\|\boldsymbol{W}\|_F^2\|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2}{d\sigma^2/\delta\|\boldsymbol{W}\|_F^2\|\boldsymbol{a}\|_2^2\|\boldsymbol{b}\|_2^2} = \frac{\delta}{d}. \tag{31}$$

Now, we will bound the perturbation of the $d$-layer deep nets by induction. For any layer $i \geq 0$, let $\boldsymbol{x}^j$ be the output at layer $j$ with original net and $\hat{\boldsymbol{x}}_i^j$ be the output at layer $j$ if the weights $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_i$ in the first layers are replaced with $\boldsymbol{W}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{W}_i + \boldsymbol{U}_i$. The induction hypothesis is then following:

Consider any $0 < \epsilon \leq 1$, the following is true with probability $1 - \frac{i\delta}{d}$ over $\boldsymbol{W}_1 + \boldsymbol{U}_1, \ldots, \boldsymbol{W}_i + \boldsymbol{U}_i$ for any $j \geq i$:

$$\|\hat{\boldsymbol{x}}_i^j - \boldsymbol{x}^j\|_2^2 \leq \sum_{l=1}^{i}\frac{c^2d\sigma^2}{\delta\mu_l^2\mu_{l\rightarrow}^2}\|\boldsymbol{x}^j\|_2^2. \tag{32}$$

For the base case $i = 0$, since we are not perturbing the input, the inequality is trivial. Now assuming that the induction hypothesis is true for $i - 1$, we consider what happens at layer $i$.

$$\|\hat{\boldsymbol{x}}_i^j - \boldsymbol{x}^j\|_2^2 = \|(\hat{\boldsymbol{x}}_i^j - \hat{\boldsymbol{x}}_{i-1}^j) + (\hat{\boldsymbol{x}}_{i-1}^j - \boldsymbol{x}^j)\|_2^2$$

$$\leq 2\|(\hat{\boldsymbol{x}}_i^j - \hat{\boldsymbol{x}}_{i-1}^j)\|_2^2 + 2\|\hat{\boldsymbol{x}}_{i-1}^j - \boldsymbol{x}^j\|_2^2 \tag{33}$$

The second term in Eq. (33) can be bounded by $\sum_{l=1}^{i-1}\frac{c^2\sigma^2}{\mu_l^2\mu_{l\rightarrow}^2}\|\boldsymbol{x}^j\|_2^2$ by induction hypothesis. Therefore, it is enough to show that the first term in Eq. (33) is bounded by $\frac{c^2\sigma^2}{\mu_i^2\mu_{i\rightarrow}^2}\|\boldsymbol{x}^j\|_2^2$. We decompose the error into two error terms one of which corresponds to the error propagation through the network if activation were fixed and the other one is the error caused by change in the activations:

$$\|(\hat{\boldsymbol{x}}_i^j - \hat{\boldsymbol{x}}_{i-1}^j)\| \tag{34}$$

$$= \|M^{i,j}((\boldsymbol{W}_i + \boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1})) - M^{i,j}(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1}))\| \tag{35}$$

$$= \| M^{i,j}((\boldsymbol{W}_i + \boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1})) - M^{i,j}(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1}))$$

$$+ J_{\boldsymbol{x}^i}^{i,j}(\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1})) - J_{\boldsymbol{x}^i}^{i,j}(\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1})) \| \tag{36}$$

$$\leq \| J_{\boldsymbol{x}^i}^{i,j}(\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1}))\| + \|M^{i,j}((\boldsymbol{W}_i + \boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1}))$$

$$- M^{i,j}(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1})) - J_{\boldsymbol{x}^i}^{i,j}(\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1})) \| . \tag{37}$$

The first term in Eq. (37) is bounded by:

$$\|J_{\boldsymbol{x}^i}^{i,j}\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1})\|_2 \tag{38}$$

$$\leq (\sqrt{d}\sigma/\sqrt{6\delta})\|J_{\boldsymbol{x}^i}^{i,j}\|_2\|\boldsymbol{W}_i\|_F\|\phi(\hat{\boldsymbol{x}}^{i-1})\|_2 \tag{39}$$

$$\leq (\sqrt{d}\sigma/\sqrt{6\delta})\|J_{\boldsymbol{x}^i}^{i,j}\|_2\|\boldsymbol{W}_i\|_F\|\hat{\boldsymbol{x}}^{i-1}\|_2 \tag{40}$$

$$\leq (\sqrt{d}\sigma/\sqrt{3\delta})\|J_{\boldsymbol{x}^i}^{i,j}\|_2\|\boldsymbol{W}_i\|_F\|\boldsymbol{x}^{i-1}\|_2 \tag{41}$$

$$\leq (c\sqrt{d}\sigma/\sqrt{3\delta})\|J^{i,j}_{\boldsymbol{x}^i}\|_2\|\boldsymbol{W}_i\|_F\|\phi(\boldsymbol{x}^{i-1})\|_2 \tag{42}$$

$$\leq (c\sqrt{d}\sigma/\sqrt{3\delta}\mu_i)\|J^{i,j}_{\boldsymbol{x}^i}\|_2\|\boldsymbol{W}_i\phi(\boldsymbol{x}^{i-1})\|_2 \tag{43}$$

$$\leq (c\sqrt{d}\sigma/\sqrt{3\delta}\mu_i\mu_{i\rightarrow})\|\boldsymbol{x}^j\|_2. \tag{44}$$

where Eq. (38) is bounded by Eq. (31), Eq. (39) is bounded by Lipschitzness of the activation function, Eq. (40) is bounded by inductive hypothesis, Eq. (41) is bounded by activation contraction, Eq. (42) is bounded by layer cushion, and Eq. (44) is bounded by interlayer cushion.

