

---

# On the Decision Boundary of Deep Neural Networks

---

**Yu Li**

KAUST  
CEMSE

[yu.li@kaust.edu.sa](mailto:yu.li@kaust.edu.sa)

**Lizhong Ding**

KAUST  
CEMSE

[lizhong.ding@kaust.edu.sa](mailto:lizhong.ding@kaust.edu.sa)

**Xin Gao**

KAUST  
CEMSE

[xin.gao@kaust.edu.sa](mailto:xin.gao@kaust.edu.sa)

## Abstract

While deep learning models and techniques have achieved great empirical success, our understanding of the source of success in many aspects remains very limited. In an attempt to bridge the gap, we investigate the decision boundary of a production deep learning architecture with weak assumptions on both the training data and the model. We demonstrate, both theoretically and empirically, that the last weight layer of a neural network converges to a linear SVM trained on the output of the last hidden layer, for both the binary case and the multi-class case with the commonly used cross-entropy loss. Furthermore, we show empirically that training a neural network as a whole, instead of only fine-tuning the last weight layer, may result in better bias constant for the last weight layer, which is important for generalization. In addition to facilitating the understanding of deep learning, our result can be helpful for solving a broad range of practical problems of deep learning, such as catastrophic forgetting and adversarial attacking.

## 1 Introduction

In recent years, deep learning has achieved impressive success in various fields [17]. Not only has it boosted the performance of the state-of-the-art methods in various areas, such as computer vision [16] and natural language processing [8], it has also enabled machines to achieve human level intelligence in specific tasks [25]. Despite its great empirical success, deep learning is often criticized for being used as a black box [5], which refers to the well-known gap between its empirical power and the theoretical understanding of it [24].

As suggested by [24], a satisfactory theoretical understanding of deep learning should cover three aspects: 1) *representation power*, 2) *optimization characteristics*, and 3) *generalization property*. The representation power of deep learning has been extensively and rigorously discussed in [29]. In terms of the second aspect, that is, the convergence analysis of stochastic gradient descent (SGD) and the property of the minima obtained, numerous recent studies have ended promising answers [12, 21, 7, 26, 30, 20, 3, 4]. For example, [12] proves the conjecture of [2], extending the result to deep nonlinear neural networks and showing the nonexistence of poor local minima. [22] also shows that all local minima are globally optimal, given reasonable assumptions. [3, 4] prove the convergence of SGD given assumptions of the input distribution.

As for the generalization mystery, the studies are still in the early stage. Through systematic experiments, [29] suggests that although the explicit regularization, such as weight decay and dropout, may be helpful, the implicit regularization of SGD may be the key for generalization. Following that direction, [4] provides the generalization guarantee for over-parameterized networks on linearly separable data, which are trained by SGD. [27] shows that, for linearly separable data, gradient descent (GD) on an unregularized logistic regression problem results in the max-margin (hard margin SVM) solution. On the other hand, [1, 13] try to demystify the generalization property via deriving the generalization bounds.

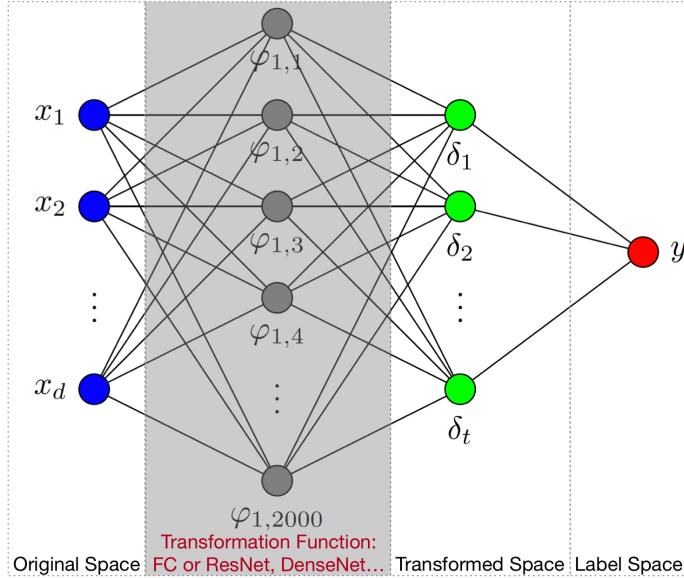


Figure 1: Network architecture. We do not oversimplify the network, with the only assumption being that the last hidden layer and the output layer are fully connected, which is practical and the common case. The transformation function can be any kind of deep learning architecture, including the legend fully connected layers or the commonly used ResNet or DenseNet *et al.* Here we show the fully connected layer for simplicity.

In this paper, we follow the direction of [29, 4, 27], investigating the implicit bias of GD and SGD. Unlike the previous studies, we do not oversimplify the model architecture. In fact, the architecture, which is shown in Fig. 1, is a productive one, which can reach the state-of-the-art performance on CIFAR-10 if we use DenseNet [11] as the transformation function. Moreover, we have little requirement for the input data distribution, only assuming that the loss converges to zero. In the Main Result section and Experiments section, we show that the direction of the neural network’s last weight layer converges to that of the SVM solution trained on the transformed data in the transformed space both theoretically and empirically. In addition, we also show that the decision boundary of the last layer is closer to the SVM decision boundary if we train the whole network, instead of only fine-tuning the last layer, in the Experiments part. We extend our result to multi-class classification problem with cross-entropy loss, which is the most common scenario in practice, on the MNIST and CIFAR10. Our study bridges the gap between the purely theoretical side, which investigates the over-simplified models and has strict requirements for the input distribution, and the practical usage of complex deep learning models. In practice, people usually owe the superior performance of deep learning to the model’s ability of learning representation and classifier simultaneously. We demystify the relationship between the learned representation and the classifier, and characterize the learned classifier in particular.

