ResNet and Batch-normalization Improve Data Separability

Yasutaka Furusho FURUSHO.YASUTAKA.FM1@IS.NAIST.JP

Kazushi Ikeda KAZUSHI@IS.NAIST.JP

Nara Institute of Science and Technology, Nara 630-0192, Japan

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

The skip-connection and the batch-normalization (BN) in ResNet enable an extreme deep neural network to be trained with high performance. However, the reasons for its high performance are still unclear. To clear that, we study the effects of the skip-connection and the BN on the class-related signal propagation through hidden layers because a large ratio of the between-class distance to the within-class distance of feature vectors at the last hidden layer induces high performance. Our result shows that the between-class distance and the within-class distance change differently through layers: the deep multilayer perceptron with randomly initialized weights degrades the ratio of the between-class distance to the within-class distance and the skip-connection and the BN relax this degradation. Moreover, our analysis implies that the skip-connection and the BN encourage training to improve this distance ratio. These results imply that the skip-connection and the BN induce high performance.

Keywords: Deep learning, ResNet, Skip-connection, Batch-normalization

1. Introduction

Deep neural networks have a high expressive power that grows exponentially with respect to the depth of the neural network (Montufar et al., 2014; Telgarsky, 2016; Raghu et al., 2017). However, the classic multilayer perceptron (MLP) cannot reduce its empirical risk by training even though it stacks more layers (He et al., 2016a). To overcome this problem, the ResNet incorporates skip-connections between layers (He et al., 2016a,b) and the batch-normalization (BN) normalizes the input of activation functions (Ioffe and Szegedy, 2015). These architectures enable an extreme deep neural network to be trained with high performance.

The property of the skip-connection is discussed from the point of view of the signal propagation, that is, the change of feature vectors through layers (Poole et al., 2016; Yang and Schoenholz, 2017). In an MLP with randomly initialized weights, the cosine distance between two feature vectors converges to a fixed point in $[0, 1]$ in an exponential order of the depth (Poole et al., 2016). The skip-connection relaxes this exponential order into a polynomial order and thus preserves the structure of the input space (Yang and Schoenholz, 2017). Their analysis used the mean-field theory, that is, they considered the neural network with infinite hidden units and approximated the sum of the activations by the Gaussian random variable. This approximation can deal with broad classes of the activation function. However, their analysis is limited to the randomly initialized neural networks and the theoretical relationship between their results and the classification performance is unclear.

In this work, we focused on the most popular activation function, the ReLU function, and analyzed the class-related signal propagation through hidden layers of the randomly initialized MLP, the ResNet, and the ResNet with BN and the effect of training because a large ratio of the between-class distance to the within-class distance of feature vectors at the last hidden layer induces high classification performance (Devroye et al., 2013). Our results show that, in randomly initialized weights, the MLP strongly decreases the between-class distance compared with the within-class distance. The skip-connection and the BN relax this decrease of the between-class distance thanks to the preservation of the angle between input vectors. Our analysis also implies that this preservation of the angle at initialization encourages training to improve the distance ratio. These results imply that the skip-connection and the BN induce high performance.

2. Preliminaries

2.1. Problem settings

We have a training set $S = \{(x(n), y(n))\}_{n=1}^{N}$. Each example is a pair of input $x(n) \in \mathbb{R}^D$ and a class label $y(n) \in \{-1, +1\}$ which are independently identically distributed from a probability distribution $\mathcal{D}$. Indices of the examples are omitted if they are clear from the context.

2.2. Neural networks

We consider DNNs, which transform an input vector $x \in \mathbb{R}^D$ into a new feature vector $h^L \in \mathbb{R}^D$ through the following L blocks. Let $h^0 = x$ and $\phi(\cdot) = \max\{0, \cdot\}$ be the ReLU activation function.

Multilayer perceptron (MLP):

$$h_i^{l+1} = \phi\left(u_i^{l+1}\right), \qquad u_i^{l+1} = \sum_{j=1}^{D} W_{i,j}^{l} h_j^{l}. \tag{1}$$

ResNet (Yang and Schoenholz, 2017; Hardt and Ma, 2017) **:**

$$h_i^{l+1} = \sum_{j=1}^{D} W_{i,j}^{l,2} \phi(u_j^{l+1}) + h_i^{l}, \qquad u_i^{l+1} = \sum_{j=1}^{D} W_{i,j}^{l,1} h_j^{l}. \tag{2}$$