The second term in Eq. (37) can be bounded as:

$$\| M^{i,j}((\boldsymbol{W}_i+\boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1})) - M^{i,j}(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1})) \\ - J^{i,j}_{\boldsymbol{x}^i}(\boldsymbol{U}^i\phi(\hat{\boldsymbol{x}}^{i-1}))\|_2 \tag{45}$$

$$=\| (M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})((\boldsymbol{W}_i+\boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1})) \\ -(M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1}))\|_2 \tag{46}$$

$$=\|(M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})((\boldsymbol{W}_i+\boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1}))\|_2 \\ +\|(M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})(\boldsymbol{W}_i\phi(\hat{\boldsymbol{x}}^{i-1}))\|_2. \tag{47}$$

Both terms in Eq. (47) can be bounded using Assumption 3. By notations we find $\boldsymbol{W}^i\phi(\hat{\boldsymbol{x}}^{i-1}) = \hat{x}^i_{i-1}$. By induction hypothesis, we have that $\|\boldsymbol{W}^i\phi(\hat{\boldsymbol{x}}^{i-1}) - \boldsymbol{x}^i\|_2^2 \leq \sum_{l=1}^{i-1}\frac{c^2d\sigma^2}{\delta\mu_l^2\mu_{l\rightarrow}^2}\|\boldsymbol{x}^i\|_2^2$. Now by interlayer smoothness property, $\|(M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})(\boldsymbol{W}^i\phi(\hat{\boldsymbol{x}}^{i-1}))\|_2^2 \leq \frac{\sum_{l=1}^{i-1}\frac{c^2d\sigma^2}{\delta\mu_l^2\mu_{l\rightarrow}^2}\|\boldsymbol{x}^j\|}{\rho_\delta} \leq (\sum_{l=1}^{i-1}\frac{c^2d\sigma^2}{\delta\mu_l^2\mu_{l\rightarrow}^2})\|\boldsymbol{x}^j\|_2^2/(3d) \simeq \frac{i-1}{3d}\frac{c^2d\sigma^2}{\delta\mu_i^2\mu_{i\rightarrow}^2}\|\boldsymbol{x}^j\|_2^2$. Similar to this term, $\|(M^{i,j}-J^{i,j}_{\boldsymbol{x}^i})((\boldsymbol{W}^i+\boldsymbol{U}_i)\phi(\hat{\boldsymbol{x}}^{i-1}))\| \leq (\sum_{l=1}^{i}\frac{c^2d\sigma^2}{\delta\mu_l^2\mu_{l\rightarrow}^2})\|\boldsymbol{x}^j\|/(3d) \simeq \frac{i}{3d}\frac{c^2d\sigma^2}{\delta\mu_i^2\mu_{i\rightarrow}^2}\|\boldsymbol{x}^j\|_2^2$. Putting everything together completes the induction with probability at least $1-\delta$ (if $i=d$).

Instead of assuming that the input domain $\mathcal{X}$ is bounded by a constant $B$, we assume that the input boundary is relative to the expected value which implies the data-distribution information: $\max_{\boldsymbol{x}}\|\boldsymbol{x}^d\|_2 \leq \mathcal{O}(\mathbb{E}_{\mathcal{D}}\|\boldsymbol{x}^d\|_2)$. According to the margin contraction property, we can use the first- and second-statistics of the margin in the last layer $\mathbb{E}_{\mathcal{D}}[\gamma_h(\boldsymbol{x},y)] = r, \mathrm{Var}_{\mathcal{D}}[\gamma_h(\boldsymbol{x},y)] = \theta^2$ to bound the perturbation instead of the worst situation:

$$\max_{\boldsymbol{x}}\|\boldsymbol{x}^d\|_2^2 \leq \mathcal{O}(\mathbb{E}_{\mathcal{D}}\|\boldsymbol{x}^d\|_2^2) \leq \mathcal{O}(\alpha^2\mathbb{E}_{\mathcal{D}}\|\gamma_h(\boldsymbol{x},y)\|_2^2) \tag{48}$$

$$\leq \mathcal{O}(\alpha^2(r^2+\theta^2)) \leq \mathcal{O}(\alpha^2\,(r+\theta)^2) \tag{49}$$

Connecting these two inequalities we prove that the equality holds with a probability at least 1/2:

$$|f_{\boldsymbol{w}+\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2^2 \leq \mathcal{O}\left(\sum_{i=1}^{d}\frac{d\alpha^2c^2(r+\theta)^2\sigma^2}{\mu_i^2\mu_{i\rightarrow}^2}\right). \quad \square \tag{50}$$

### 5.2. Proof of Lemma 4

**Proof of Lemma 4.** Let $\boldsymbol{w}' = \boldsymbol{w}+\boldsymbol{u}$, Let $\mathcal{S}_{\boldsymbol{w}}$ be the set of perturbations with the following property:

$$\mathcal{S}_{\boldsymbol{w}} \subseteq \left\{\boldsymbol{w}' \,\middle|\, \max_{\boldsymbol{x}\in\mathcal{X}}|f_{\boldsymbol{w}'}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2 < \frac{r-\theta}{8\sqrt{\rho}}\right\}, \tag{51}$$

then we will have $\max_{\boldsymbol{x}\in\mathcal{X}}|f_{\boldsymbol{w}'}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_\infty < \sqrt{\rho}\max_{\boldsymbol{x}\in\mathcal{X}}|f_{\boldsymbol{w}'}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2 < \frac{r-\theta}{8}$.

Let $q$ be the probability density function over the parameters $\boldsymbol{w}'$. We construct a new distribution $\tilde{Q}$ over predictors $f_{\tilde{\boldsymbol{w}}}$ where $\tilde{\boldsymbol{w}}$ is restricted to $\mathcal{S}_{\boldsymbol{w}}$ with the probability density function:

$$\tilde{q}(\tilde{\boldsymbol{w}}) = \frac{1}{Z}\begin{cases} q(\tilde{\boldsymbol{w}}) & \tilde{\boldsymbol{w}}\in\mathcal{S}_{\boldsymbol{w}} \\ 0 & \text{otherwise}\end{cases} \tag{52}$$

According to the lemma assumption, we have $Z = \mathbb{P}\left[\boldsymbol{w}'\in\mathcal{S}_{\boldsymbol{w}}\right] \geq \frac{1}{2}$. Therefore, we can bound the change of the margins for any instance:

$$\max_{i,j\in[k],\boldsymbol{x}\in\mathcal{X}}|(|f_{\tilde{\boldsymbol{w}}}(\boldsymbol{x})[i] - f_{\tilde{\boldsymbol{w}}}(\boldsymbol{x})[j]|) - (|f_{\boldsymbol{w}}(\boldsymbol{x})[i] - f_{\boldsymbol{w}}(\boldsymbol{x})[j]|)| < \frac{r-\theta}{2} \tag{53}$$

Here we define a perturbed loss function as:

$$L'_{r,\theta}(h) = \Pr_{\mathcal{D}}\left[\gamma_h(\boldsymbol{x},y) \leq \frac{r-\theta}{2}\right] + \Pr_{\mathcal{D}}\left[\gamma_h(\boldsymbol{x},y) > r+\theta+\frac{r-\theta}{2}\right]. \tag{54}$$

We can get the following:

$$L_0\,(f_{\boldsymbol{w}}) \leq L'_{r,\theta}\,(f_{\tilde{\boldsymbol{w}}}) \tag{55}$$

$$\widehat{L}'_{r,\theta}\,(f_{\tilde{\boldsymbol{w}}}) \leq \widehat{L}_{r,\theta}\,(f_{\boldsymbol{w}}) \tag{56}$$