## 2 Problem Formulation

Unlike the setting of previous studies [27, 4, 3], which assume the training data is linearly separable or follows a certain distribution, we do not have such a requirement. Formally, for binary classification, we consider a dataset  $\{\mathbf{x}_n, y_n\}_{n=1}^N$ , with  $\mathbf{x}_n \in \mathbb{R}^d$ , and binary labels  $y_n \in \{-1, 1\}$ . We use  $\mathbf{X} \in \mathbb{R}^{d \times N}$  to denote the data matrix. For multi-class classification, we have  $y_n \in [K] := \{1, 2, \dots, K\}$  and  $K$  is the number of classes.

Regarding the neural network model, we do not restrict to any specific the architecture neither. Consider a neural network with the architecture shown in Fig. 1, which is basically a production network with practical usage. We divide the neural network into four components. The original space and label space are the training interface. The transformation function combined with the transformed space (the output of the last hidden layer) is one of the reasons why the deep learning’s performance is being continuously improved. For the sake of analysis, we take the transformed space

as an independent component which is fully connected with the label space. Formally, we denote the output of the last hidden layer on example  $\mathbf{x}_n$  as  $\delta_n$ , with  $\delta_n \in \mathbb{R}^t$ .

We denote the entire parameter set of the network as  $\theta$ . The network defines a function  $f(\mathbf{x}; \theta) : \mathbb{R}^d \rightarrow \{-1, 1\}$  for the binary case. The transformation function is  $\delta_n = h(\mathbf{x}_n; \phi)$ , where  $\phi$  is the parameter set of the transformation function. Notice that from  $\delta_n$  to the final output, the last weight layer defines a linear transformation, which has the following form:

$$g(\delta_n; \mathbf{W}) = \mathbf{W}\delta_n, \quad (1)$$

where  $\mathbf{W} \in \mathbb{R}^{t*k}$  is the weight vector of the last layer (notice that for the binary case,  $k = 1$ ). We use  $W_i \in \mathbb{R}^{t*1}$  to denote the  $i$ -th row of it. So, we have  $\theta = (\phi, \mathbf{W})$ .

In general, the empirical loss over the training dataset has the following form:

$$L(\theta) = \sum_{n=1}^N l(f(\mathbf{x}_n; \theta), y_n), \quad (2)$$

where  $l$  is the specified loss function (e.g., exponential loss, cross-entropy, ...). For example, with the exponential loss,  $l(t, y_n) = e^{-y_n t}$ , the empirical risk is given by

$$\begin{aligned} L_{exp}(\theta) &= \sum_{n=1}^N e^{-y_n f(\mathbf{x}_n; \theta)}; \\ L_{exp}(\mathbf{W}, \phi) &= \sum_{n=1}^N e^{-y_n \mathbf{W}\delta_n(\mathbf{x}_n; \phi)}, \end{aligned} \quad (3)$$

where the second expression emphasizes the last weight layer.

For multi-class classification, the commonly used loss function is cross-entropy loss:

$$L_{cross-entropy}(\mathbf{W}, \phi) = - \sum_{n=1}^N \log \left( \frac{\exp(W_{y_n} \delta_n(\mathbf{x}_n; \phi))}{\sum_{l=1}^K \exp(W_l \delta_n(\mathbf{x}_n; \phi))} \right), \quad (4)$$

where  $W_l$  is the  $l$ -th component of  $\mathbf{W}$ , which is the weight for a certain class  $l$ ;  $W_{y_n}$  is the component of  $\mathbf{W}$  for the class represented by  $y_n$ .

The goal of performing optimization is to find:

$$\arg \min_{\theta} L(\theta).$$

In the following, we focus on minimizing Equation (3) using GD algorithm with a constant learning rate  $\eta$  for the binary case and Equation (4) for the multi-class case. At iteration  $t$ , the update rule has the following form:

$$\theta_t = \theta_{t-1} - \eta \nabla L(\theta_{t-1}). \quad (5)$$

### 3 Main Result

In this section, we start with the result in [27] for linearly separable data in logistic regression and then obtain the result for the neural network in Fig. 1. Finally, we extend the result from the binary case to the multi-class case.

In [27], the authors investigate the following problem.

**Definition 1.** For a logistic regression problem, whose weight vector is  $\mathbf{w} \in \mathbb{R}^d$ , the loss has the following form:

$$L_{logistic}(\mathbf{w}) = \sum_{n=1}^N l(y_n \mathbf{w}^\top \mathbf{x}_n).$$

For this binary case, assuming all the labels are positive  $\forall n : y_n = 1$  (we can re-define  $y_n \mathbf{x}_n$  as  $\mathbf{x}_n$ ), we have the GD update for that loss function at iteration  $t$  having the following form:

$$\begin{aligned} \mathbf{w}_t &= \mathbf{w}_{t-1} - \eta \nabla L_{logistic}(\mathbf{w}_{t-1}) \\ &= \mathbf{w}_{t-1} - \eta \sum_{n=1}^N l'(\mathbf{w}_{t-1}^\top \mathbf{x}_n) \mathbf{x}_n. \end{aligned}$$

The authors show that  $\mathbf{w}_t$  finally diverges [27]:

**Lemma 1.** Let  $\mathbf{w}_t$  be the iterates of gradient descent in Definition 1 with  $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$ , where  $\beta$  is the smoothness of  $l$  and  $\sigma_{\max}(X)$  is the maximal singular value of the data matrix  $X \in \mathbb{R}^{d \times N}$  and any starting point  $\mathbf{w}_0$ . For linearly separable data and  $\beta$ -smooth decreasing loss function, we have: (1)  $\lim_{t \rightarrow \infty} L_{\text{logistic}}(\mathbf{w}_t) = 0$ , (2)  $\lim_{t \rightarrow \infty} \|\mathbf{w}_t\| = \infty$  and (3)  $\forall n : \lim_{t \rightarrow \infty} \mathbf{w}_t^\top \mathbf{x}_n = \infty$ .