ResNet with batch-normalization (BN):

$$\begin{aligned} h_i^{l+1} &= \sum_{j=1}^{D} W_{i,j}^{l,2} \phi\left(\mathrm{BN}(u_j^{l+1})\right) + h_i^{l}, \\ \mathrm{BN}(u_i^{l+1}) &= \frac{u_i^{l+1} - \mathbb{E}\left[u_i^{l+1}\right]}{\sqrt{\mathrm{Var}\left(u_i^{l+1}\right)}}, \qquad u_i^{l+1} = \sum_{j=1}^{D} W_{i,j}^{l,1} h_j^{l}, \end{aligned} \tag{3}$$

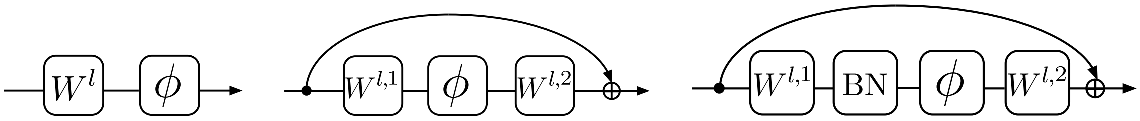


Figure 1: The blocks of the MLP, the ResNet, and the ResNet with BN.

where the expectation is taken under the distribution of input vectors in the mini-batch of the stochastic gradient descent (SGD). Without loss of generality, we assume that the variance of input vectors in the mini-batch is one, $\mathrm{Var}\,(x_d) = 1$ for all $d \in [D]$.

The above DNNs predict the corresponding output for an input based on the feature vector at the last block.

$$\hat{y} = v^T h^L + b. \tag{4}$$

We analyzed the average behaviors of these neural networks when the weights were randomly initialized as follows. In the MLP, the weights were initialized by the He initialization (He et al., 2015) because the activation function is the ReLU function.

$$W^l_{i,j} \sim \mathcal{N}\left(0, \frac{2}{D}\right). \tag{5}$$

In the ResNet and the ResNet with BN, the first internal weights were initialized by the He initialization, but the second internal weights were initialized by the Xavier initialization (Glorot and Bengio, 2010) because the second internal activation function is the identity.

$$W^{l,1}_{i,j} \sim \mathcal{N}\left(0, \frac{2}{D}\right), \quad W^{l,2}_{i,j} \sim \mathcal{N}\left(0, \frac{1}{D}\right). \tag{6}$$

2.3. Classification error and the feature vectors

A large ratio of the between-class distance to the within-class distance of feature vectors h^L at the last block induces a small classification error.

Theorem 1 *Theorem 4.4 of (Devroye et al., 2013). Let m_{+1} and S_{+1} be the population mean and the population covariance matrix of the feature vector h^L for class $+1$. We also define m_{-1} and S_{-1} for class -1 in the same way. Then, for any $v \in \mathbb{R}^D$,*

$$\inf_{b\in\mathbb{R}} \Pr_{(x,y)\sim\mathcal{D}}[y \cdot (v^T h^L + b) \leq 0] \leq \left[1 + \left(\frac{v^T(m_{+1} - m_{-1})}{\sqrt{v^T S_{+1} v} + \sqrt{v^T S_{-1} v}}\right)^2\right]^{-1}. \tag{7}$$

We can apply this result into the classification error in the training set by replacing the population mean and the population covariance matrix with the sample mean and the sample covariance matrix.

2.4. Signal propagation

To clear the effect of the architecture on the distance between feature vectors $h^l(n), h^l(m)$,

$$\|h^l(n) - h^l(m)\|^2 = \|h^l(n)\|^2 + \|h^l(m)\|^2 - 2\|h^l(n)\|\|h^l(m)\| \cos\angle\left(h^l(n), h^l(m)\right), \tag{8}$$

we calculated the length $q^l(n)$ and the angle $\angle^l(n,m)$,

$$\begin{aligned} q^l(n) &= \mathbb{E}\left[\|h^l(n)\|^2\right], & \angle^l(n,m) &= \arccos c^l(n,m), \\ q^l(n,m) &= \mathbb{E}\left[h^l(n)^T h^l(m)\right], & c^l(n,m) &= \frac{q^l(n,m)}{\sqrt{q^l(n)q^l(m)}}, \end{aligned} \tag{9}$$

where $q^l(n,m)$ is the inner product and $c^l(n.m)$ is the cosine similarity. Note that the expectation is taken under the probability distribution of randomly initialized weights.

3. Main results

3.1. Signal propagation

The distance between feature vectors, which can be written as the length and the angle, is related to the classification error. To clear the effect of the skip-connection and the BN on the classification error, we derive the recurrence relations of the length and the angle in the MLP, the ResNet, and the ResNet with BN. The proofs of theorems are in the appendix.