Finally, using the proof of Neyshabur et al. (2018, Lemma 1), with probability $1-\delta$ over the training set we have:

$$L_0\,(f_{\boldsymbol{w}}) \leq \mathbb{E}_{\tilde{\boldsymbol{w}}}\left[L'_{r,\theta}\,(f_{\tilde{\boldsymbol{w}}})\right] \tag{57}$$

$$\leq \mathbb{E}_{\tilde{\boldsymbol{w}}}\left[\widehat{L}'_{r,\theta}\,(f_{\tilde{\boldsymbol{w}}})\right] + 2\sqrt{\frac{2\left(D_{\mathrm{KL}}(\tilde{\boldsymbol{w}}\parallel P) + \ln\frac{2m}{\delta}\right)}{m-1}} \tag{58}$$

$$\leq \widehat{L}_{r,\theta}\,(f_{\boldsymbol{w}}) + 2\sqrt{\frac{2\left(D_{\mathrm{KL}}(\tilde{\boldsymbol{w}}\parallel P) + \ln\frac{2m}{\delta}\right)}{m-1}} \tag{59}$$

$$\leq \widehat{L}_{r,\theta}\,(f_{\boldsymbol{w}}) + 4\sqrt{\frac{D_{\mathrm{KL}}\,(\boldsymbol{w}'\parallel P) + \ln\frac{6m}{\delta}}{m-1}} \quad \square \tag{60}$$

### 5.3. Proof of Theorem 5

**Proof of Theorem 5.** Since Lemma 3 proves that the perturbation caused by random vector $\boldsymbol{u}$ is bounded by a term relative to the variance $\sigma$, we can preset the value of $\sigma$ to make the random perturbation satisfy the condition for Lemma 4. Bounding the Kullback–Leibler divergence term by $\|\boldsymbol{w}\|_2^2/\|\boldsymbol{u}\|_2^2$ in PAC-Bayesian theorem, we can attain the generalization bound based on a specific margin distribution.

The proof involves chiefly two steps. In the first step we bound the maximum value of perturbation of parameters to satisfy the condition that the change of output restricted by hyper-parameters of margin $r$ and $\theta$, using Lemma 3. In the second step we prove the final margin generalization bound through Lemma 4 with the value of Kullback–Leibler divergence term calculated based on the bound in the first step.

$$|f_{\boldsymbol{w}+\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{w}}(\boldsymbol{x})|_2^2 \leq \mathcal{O}\left(\sum_{i=1}^{d}\frac{d\alpha^2c^2(r+\theta)^2\sigma^2}{\mu_i^2\mu_{i\rightarrow}^2}\right) \\ = (\frac{r-\theta}{8\sqrt{\rho}})^2$$

We can derive $\sigma = \frac{r-\theta}{8\alpha cd\sqrt{\rho}(r+\theta)\sqrt{\sum_{i=1}^{d}\frac{1}{\mu_i^2\mu_{i\rightarrow}^2}}}$ from the above inequality. Naturally, we can calculate the Kullback–Leibler divergence in Lemma 3 with the chosen distributions for $P \sim \mathcal{N}(0,\sigma^2\mathbf{I})$.

$$D_{\mathrm{KL}}(\boldsymbol{w}+\boldsymbol{u}\parallel P) \leq \frac{|\boldsymbol{w}|^2}{2|\boldsymbol{w}|^2|\eta\eta^\top|^2} = \frac{1}{2\sigma^2} \tag{61}$$

$$\leq \mathcal{O}\left(\frac{(r+\theta)^2}{(r-\theta)^2}\sum_{i=1}^{d}\frac{d\rho\alpha^2c^2}{\mu_i^2\mu_{i\rightarrow}^2}\right) \tag{62}$$

Put it in Lemma 4 and let $\lambda = \theta/r$, with probability at least $1-\delta$ and for all $\boldsymbol{w}$ such that, we have:

$$L_0(h) \le \widehat{L}_{r,\theta}(h) + \mathcal{O}\left(\sqrt{\frac{\frac{(1+\lambda)^2}{(1-\lambda)^2}\sum_{i=1}^{d}\frac{d\rho\alpha^2c^2}{\mu_i^2\mu_{i\to}^2} + \ln\frac{dm}{\delta}}{m}}\right). \quad \square \qquad (63)$$

## 6. Optimizing margin distribution measure

The generalization theory shows the importance of optimizing the margin distribution ratio $\lambda$. The result inspires us to find a margin distribution band $(r-\theta \le \gamma_h(\boldsymbol{x}, y) < r+\theta)$ containing as many training samples as possible to minimize the empirical estimate loss $\widehat{L}_{r,\theta}$, but also a ratio $\lambda = \theta/r$ as small as possible to minimize the generalization gap $L_0(h) - \widehat{L}_{r,\theta}(h)$. This type of loss function was first proposed by Zhang and Zhou (2019) to optimize the first- and second-order statistics of margin distribution. We formulate a convex margin distribution loss function for DNNs:

**Definition 2** (*Convex Margin Distribution Loss Function*)**.** For a labeled sample $(\boldsymbol{x}, y) \in \mathcal{D}$, we denote its margin by $\gamma_h$ which is defined as Eq. (1). We define the margin distribution loss for networks (mdNet loss) as:

$$\ell_{r,\theta,\eta}(h(\boldsymbol{x}), y) = \begin{cases} \frac{(r-\theta-\gamma_h)^2}{(r-\theta)^2} & \gamma_h \le r-\theta \\ 0 & r-\theta < \gamma_h \le r+\theta \\ \frac{\eta(r+\theta-\gamma_h)^2}{(r+\theta)^2} & \gamma_h > r+\theta, \end{cases} \qquad (64)$$

where $r$ is the margin mean parameter, $\theta$ is the margin variance parameter and $\eta$ is a parameter to trade off two different kinds of deviation (keeping the balance on both sides of the margin mean). Fig. 1(c) shows the shape of this convex loss function.

Eq. (64) will produce a square loss when the margin satisfies $\gamma_h \le r-\theta$ or $\gamma_h \ge r+\theta$. Therefore, our margin loss function will force the zero-loss band to contain as many sample points as possible. The ratio of hyper-parameters $\lambda = \theta/r$ can control the capacity measure, which implies our measure is dependent to our specific learning algorithm (loss function with specific hyper-parameters). Since our loss function aims at finding a decision boundary which is determined by the entire margin distribution, instead of the minority samples that have minimum margins, we name our method as margin distribution Networks (mdNet).