But the direction of the above solution converges to that of the hard margin SVM solution [27].

**Lemma 2.** For any dataset which is linearly separable, any  $\beta$ -smooth decreasing loss function with an exponential tail (the loss function tail is bounded by two exponential functions), any step size  $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$  and any starting point  $\mathbf{w}_0$ , the gradient descent iterations will behave as:

$$\mathbf{w}_t = \hat{\mathbf{w}} \log t + \rho_t, \quad (6)$$

where  $\hat{\mathbf{w}}$  is the  $L_2$  max margin vector:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{w}\|^2 \\ \text{subject to } &\mathbf{w}^\top \mathbf{x}_n \geq 1, \end{aligned} \quad (7)$$

and the residual grows at most as  $\|\rho_t\| = O(\log \log(t))$ , and so

$$\lim_{t \rightarrow \infty} \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}.$$

Furthermore, except for measuring zero, the residual  $\rho_t$  is bounded.

As for our problem, we have the following assumption:

**Assumption 1.** The loss in Equation (2) converges to zero:  $\lim_{t \rightarrow \infty} L(\theta_t) = 0$ .

This assumption is a reasonable assumption. It could be satisfied as long as the data is linearly or non-linearly separable, with no wrongly labeled data points and the model has enough capacity, which is usually the case for deep learning models. Based on Assumption 1, we have the following lemma:

**Lemma 3.** Under Assumption 1, for the neural network with architecture as in Fig. 1, even if the dataset  $\{\mathbf{x}_n, y_n\}_{n=1}^N$  is not linearly separable, the transformed dataset  $\{\delta_n, y_n\}_{n=1}^N$  is linearly separable:  $\exists \mathbf{W}^*$  such that  $\forall n : y_n \mathbf{W}^* \delta_n > 0$ .

In fact, since the last weight layer is a linear transformation, if  $\{\delta_n, y_n\}_{n=1}^N$  is not linearly separable, the classification error can never reach zero, let alone the loss. Following Definition 1, let us re-define  $y_n \delta_n$  as  $\delta_n$ . Based on Lemma 2 and Lemma 3, we obtain the first main result:

**Theorem 1.** For any neural network for binary classification, any  $\beta$ -smooth decreasing loss function with an exponential tail, small enough step size  $\eta < 2\beta^{-1}\sigma_{\max}^{-2}(X)$  and any start point  $\mathbf{W}_0$ , as long as  $\lim_{t \rightarrow \infty} L(\theta_t) = 0$ , the direction of the neural network's last weight layer converges:

$$\lim_{t \rightarrow \infty} \frac{\mathbf{W}_t}{\|\mathbf{W}_t\|} = \frac{\hat{\mathbf{W}}}{\|\hat{\mathbf{W}}\|}, \quad (8)$$

where  $\hat{\mathbf{W}}$  is the  $L_2$  max margin vector:

$$\begin{aligned} \hat{\mathbf{W}} &= \arg \min_{\mathbf{W} \in \mathbb{R}^{t \times 1}} \|\mathbf{W}\|^2 \\ \text{subject to } &\mathbf{W} \delta_n \geq 1, \end{aligned}$$

in which  $\delta_n$  is the re-defined input of the last weight layer.

It is true that the convergence of the transformation function can also affect the last layer decision boundary. However, since the loss converges to zero, the variance of the transformation function is bounded after long enough training time, which makes the theorem hold.

As for the multi-class classification problem, we have the following lemma from [27]:

**Lemma 4.** For a logistic regression problem in which we learn a predictor  $\mathbf{w}_k$  for each class  $k \in [K]$  in a linearly separable multi-class dataset, any starting point  $\mathbf{w}_{k,0}$  and any small enough step size, under most circumstances (i.e., except for a measure zero), the iterates of gradient descent on the cross-entropy loss will behave as:

$$\mathbf{w}_{k,t} = \hat{\mathbf{w}}_k \log(t) + \rho_{k,t}, \quad (9)$$

where the residual  $\rho_{k,t}$  is bounded and  $\hat{\mathbf{w}}_k$  is the solution of the  $K$ -class SVM:

$$\begin{aligned} & \arg \min_{\mathbf{w}_1, \dots, \mathbf{w}_K} \sum_{k=1}^K \|\mathbf{w}_k\|^2 \\ & \text{subject to} \\ & \forall n, \forall k \neq y_n : \mathbf{w}_{y_n}^\top \mathbf{x}_n \geq \mathbf{w}_k^\top \mathbf{x}_n + 1. \end{aligned} \quad (10)$$