Theorem 2 *The transformation by the randomly initialized MLP remains the length for any feature vector and strongly decreases a large angle compared with a small angle (Fig. 3).*

$$q^{l+1}(n) = \|h^l(n)\|^2, \qquad \angle^{l+1}(n,m) = \arccos\psi(\angle(h^l(n), h^l(m))) \tag{10}$$

where $\psi(\theta) = \frac{1}{\pi}\{\sin\theta + (\pi - \theta)\cos\theta\}$.

Remark 3 *The angle between input vectors which belong to different classes has a larger angle than the angle between feature vectors which belong to the same class in real-life applications (Yamaguchi et al., 1998; Wolf and Shashua, 2003). Therefore, $\psi(\theta)$ strongly decreases the between-class angle compared with the within-class angle (Fig. 2).*

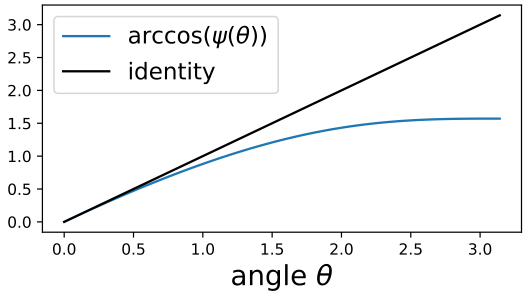


Figure 2: Strong decrease of the between-class angle by $\psi(\theta)$.

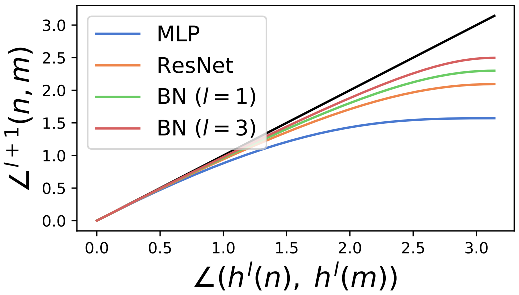


Figure 3: Recurrence relation of the angle by the lth block of the DNN.

Remark 4 *Because of the strong decrease of the between-class angle (Theorem 2 and Fig. 3), the MLP degrades the ratio of the between-class distance to the within-class distance, which is undesirable property for the classification.*

Next theorems show that the skip-connection and the BN relax this degradation.

Theorem 5 *The transformation by the randomly initialized ResNet increases the length in the same scale for any feature vector and relaxes the strong decrease of the large angle compared with the MLP (Fig. 3).*

$$\begin{aligned} &q^{l+1}(n) = 2\|h^l(n)\|^2, \\ &\angle^{l+1}(n,m) = \arccos\left\{\frac{1}{2}\psi(\angle(h^l(n),h^l(m))) + \frac{1}{2}\cos\angle(h^l(n),h^l(m))\right\}. \end{aligned} \tag{11}$$

Theorem 6 *The transformation by the randomly initialized ResNet with BN increases the length in the same scale for any feature vector and relaxes the strong decrease of the large angle compared with the MLP and the ResNet (Fig. 3).*

$$\begin{aligned} &q^{l+1}(n) = \frac{l+3}{l+2}\|h^l(n)\|^2, \\ &\angle^{l+1}(n,m) = \arccos\left\{\frac{1}{l+3}\psi(\angle(h^l(n),h^l(m))) + \frac{l+2}{l+3}\cos\angle(h^l(n),h^l(m))\right\}. \end{aligned} \tag{12}$$

Remark 7 *The skip-connection and the BN relax the degradation of the distance ratio compared with the MLP thanks to the preservation of the angle (Theorem 5, 6, and Fig. 3).*

The above theorems also provide us with a clear interpretation of how the skip-connection and the BN preserve the angle. The ReLU activation function contracts the angle because the ReLU activation function truncates the negative value of its input. The skip-connection bypasses the ReLU activation function and thus reduces the effect of the ReLU activation function to the half. Moreover, the BN reduces the effect of the ReLU activation function to the reciprocal of the depth.

Corollary 8 *Suppose that the change of the angle by each block follows the Theorem 3, 5, and 6. When the angle between input vectors is sufficiently small such that* $\arccos\psi(\angle(x(n),x(m))$ *can be approximated by the linear function* $a\cdot\angle(x(n),x(m))$ *where* $0<a<1$ *is a constant, we can obtain the angle dynamics (Table 1).*

Table 1: Angle dynamics when the input angle is sufficiently small.