## 7. Experiments

In Section 7.1, we introduce the configuration of datasets and models. In Section 7.2, we design an ablation experiment to verify the superiority of our method. In Section 7.3, we show the correlation between separability of representations and margin ratio via visualization. In Section 7.4, we design experiments to confirm that our method can control the capacity of deep nets. In Section 7.5, we discuss the influence of the different hyper-parameters on the test accuracy.

### 7.1. Configuration

Since our method only works on the loss function part of deep models and does not change the architecture of deep neural networks, we can verify the effectiveness of mdNet on the classic CNNs (convolutional neural networks) and image classification benchmark datasets. We consider the following architectures and datasets: a LeNet architecture for MNIST dataset (LeCun, Bottou, Bengio, & Haffner, 1998), an AlexNet architecture (Krizhevsky, Sutskever, & Hinton, 2012) for CIFAR-10 dataset (Krizhevsky, 2009) and a ResNet-18 architecture (He, Zhang, Ren, & Sun, 2016) for ImageNet dataset (Russakovsky et al., 2015). From the literature, these datasets come pre-divided into training and testing sets, therefore in our experiments, we use them in their original format. The loss functions used for comparison in the experiments are as follows: cross-entropy loss (abbr., xent), hinge loss and soft hinge loss. Hinge loss (Cortes & Vapnik, 1995) and soft hinge loss (Liu et al., 2016) are loss functions specially proposed to optimize the minimum margin, both of them are inspired the traditional margin theory.

As for details about the architecture, we remove the weight decay (Krogh & Hertz, 1992), dropout (Srivastava et al., 2014) and batch normalization (BN) (Ioffe & Szegedy, 2015) from all the models, because the batch normalization operation and weight decay will shift the data distribution. The notable dropout technique, in which some neurons are dropped from the DNNs in each iteration, can also be viewed as an ensemble method composed of different neural networks, with different dropped neurons (Baldi & Sadowski, 2013). It is hard to analyze the influence of the ensemble structure on the margin distribution, so we remove this technique in these architectures in the experiments except to understand the contribution of the components to the whole models in the ablation study.

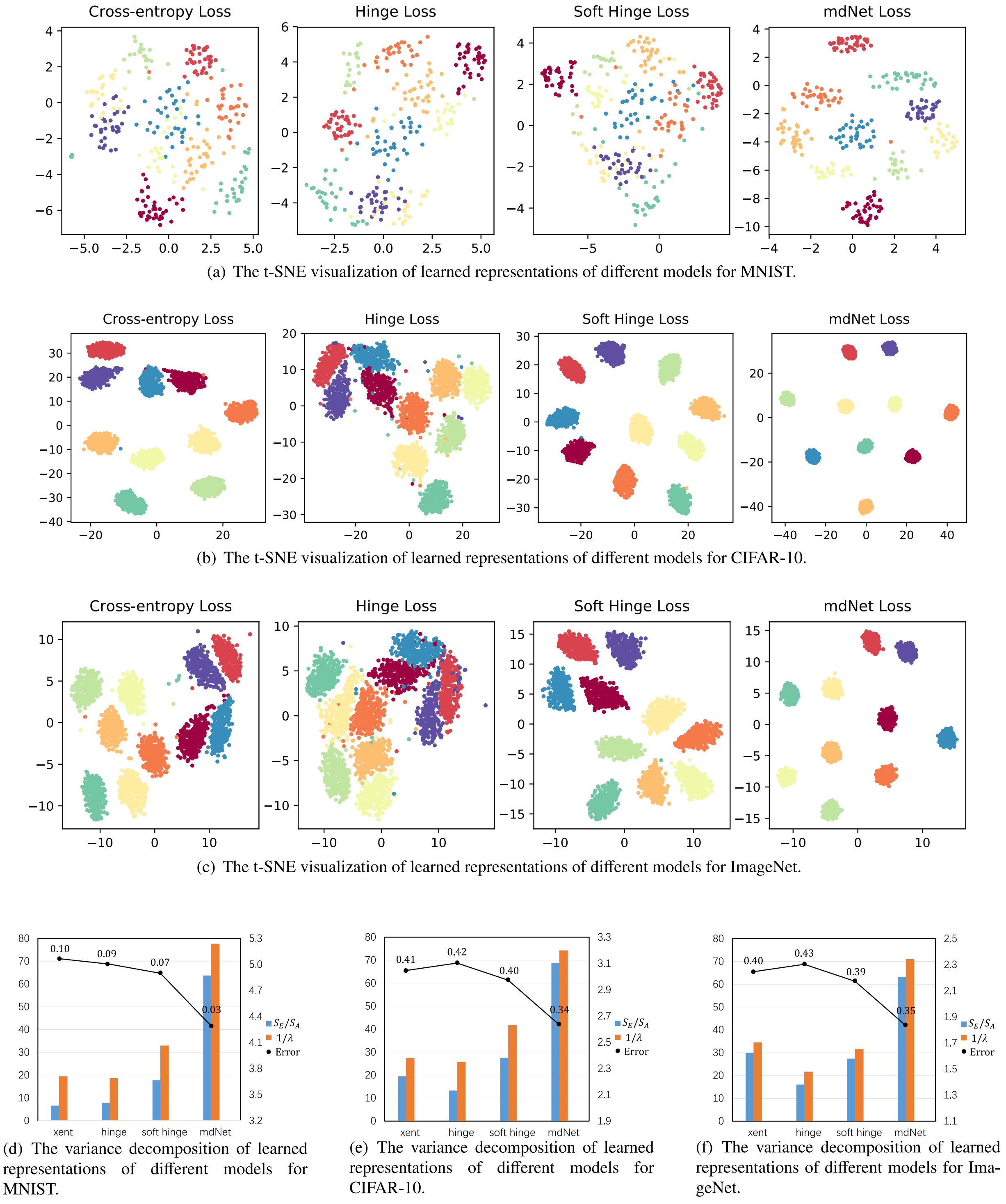
For *special hyper-parameters*, including the expected margin parameter and margin variance parameter for mdNet loss model, and margin parameter for hinge loss model, we perform hyper-parameter search. We hold out 5000 samples of the training set as a validation set, and use the remaining samples to train models with different special hyper-parameters values on all datasets. As for the common hyper-parameters, such as learning rate and momentum, we set them as the default commonly used values in PyTorch (Paszke et al., 2019) for all the models. We chose batch stochastic gradient descent as the optimizer. We run all the experiments on four K80 GPU machines. As for the influence of the different hyper-parameters on the test accuracy, we discuss it empirically in Section 7.5.