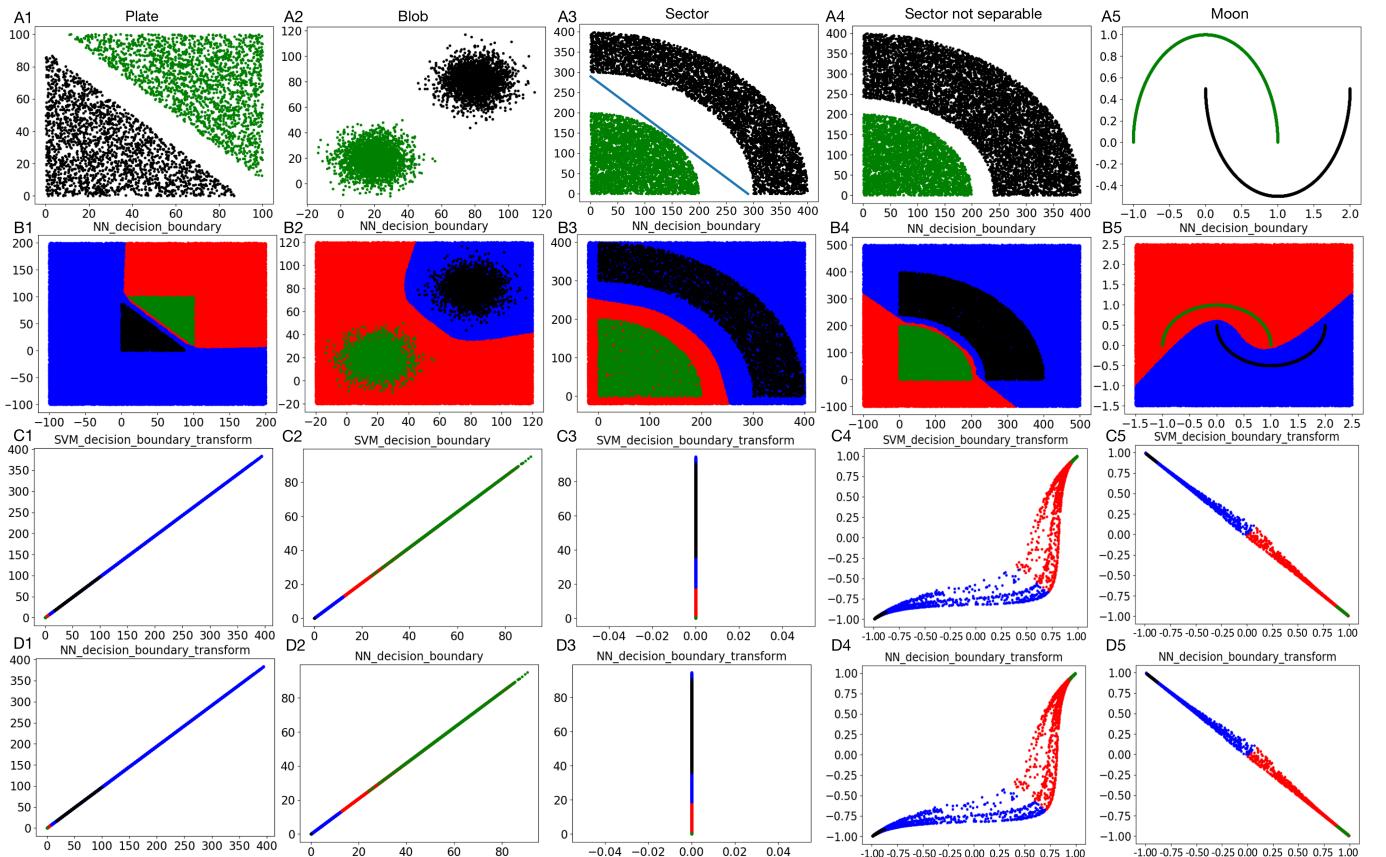


Figure 2: Results on simulated datasets. The five columns are five datasets. The first row is the training datasets in the original input space. In the last three rows, red points are random testing points classified with the same label as the green training data points and blue points are random testing points classified with the same label as the black training data points. The interface between red dots and blue dots is the decision boundary. The second row figures show the decision boundary of a trained neural network in the original space. The third and the fourth rows are in the transformed space. The third row shows the decision boundary of linear SVM trained with the transformed data in the transformed space. The last row shows the decision boundary of the neural network's last weight layer.

Similar to Theorem 1, we can derive the following result for the multi-class case with cross-entropy loss.

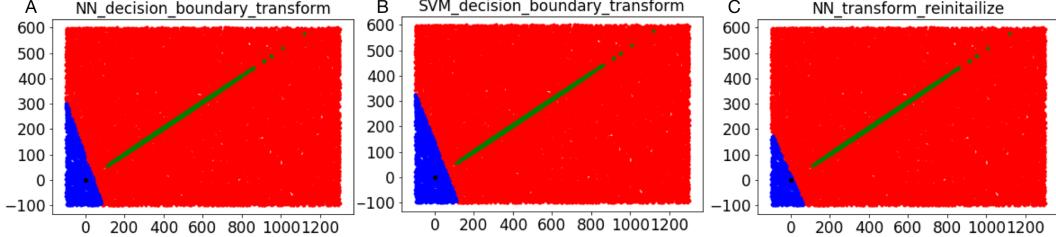


Figure 3: MNIST binary classification decision boundary. We randomly chose two classes from the MNIST dataset (“0” and “1” for the above figures) and trained a neural network with ResNet as the transformation function in Fig. 1. We set  $t$  as 2 for visualization purpose. The above figures show the decision boundary of SVM trained with transformed data (B) and the last weight layer of the neural network (A) in the transformed space. After model converged, we reinitialized the last weight layer and retrained the model with the weights in the transformation function being fixed, resulting in (C).