### 7.2. Ablation study

Since the optimization algorithm proposed in this paper only focuses on the improvement of loss function, we design an ablation experiment to study the performance of our proposed mdnet method and the traditional benchmark loss functions under different regularizations in Table 1. The mdNet loss outperforms the others consistently across different situations, no matter whether dropout, batch normalization or the entire dataset are used or not. The experiments are evaluated on three MNIST (LeNet), CIFAR-10 (AlexNet) and ImageNet (ResNet-18) datasets. Specifically, when the amount of training samples is small, the advantage of mdNet loss is significant. Moreover, the mdNet loss function can cooperate with both batch normalization and dropout, achieving the best performance in Table 1, which is highlighted in bold red text. Unlike dropout and batch normalization which lack solid theoretical grounds, the mdNet loss function is inspired by the margin distribution bound in Theorem 5, which guides us to find a suitable margin ratio to restrict the model capacity and alleviate the overfitting problem efficiently.

### 7.3. Feature visualization

In this experiment, we use t-SNE method to visualize the learned representations on the last hidden layer. Figs. 4(a), 4(b) and 4(c) plot the 2D t-SNE (van der Maaten & Hinton, 2008) embedding image on limited datasets, including MNIST (LeNet), CIFAR-10 (AlexNet) and 10-class ImageNet (ResNet-18). Consistently, we can find that the result of mdNet loss model is better



(a) The t-SNE visualization of learned representations of different models for MNIST.

(b) The t-SNE visualization of learned representations of different models for CIFAR-10.

(c) The t-SNE visualization of learned representations of different models for ImageNet.

(d) The variance decomposition of learned representations of different models for MNIST.

(e) The variance decomposition of learned representations of different models for CIFAR-10.

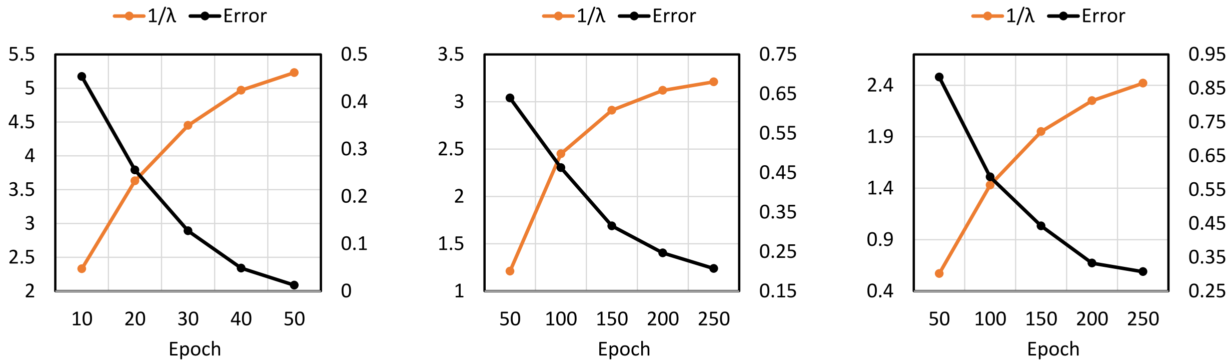
(f) The variance decomposition of learned representations of different models for ImageNet.

**Fig. 4.** The quality of feature representations generated by different models on the MNIST, CIFAR-10 and ImageNet datasets.

than all the others, the distribution of samples which have the same label are more compact. To quantify the degree of separability of data distribution, we perform a variance decomposition on the data in the embedding space. By comparing the ratio of inter-class variance $S_E$ to intra-class variance $S_A$ in Figs. 4(d), 4(e) and 4(f), we see that the mdNet loss always attain the most separable distribution among these four loss functions. Moreover, the visualization result is consistent with the margin distribution ratio $1/\lambda$ of these four models, which means that optimizing the margin distribution (searching an appropriate margin ratio $\lambda$) is helpful to attain a good learned representation space. This representation features space can further alleviate the overfitting problem of deep learning, we verify empirically that a network trained with mdNet loss shows stronger clustering. Specially, Figs. 4(d), 4(e) and 4(f) show the relationship between the margin ratio and test error. Moreover, Fig. 5 plots the test error of mdNet and the margin ratio across the different epochs. We can see

**Table 1**
Test accuracy of LeNet on MNIST, AlexNet on CIFAR-10 and ResNet-18 on ImageNet datasets with different regularization methods and different fractions of training set. The best accuracy on each training dataset is highlighted in bold red type. The bold black text indicates the better accuracy between the two losses with the same regularization.

| | | | | |
|---|---|---|---|---|
| MNSIT | | | | |
| Accuracy(%) | xent | Hinge | Soft hinge | mdNet |
| | Batch normalization | | | |
| 100%_Dropout | 99.095 ± 0.083 | 98.593 ± 0.164 | 99.148 ± 0.039 | **99.161 ± 0.073** |
| 100%_Non_Dropout | 98.384 ± 0.072 | 97.571 ± 0.178 | 98.475 ± 0.064 | **98.837 ± 0.091** |
| 5%_Dropout | 97.001 ± 0.131 | 96.527 ± 0.219 | 97.112 ± 0.092 | **97.268 ± 0.113** |
| 5%_Non_Dropout | 83.364 ± 0.452 | 83.292 ± 0.721 | 83.749 ± 0.273 | **84.483 ± 0.348** |
| | Non batch normalization | | | |
| 100%_Dropout | 98.228 ± 0.079 | 98.029 ± 0.184 | 98.271 ± 0.055 | **98.342 ± 0.069** |
| 100%_Non_Dropout | 91.728 ± 0.117 | 90.237 ± 0.318 | 91.029 ± 0.098 | **92.274 ± 0.121** |
| 5%_Dropout | 77.842 ± 0.489 | 76.938 ± 0.827 | 77.727 ± 0.411 | **78.173 ± 0.619** |
| 5%_Non_Dropout | 58.023 ± 0.951 | 57.822 ± 1.280 | 59.384 ± 0.827 | **61.379 ± 0.588** |
| CIFAR-10 | | | | |
| Accuracy(%) | xent | Hinge | Soft hinge | mdNet |
| | Batch normalization | | | |
| 100%_Dropout | 85.782 ± 0.198 | 84.234 ± 0.748 | 86.744 ± 0.294 | **87.644 ± 0.151** |
| 100%_Non_Dropout | 81.491 ± 0.143 | 80.938 ± 0.812 | 86.032 ± 0.298 | **86.233 ± 0.244** |
| 5%_Dropout | 61.955 ± 1.945 | 58.363 ± 2.450 | 59.441 ± 1.316 | **67.636 ± 1.633** |
| 5%_Non_Dropout | 57.753 ± 2.228 | 54.289 ± 3.482 | 56.839 ± 2.318 | **64.173 ± 1.982** |
| | Non batch normalization | | | |
| 100%_Dropout | 83.517 ± 0.322 | 82.153 ± 1.236 | 81.961 ± 0.293 | **84.643 ± 0.255** |
| 100%_Non_Dropout | 72.223 ± 1.284 | 69.379 ± 2.907 | 75.267 ± 1.027 | **76.793 ± 1.279** |
| 5%_Dropout | 50.747 ± 3.735 | 42.739 ± 6.763 | 52.847 ± 1.823 | **58.739 ± 1.348** |
| 5%_Non_Dropout | 36.293 ± 4.872 | 30.984 ± 7.736 | 43.265 ± 4.263 | **47.056 ± 3.927** |
| ImageNet | | | | |
| Accuracy(%) | xent | Hinge | Soft hinge | mdNet |
| | Batch normalization | | | |
| 100%_Dropout | 70.238 ± 1.221 | 69.782 ± 1.933 | 70.284 ± 1.022 | **70.758 ± 1.014** |
| 100%_Non_Dropout | 68.484 ± 1.265 | 67.918 ± 2.166 | 68.83 ± 1.151 | **69.447 ± 1.124** |
| 5‰_Dropout | 60.176 ± 2.045 | 57.475 ± 2.023 | 61.379 ± 1.053 | **65.080 ± 2.373** |
| 5‰_Non_Dropout | 59.574 ± 2.747 | 56.621 ± 2.253 | 60.068 ± 1.773 | **63.529 ± 2.012** |
| | Non batch normalization | | | |
| 100%_Dropout | 66.924 ± 1.552 | 67.387 ± 1.764 | 67.462 ± 1.017 | **68.655 ± 1.732** |
| 100%_Non_Dropout | 64.481 ± 2.183 | 61.820 ± 2.947 | 64.334 ± 2.367 | **65.838 ± 2.481** |
| 5‰_Dropout | 54.961 ± 3.382 | 52.543 ± 3.722 | 55.757 ± 2.357 | **58.774 ± 3.841** |
| 5‰_Non_Dropout | 47.374 ± 3.265 | 45.741 ± 5.349 | 48.798 ± 3.392 | **53.727 ± 4.235** |



(a) LeNet for MNIST. (b) AlexNet for CIFAR-10. (c) ResNet-18 for ImageNet.

**Fig. 5.** Test error and margin ratio across epochs on mdNet models for MNIST, CIFAR-10 and ImageNet datasets.

that more compact margin distribution gets better prediction performance across different models and epochs. This exhibits that optimizing margin distribution can indeed improve the learning ability of deep nets.

### 7.4. Controlling capacity

The first two experiments have demonstrated that our mdNet can outperform other classical loss functions and our method can learn a more separable feature representation, as the corresponding margin ratio is also smaller. However, it leads to the last question:

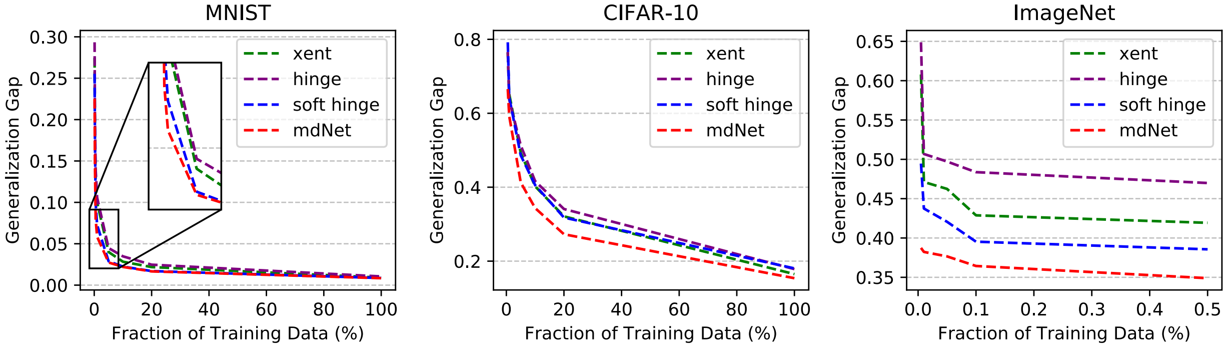
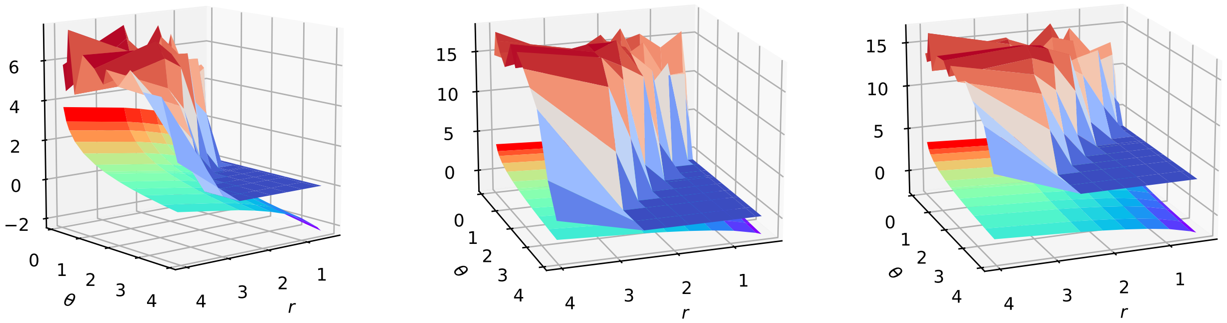


**Fig. 6.** Compare the convergence rate of generalization gap with the increase of training samples under different loss functions on MNIST, CIFAR-10 and ImageNet datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) LeNet for MNIST. (b) AlexNet for CIFAR-10. (c) ResNet-18 for ImageNet.

**Fig. 7.** The test accuracy varying with two hyper-parameter $r$ and $\theta$ on MNIST, CIFAR-10 and ImageNet datasets. The logarithm of ratio $\ln(1/\lambda) = \ln(r/\theta)$ is the lower surface with rainbow colors and the test accuracy is the upper surface with warm-cool colors. The test accuracy is rescaled to be within the same range as the ratio. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

*Can smaller margin ratio reduce the capacity of models and accelerate the convergence of generalization gap?*

Let us go back to the theoretical result obtained by Theorem 5. The generalization gap of the model is bounded by $\Lambda_{\lambda,\boldsymbol{w}}/\sqrt{m}$, where the margin distribution measure $\Lambda_{\lambda,\boldsymbol{w}} \propto \frac{1+\lambda}{1-\lambda}$ determines the worse case of generalization gap when the number of samples are equal:

$$\text{Generalization Gap} \le \mathcal{O}(\frac{\Lambda_{\lambda,\boldsymbol{w}}}{\sqrt{m}}). \tag{65}$$

Therefore, we design experiments to compare the empirical value of generalization gap with the increase of training samples under different loss functions. In Fig. 6, it shows that the red dotted line representing the convergence curve of our method always converges faster than the other lines under different datasets and models. It also demonstrates that our method can effectively control the capacity of the model by optimizing the margin distribution ratio, so that the trained model has better generalization performance.