**Theorem 2.** *For any neural network, small enough step size  $\eta$  and any starting point  $\mathbf{W}_0$ , as long as the dataset makes  $\lim_{t \rightarrow \infty} L(\theta_t) = 0$ , the iterates of gradient descent on the cross-entropy loss of the last weight layer  $\mathbf{W}$  will behave as:*

$$W_{k,t} = \hat{W}_k \log(t) + \rho_{k,t}, \quad (11)$$

where the residual  $\rho_{k,t}$  is bounded,  $W_{k,t}$  is the weight for class  $k$  at iteration  $t$  and  $\hat{W}_k$  is the solution of the  $K$ -class SVM:

$$\begin{aligned} & \arg \min_{W_1, \dots, W_K} \sum_{k=1}^K \|W_k\|^2 \\ & \text{subject to} \\ & \forall n, \forall k \neq y_n : W_{y_n}^\top \delta_n \geq W_k^\top \delta_n + 1, \end{aligned}$$

and so:

$$\lim_{t \rightarrow \infty} \frac{W_{t,k}}{\|W_{t,k}\|} = \frac{\hat{W}_k}{\|\hat{W}_k\|}. \quad (12)$$

## 4 Experiments

### 4.1 Experimental setting

There are seven datasets in our experiments, including five simulated 2D datasets and two real datasets. The five simulated datasets can be referred to Fig. 2 (A1-A5). The first three (Plate, Blob, and Sector) are linearly separable. The last two (Sector not separable and Moon) are non-linearly separable. There are 5000 points within each simulated dataset. The two real datasets are MNIST [15] and CIFAR-10 [18]. Since MNIST and CIFAR-10 are multi-class datasets, we randomly chose two classes out of the 10 classes for each one for the binary classification case. We used the network architecture in Fig. 1 for all the experiments. The only difference is the transformation function. We used a fully connected layer with 2000 nodes as the transformation function for the five simulated datasets; ResNet [10] for MNIST; and DenseNet [11] for CIFAR-10. For visualization purpose, we set  $t$  as 2. We used cross-entropy loss as the loss function and ReLU as the activation function. For multi-class classification problem, we set the number of nodes in the output layer the same as the number of classes. We used GD for the simulated datasets and SGD for MNIST and CIFAR-10. We turned off all the commonly used explicit regularizers, such as weight decay and dropout, for all the experiments.

## 4.2 Simulated datasets

The results are summarized in Fig. 2 (additional results can be found in the Appendices). The decision boundary of neural networks in the original input space can be referred to Fig. 2 (B1-B5). The green and black dots are the training data points. We sampled test data points uniformly across the whole space so that we can visualize the decision boundary of the trained neural networks. The blue points are the ones predicted by the model with the same label as the black training data while the red points are the ones predicted with the same label as the green training data. The curve that separates the blue points and red points can be considered as the decision boundary of the network. Although it is difficult to gain insight from the original space, as suggested by the analysis in the Main Result section, the transformed space is more interesting. Fig. 2 (D1-D5) shows the training data and testing data in the transformed space. As a comparison, we trained a linear SVM with the transformed training data and labeled the same testing data points with the SVM classifier, whose results are shown in Fig. 2 (C1-C5). As shown in the figure, the direction of the neural network’s last layer decision boundary trained with GD converges to that of the linear SVM solution, which verifies Theorem 1. Furthermore, the two kinds of decision boundaries are very close to each other, not only in the direction but also in the constant bias term. We further discuss this phenomenon in the next subsection.

## 4.3 MNIST binary

After training a residual network with the MNIST data, we mapped the data into the transformed space. Within that space, we sampled test data uniformly and labeled those test data points using the last layer of the network in Fig. 1, which results in the decision boundary in Fig. 3 (A). Utilizing the training data in the transformed space, we trained a linear SVM classifier and plotted out the decision boundary of that classifier in Fig. 3 (B). As shown in the figures, after mapping the data into the transformed space, the direction of the first decision boundary is very close to that of the second decision boundary, which further supports Theorem 1. Furthermore, with the transformation function fixed, we reinitialized the last layer and retrained the last layer, whose result is shown in Fig. 3 (C). It suggests that our result still holds. On the other hand, the original boundary obtained by training the network as a whole is closer to the SVM boundary in terms of the bias constant, which suggests the whole network training may have better initialization for the last layer and thus make the model generalize better. Notice that although we turned off dropout and the model had been trained for a very long time to make it completely fit to the training data, the trained model still has very impressive generalization property with the testing accuracy being as high as 99.7%.

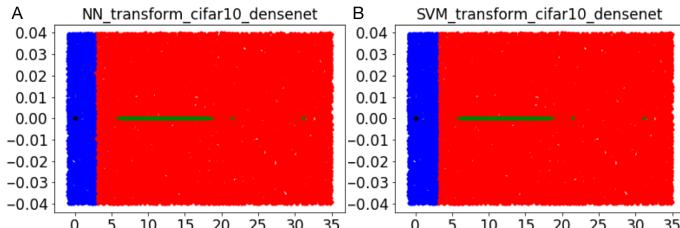


Figure 4: CIFAR-10 binary decision boundary result. We randomly chose two classes from the CIFAR-10 dataset and trained a neural network with DenseNet as the transformation function in Fig. 1, setting  $t$  as 2. (A) shows the decision boundary of the last weight layer in the transformed space and (B) shows the decision boundary of the linear SVM trained with the transformed data in the transformed space.