Given a fixed number of samples $m$, we find that the worst case of generalization gap is proportional to the model capacity. When $m$ is large enough, the scale factor $1/\sqrt{m}$ will be close to 0, and the difference of sample complexity is not significant. The advantage of optimizing margin distribution is relatively significant when $\sqrt{m}$ is relatively small. Therefore, in the right of Fig. 6 (the ImageNet experiment), we specially truncate the most significant result of convergence rate (form 0.1‰ to 5‰ of training set), which shows that optimizing margin distribution can control the capacity of the model even on such a complex dataset.

### 7.5. Influence of the hyper-parameters

Fig. 7 plots the 3D surface figure for the test accuracy on the MNIST, CIFAR-10 and ImageNet datasets varying with two hyper-parameter $r$ and $\theta$. It shows that the ratio $1/\lambda = r/\theta$ (the lower surface with rainbow colors) increases with $r$ increasing and $\theta$ decreasing. As for the test accuracy (the upper surface with warm-cool colors), we find that its trend is consistent with the ratio $1/\lambda$. Therefore, the influence of the hyper-parameters demonstrates that our theoretical result. Within a certain range, getting a smaller ratio $\lambda$ through specific optimization (the margin distribution loss function) will effectively reduce the size of the hypothesis set for deep nets (returned by the specific algorithm), so as to improve the generalization ability of the learned model. In other words, the test accuracy changes consistently with the ratio of hyper-parameters (an estimation of the ratio of margin distribution). The parameter $\eta$ to trade off two different kinds of deviation (keep the balance on both sides of the margin mean) is always fixed to 0.1 in practice.

## 8. Conclusion

This paper proves generalization bound for deep neural networks by considering the margin distribution at the last layer instead of the minimum margin. The theoretical result inspires us to utilize a margin distribution loss function to improve the generalization performance of neural networks. Experimental results show that our method can effectively control the model capacity by optimizing the margin distribution measure, so that the trained model learns more separable representations and has better generalization performance. In future work, we will explore the effectiveness of regularization methods from a *margin theory* perspective.

## Acknowledgement

This research was supported by the National Science Foundation of China (61921006) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. The authors would like to thank the anonymous reviewers for constructive suggestions, as well as Yi-Xiao He and Yi-He Chen for helpful discussions.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Arora, Sanjeev, Ge, Rong, Neyshabur, Behnam, & Zhang, Yi (2018). Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th international conference on machine learning* (pp. 254–263).

Azulay, Aharon, & Weiss, Yair (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, *20*(184), 1–25.

Baldi, Pierre, & Sadowski, Peter J. (2013). Understanding dropout. In *Advances in neural information processing systems, Vol. 27* (pp. 2814–2822).

Bartlett, Peter L., Foster, Dylan J., & Telgarsky, Matus J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in neural information processing systems, Vol. 31* (pp. 6241–6250).

Bartlett, Peter L., Maiorov, Vitaly, & Meir, Ron (1998). Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, *10*(8), 2159–2173.

Becker, Carlos J., Christoudias, C. Mario, & Fua, Pascal (2013). Non-linear domain adaptation with boosting. In *Advances in neural information processing systems, Vol. 26* (pp. 485–493).

Breiman, Leo (1999). Prediction games and arcing algorithms. In *Neural Computation Neural Computation*, *11*(7), 1493–1517.

Chan, Tsung-Han, Jia, Kui, Gao, Shenghua, Lu, Jiwen, Zeng, Zinan, & Ma, Yi (2015). PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, *24*(12), 5017–5032.

Chen, Yuansi, Jin, Chi, & Yu, Bin (2018). Stability and convergence trade-off of iterative optimization algorithms. CoRR, abs/1804.01619.

Cortes, Corinna, & Vapnik, Vladimir (1995). Support-vector networks. In *Machine Learning Machine Learning*, *20*(3), 273–297.

Dinh, Laurent, Pascanu, Razvan, Bengio, Samy, & Bengio, Yoshua (2017). Sharp minima can generalize for deep nets. In *Proceedings of the 34th international conference on machine learning, Vol. 70* (pp. 1019–1028).

Dubey, Abhimanyu, Ramanathan, Vignesh, Pentland, Alex, & Mahajan, Dhruv (2021). Adaptive methods for real-world domain generalization. In *IEEE conference on computer vision and pattern recognition* (pp. 14340–14349).

Gao, Wei, & Zhou, Zhi-Hua (2013). On the doubt about margin explanation of boosting. *Artificial Intelligence*, *203*, 1–18.

Ghifary, Muhammad, Kleijn, W. Bastiaan, Zhang, Mengjie, & Balduzzi, David (2015). Domain generalization for object recognition with multi-task autoencoders. In *IEEE international conference on computer vision* (pp. 2551–2559).

Glantz, Stanton, & Slinker, Bryan (2001). *Primer of applied regression & analysis of variance, Ed.* New York: McGraw-Hill, Inc..

Goodfellow, Ian J., Shlens, Jonathon, & Szegedy, Christian (2015). Explaining and harnessing adversarial examples. In *3rd international conference on learning representations*.

Gori, Marco (2022). Ten questions for a theory of vision. *Frontiers in Computer Science*, *3*, 701248. http://dx.doi.org/10.3389/fcomp.2021.701248.

Grønlund, Allan, Kamma, Lior, Larsen, Kasper Green, Mathiasen, Alexander, & Nelson, Jelani (2019). Margin-Based Generalization Lower Bounds for Boosted Classifiers. In *Advances in Neural Information Processing Systems 32* (pp. 11940—11949).

Gunasekar, Suriya, Lee, Jason D., Soudry, Daniel, & Srebro, Nathan (2018a). Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th international conference on machine learning, Vol. 80* (pp. 1827–1836).

Gunasekar, Suriya, Lee, Jason D., Soudry, Daniel, & Srebro, Nati (2018b). Implicit bias of gradient descent on linear convolutional networks. In *Advances in neural information processing systems, Vol. 31* (pp. 9482–9491).

Hardt, Moritz, Recht, Ben, & Singer, Yoram (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd international conference on machine learning, Vol. 48* (pp. 1225–1234).

Harvey, Nick, Liaw, Christopher, & Mehrabian, Abbas (2017). Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 30th annual conference on learning theory* (pp. 1064–1068).