## 4.4 CIFAR-10 binary

We trained a model with DenseNet transformation function on the CIFAR-10 dataset. The decision boundary results of this dataset could be referred to Fig. 4. As shown in the figure, similar to the result on MNIST, the directions of those two boundaries are very close to each other, which further supports Theorem 1. Furthermore, in addition to being close in terms of direction, the neural network boundary is very close to the midpoint of the two clusters, if it does not cross the midpoint, where the SVM boundary should pass theoretically. This phenomenon is consistent with the result of the

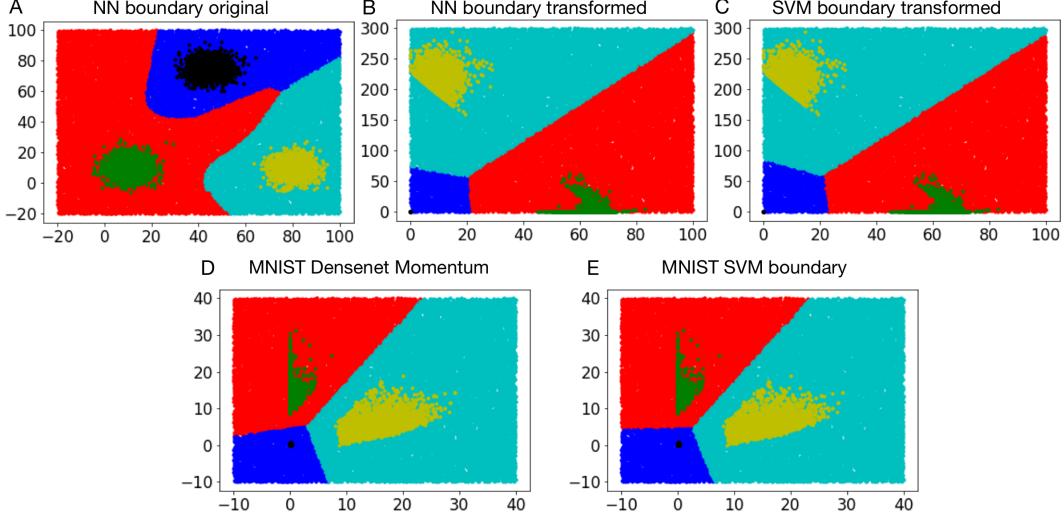


Figure 5: The multi-class experiment result. (A,B,C) show the decision boundary results on a simulated 3-class Blob dataset. (D,E) show the 3-class MNIST (“0”, “1” and “2” for the above figures) result trained with DenseNet and Momentum.

simulated datasets and the MNIST dataset, suggesting that training the whole neural network using GD or SGD may result in a decision boundary with good bias constant. In terms of the trained model’s generalization property, although we turned off explicit regularizers, the model can still have 92.6% testing accuracy for this CIFAR-10 dataset, which is within the performance range of a productive deep learning model.

#### 4.5 Multi-class classification

In practice, deep learning is usually used for multi-class classification with cross-entropy loss. We investigated the multi-class classification case in this section. We performed experiments on a simulated three class Blob dataset. The neural network decision boundary in the original space and the transformed space can be referred to Fig. 5 (A,B), respectively. As a comparison, the SVM decision boundary on the transformed data in the transformed space is shown in Fig. 5 (C). Those results, which show the decision boundary direction of the neural network last weight layer converges to that of SVM, verify Theorem 2. We also performed such experiment on the MNIST data with DenseNet transformation function. During the training, we also tried other optimizers other than just SGD, such as Momentum. The results are shown in Fig. 5 (D,E). From the two figures, we can find that the corresponding decision boundary directions of the neural network last layer and SVM are very close to each other. Besides, similar to the previous result, the decision boundary of neural network is very close to the midpoint between different clusters. Those experiments further support Theorem 2, which also shows that our hypothesis may be generalized to other optimizers, such as Momentum.

#### 4.6 Real task: Fashion MNIST

We also investigated the decision boundary of the DenseNet’s last layer, which is used to perform 10-class classification on Fashion MNIST. We used the same architecture from [11], except for that we added an additional layer to make the last hidden layer in 2D space for visualization purpose. We turned off the commonly used techniques for improving performance, such as data augmentation and dropout. We deployed Momentum as the optimizer. After the model being trained for 1,000 epoches, the loss oscillated around  $6 \times 10^{-4}$ . The testing accuracy is around 91.8%, which is within the known performance range of the deep learning model on this dataset. We show the decision boundary comparison of the network’s last layer and the multi-class linear SVM solution in Fig. 6. As shown in the figure, although the experiment setting is not exactly the same as the assumptions in our main result, the decision boundary of the trained neural network still worths investigating. In fact, the decision boundary shown on the up-left of Fig. 6 (A) is very similar to that of Fig. 6 (B). On the

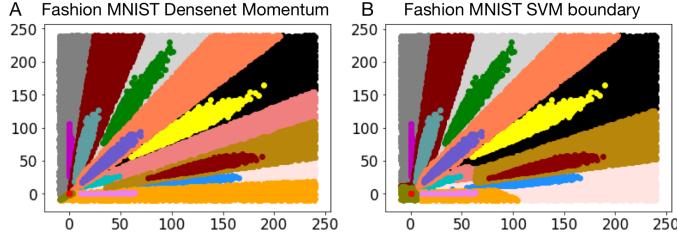


Figure 6: The real task result. We trained a DenseNet for the 10-class Fashion MNIST classification using Momentum. After 1,000 epoches, the loss is around  $6 * 10^{-4}$ . (A) shows the decision boundary of the neural network’s last layer. (B) shows the decision boundary of SVM trained with the transformed dataset.

other hand, the transformed representation of the blue class has very complex spatial relationship with the other three classes around it, which causes the neural network get stuck in local minimum and diverge from the multi-class linear SVM solution. Although this real task does not completely fit our assumption and our result, the experiment shows that the margin theory can have the potential to explain the generalization property of deep learning.

## 5 Discussion

The result of this paper can be useful for solving several practical problems related to deep learning, such as catastrophic forgetting [14] and the data-hungry challenge [6]. We take these two as examples. On the other hand, we believe that investigating the transformation function would be helpful for solving adversarial attacking [23] and studying the last layer can push out new ways of introducing uncertainty into supervised deep learning [9].