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

Ioffe, Sergey, & Szegedy, Christian (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd international conference on machine learning* (pp. 448–456).

Ji, Ziwei, & Telgarsky, Matus (2019). Gradient descent aligns the layers of deep linear networks. In *7th international conference on learning representations*.

Jiang, Yiding, Krishnan, Dilip, Mobahi, Hossein, & Bengio, Samy (2019). Predicting the generalization gap in deep networks with margin distributions. In *7th international conference on learning representations*.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, & Tang, Ping Tak Peter (2017). On large-batch training for deep learning: Generalization gap and sharp minima. In *5th international conference on learning representations*.

Koniusz, Piotr, Tas, Yusuf, & Porikli, Fatih (2017). Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In *IEEE conference on computer vision and pattern recognition* (pp. 7139–7148).

Krizhevsky, Alex (2009). *Learning multiple layers of features from tiny images*: *Technical report*.

Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems, Vol. 26* (pp. 1097–1105).

Krogh, Anders, & Hertz, John A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems, Vol. 5* (pp. 950–957).

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey (2015). Deep learning. *Nature*, *521*(7553), 436.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, & Haffner, Patrick (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Yuanzhi, Ma, Tengyu, & Zhang, Hongyang (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on learning theory, Vol. 75* (pp. 2–47).

Liu, Weiyang, Wen, Yandong, Yu, Zhiding, & Yang, Meng (2016). Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd international conference on machine learning* (pp. 507–516).

Mansour, Yishay, Mohri, Mehryar, & Rostamizadeh, Afshin (2009). Domain adaptation: Learning bounds and algorithms. In *The 22nd conference on learning theory*.

Mansour, Yishay, & Schain, Mariano (2014). Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, *71*(4), 365–380.

Matskevych, Alex, Wolny, Adrian, Pape, Constantin, & Kreshuk, Anna (2022). From shallow to deep: exploiting feature-based classifiers for domain adaptation in semantic segmentation. *Frontiers in Computer Science*, *4*, 805166. http://dx.doi.org/10.3389/fcomp.2022.805166.

Mohri, Mehryar, Rostamizadeh, Afshin, & Talwalkar, Ameet (2018). *Foundations of machine learning*. MIT Press.

Mou, Wenlong, Wang, Liwei, Zhai, Xiyu, & Zheng, Kai (2018). Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In *Conference on learning theory, Vol. 75* (pp. 605–638).

Neyshabur, Behnam, Bhojanapalli, Srinadh, McAllester, David, & Srebro, Nati (2017). Exploring generalization in deep learning. In *Advances in neural information processing systems, Vol. 30* (pp. 5947–5956).

Neyshabur, Behnam, Bhojanapalli, Srinadh, & Srebro, Nathan (2018). A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *6th international conference on learning representations*.

Neyshabur, Behnam, Tomioka, Ryota, & Srebro, Nathan (2015). Norm-based capacity control in neural networks. In *Proceedings of the 28th Annual Conference on Learning Theory* (pp. 1376–1401).

Pan, Sinno Jialin, Tsang, Ivor W., Kwok, James T., & Yang, Qiang (2009). Domain adaptation via transfer component analysis. In *Proceedings of the 21st international joint conference on artificial intelligence* (pp. 1187–1192).

Papernot, Nicolas, McDaniel, Patrick D., Goodfellow, Ian J., Jha, Somesh, Celik, Z. Berkay, & Swami, Ananthram (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on asia conference on computer and communications security* (pp. 506–519).

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems, Vol. 32* (pp. 8024–8035).

Reyzin, Lev, & Schapire, Robert E. (2006). How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd international conference on machine learning* (pp. 753–760).

Rozantsev, Artem, Salzmann, Mathieu, & Fua, Pascal (2019). Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(4), 801–814.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schapire, Robert E., Freund, Yoav, Barlett, Peter, & Lee, Wee Sun (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the 14th international conference on machine learning* (pp. 322–330).

Schroff, Florian, Kalenichenko, Dmitry, & Philbin, James (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 815–823).

Soudry, Daniel, Hoffer, Elad, Nacson, Mor Shpigel, Gunasekar, Suriya, & Srebro, Nathan (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, *19*(70), 1–57.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, & Salakhutdinov, Ruslan (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tan, Zhi-Hao, Tan, Peng, Jiang, Yuan, & Zhou, Zhi-Hua (2020). Multi-label optimal margin distribution machine. *Machine Learning*, *109*(3), 623–642.

van der Maaten, Laurens, & Hinton, Geoffrey (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Wang, Jindong, Lan, Cuiling, Liu, Chang, Ouyang, Yidong, & Qin, Tao (2021). Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the 30th international joint conference on artificial intelligence* (pp. 4627–4635).

Wei, Colin, Lee, Jason D., Liu, Qiang, & Ma, Tengyu (2018). On the margin theory of feedforward neural networks. CoRR, abs/1810.05369.

Wu, Wei, Jing, Xiaoyuan, Du, Wencai, & Chen, Guoliang (2021). Learning dynamics of gradient descent optimization in deep neural networks. *Science China Information Science*, *64*(5), 15102. http://dx.doi.org/10.1007/s11432-020-3163-0.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, & Vinyals, Oriol (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Zhang, Chao, Zhang, Lei, & Ye, Jieping (2012). Generalization bounds for domain adaptation. *Advances in Neural Information Processing Systems*, *25*(4), 3320.

Zhang, Teng, Zhao, Peng, & Jin, Hai (2020). Optimal margin distribution learning in dynamic environments. In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 6821–6828).

Zhang, Teng, & Zhou, Zhi-Hua (2017). Multi-class optimal margin distribution machine. In *Proceedings of the 34th international conference on machine learning* (pp. 4063–4071).

Zhang, Teng, & Zhou, Zhi-Hua (2018a). Semi-supervised optimal margin distribution machines. In *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 3104–3110).

Zhang, Teng, & Zhou, Zhi-Hua (2018b). Optimal margin distribution clustering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (pp. 4474–4481).

Zhang, Teng, & Zhou, Zhi-Hua (2019). Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, *32*(6), 1143–1156.

Zhou, Zhi-Hua (2014). Large Margin Distribution Learning. In *Artificial Neural Networks in Pattern Recognition* (pp. 1–11).

Zhou, Zhi-Hua (2021). Why over-parameterization of deep neural networks does not overfit?. *Science China Information Sciences*, *64*(1), 1–3.

Zhu, Zhanxing, Wu, Jingfeng, Yu, Bing, Wu, Lei, & Ma, Jinwen (2019). The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of the 36th international conference on machine learning, Vol. 97* (pp. 7654–7663).