### 5.1 Catastrophic forgetting

Catastrophic forgetting [14], which means the neural network does not have the ability of learning new knowledge without forgetting the learned knowledge, is one of the bottlenecks of deep learning. Recently, a rehearsal framework, called SupportNet [19], was proposed to deal with catastrophic forgetting when performing class incremental learning. In short, it maintains a subset of the old data, which is chosen based on the support vector information obtained by using SVM to approximate the last layer, and feeds the subset together with the new data to the model when incorporating the new classes into the model. Despite the lack of theoretical analysis in the paper, the framework works quite well in practice, even achieving nearly optimal performance on some datasets. In fact, according to Lemma 1 and Theorem 2, we can write  $W_{k,t} = c(t)\hat{W}_k + \rho_{k,t}$  such that  $c(t) \rightarrow \infty$  and  $\rho_{k,t}$  is bounded. The gradient of the exponential loss for  $W_{k,t}$  can then be formulated as:

$$\begin{aligned} -\nabla L_{exp}(W_{k,t}) &= \sum_{n=1}^N \exp(-W_{k,t}\delta_n) \delta_n \\ &= \sum_{n=1}^N \exp(-c(t)\hat{W}_k\delta_n) \exp(-\rho_{k,t}\delta_n) \delta_n, \end{aligned} \tag{13}$$

when the model converges and  $c(t) \rightarrow \infty$ , only those data with the largest exponents, that is,  $\hat{W}_k\delta_n$  should be the smallest, will contribute to the gradients. Those samples are exactly the support vectors of the SVM trained on the transformed data, which are selected by SupportNet. Using those data for future tuning, the model is very likely to learn the same boundary for the old classes. Our results partially explains why that rehearsal method works very well in practice.

### 5.2 Reducing the training data size and transfer learning

It is always desirable to reduce the training data size for the data-hungry deep learning method, without too much performance compromise. In practice, especially in the computer vision field, when

the data size is not large enough, people usually take advantage of transfer learning [28], fine-tuning the last one or two layers of a pre-trained model with the training data. In fact, based on our result in the Main Result section and the analysis in the previous subsection, it is not data-hungry from the transformed space to the label space since only the support vector samples matter, which means the drawback property of deep learning comes from the transformation function component. The transfer learning technique, taking advantage of an existing transformation function and avoiding the data size requirement of that component, can thus learn a useful model with limited data.

## 6 Conclusion

Bridging the gap between the theoretical research and the practical power of deep learning is a fascinating research direction. In this paper, we investigate the decision boundary of a productive deep learning architecture with weak assumption on both the training data and the model. Through comprehensive theoretical analysis and experiments, we show that the direction of the neural network's last weight layer converges to that of a linear SVM trained on the transformed data if the loss converges to zero, for both the binary case and the multi-class case with the commonly used cross-entropy loss. In addition, we show it empirically that training a neural network as a whole may result in better bias constant for the last weight layer, which is important for the generalization property of deep learning models. In addition to facilitating the understanding of deep learning and thus further improving its performance, our result can be useful for solving a broad range of practical problems in the deep learning field, such as catastrophic forgetting, reducing the data size requirement of deep learning, adversarial attacking, and introducing uncertainty into deep learning.

## References

- [1] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.
- [2] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.
- [3] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.
- [4] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.
- [5] Davide Castelvecchi. Can we open the black box of ai? *Nature*, 538(7623):20–23, 2016.
- [6] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525, 2014.
- [7] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. *Artificial Intelligence and Statistics*, pages 192–204, 2015.
- [8] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.
- [9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, pages 1050–1059, 2016.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016.
- [11] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1(2):3, 2017.
- [12] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, pages 586–594, 2016.
- [13] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.
- [14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- [15] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.
- [17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [19] Yu Li, Zhongxiao Li, Lizhong Ding, Peng Yang, Yuhui Hu, Wei Chen, and Xin Gao. Supportnet: solving catastrophic forgetting in class incremental learning with support data. *arXiv preprint arXiv:1806.02942*, 2018.
- [20] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *Advances in Neural Information Processing Systems*, pages 597–607, 2017.
- [21] Qianli Liao and Tomaso Poggio. Theory of deep learning ii: Landscape of the empirical risk in deep learning. *arXiv preprint arXiv:1703.09833*, 2017.
- [22] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.
- [23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387, 2016.
- [24] Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.
- [25] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.
- [26] Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.
- [27] Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.
- [28] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.
- [29] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.
- [30] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

# Appendices

## A Additional results of binary classification

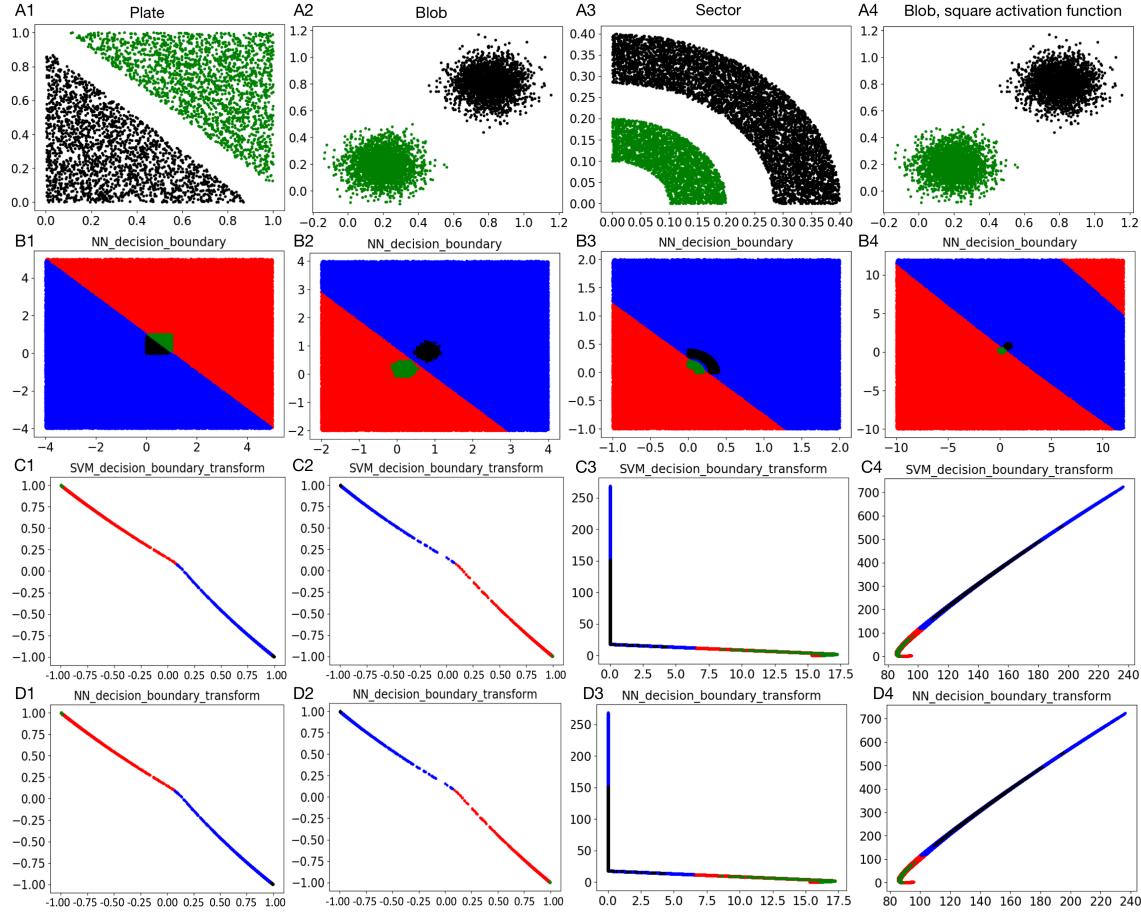


Figure 7: Additional results of the simulated data, whose data range is much smaller than that in the main text. The four rows have the same meanings as the figure in the main text. The first three columns are the results of the neural networks with ReLU activation function on three linearly separable datasets. The last column is the result of the neural network with square activation function trained on Blob dataset.

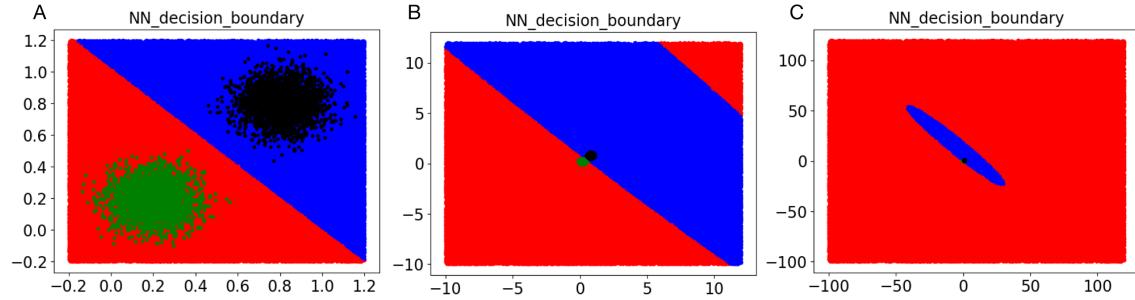


Figure 8: The decision boundary of the neural network with square activation function trained on the Blob dataset in the original space in different scales.

We first want to clarify why in the main text, we chose the data range of the simulated linearly separable datasets to be relatively large, from 0 to around 100 or 400. Here we provide the results of the small-range linearly separable datasets (from 0 to around 1), which can be referred to Fig. 7. Intuitively, the decision boundary in the original space (Fig. 7 (B1-B3)) is very surprising because the highly over-parameterized multi-layer neural network seems to learn a linear decision boundary. We argue that it is because of the small range of the datasets and also the shape of the activation function. As we know, a very large part of the ReLU activation function is linear. If the data range is very small, it is very likely that during training, the nonlinear part of the activation function is not used. As a result, the whole network becomes a linear classifier, which makes the decision boundary linear. We demonstrate that by performing an additional experiment on the small-range Blob dataset with the neural network having the following square activation function:

$$a(u) = u^2.$$

Within this function, there is no linear part. So even the data range is small, the decision boundary of the neural network should still be nonlinear. The experimental results of this setting are shown in the last column of Fig. 7. From Fig. 7 (B4), we can see that the decision boundary in the original space is a nonlinear one, which is as expected. On the other hand, we also show the decision boundary in the original space in different scales in Fig. 8. As shown in Fig. 8 (A, B), although the boundary is nonlinear globally (Fig. 8 (C)), it is very similar to a linear boundary if we only consider its local shape (i.e. from  $-1$  to  $2$ ), which supports our assumption, that is, if the data range is small, the nonlinear power of the activation function is used limitedly. This experiment demonstrates that the data range combining with the activation function can have a significant impact on the decision boundary in the original space. To eliminate the potential misunderstanding and misleading results caused by the datasets and emphasize the main results, we chose the large-range datasets in the main text.

On the other hand, if we investigate the results of the neural network (Fig. 7 (D1-D4)) and the linear SVM (Fig. 7 (C1-C4)) in the transformed space on those small-range datasets, we can find that the results are similar to those on the large-range datasets in the main text, which further supports our main results.