Model	Angle $\angle^L(n,m)$
MLP	$\simeq a^L\cdot\angle(x(n),x(m))$
ResNet	$\gtrapprox \left(\frac{1+a}{2}\right)^L\cdot\angle(x(n),x(m))$
BN	$\gtrapprox \frac{2}{L+2}\cdot\angle(x(n),x(m))$

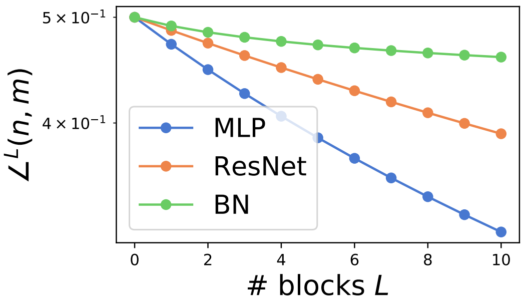


Figure 4: Angle dynamics when the input angle is $\angle(x(n),x(m))=0.5$.

3.2. Role of training and its relation to the initialization

Feature vectors should have a large between-class angle and a small within-class angle for small classification error. However, the randomly initialized neural networks decrease the between-class angle. Wang et al. (2018) showed that minimizing the softmax loss corresponds to increasing the between-class angle and decreasing the within-class angle under some assumptions. Besides the above property of the softmax loss, our analysis provides us with an insight into how training tackles this problem from the point of view of the network architecture.

Theorem 9 *In the MLP, the cosine similarity* $\cos\angle(h^{l+1}(n), h^{l+1}(m))$ *is proportional to*

$$\sum_{i=1}^{D} \delta_i^l(n,m) \cdot \|W_{i,\bullet}^l\|^2 \cdot \cos\angle(W_{i,\bullet}^l, h^l(n)) \cos\left(\angle(W_{i,\bullet}^l, h^l(n)) - \angle(h^l(n), h^l(m))\right) \tag{13}$$

where $W_{i,\bullet}^l = \left(W_{i,1}^l, ..., W_{i,D}^l\right)^T$ *is a weight vector, which can be controlled by training, and* $\delta_i^l(n,m) = 1[W_{i,\bullet}^l h^l(n) > 0] \cdot 1[W_{i,\bullet}^l h^l(m) > 0]$ *is the co-activation state.*

Similar theorems hold in the cases of the ResNet and the ResNet with BN.

Theorem 9 implies that training can find a good weight W^l making the within-class angle smaller and making the between-class angle larger as follows.

Remark 10 $\cos\angle(W_{i,\bullet}^l, h^l(n)) \cos\left(\angle(W_{i,\bullet}^l, h^l(n)) - \angle(h^l(n), h^l(m))\right)$ *in Eq.13 and its plot (Fig. 5) imply training can find a weight vector* $W_{i,\bullet}^l$ *making the angle smaller or larger. However, it seems difficult to make the within-class angle smaller and the between-class angle larger at the same time. The co-activation states* $\{\delta_i^l(n,m)\}_{i=1}^D$ *overcome this problem by being activated class-wisely such that part of it becomes active if* $x(n), x(m)$ *belong to the same class and the other part becomes active if* $x(n), x(m)$ *belong to difference classes.*

The above discussion and the following theorem show the relationship between training and the preservation of the angle at initialization.

Theorem 11 *A small angle* $\angle(h^l(n), h^l(m))$ *encourages the co-activation at initialization.*

$$P\left(\delta_i^l(n,m) = 1\right) = \frac{1}{2}\left(1 - \frac{\angle(h^l(n), h^l(m))}{\pi}\right). \tag{14}$$

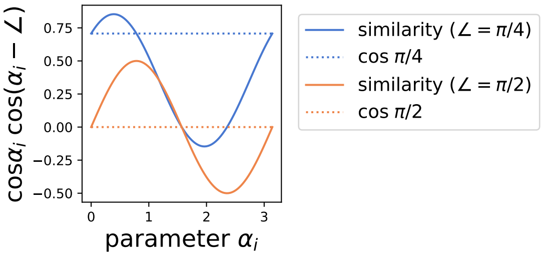
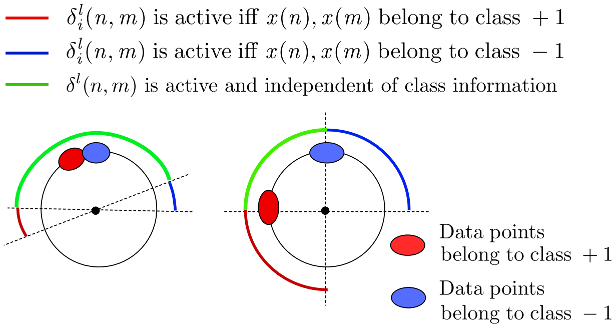


Figure 5: Change of the similarity by changing $\alpha_i = \angle(W_{i,\bullet}^l, h^l(n))$. Let $\angle = \angle(h^l(n), h^l(m))$.



δ_i^l(n, m) is active iff x(n), x(m) belong to class +1
δ_i^l(n, m) is active iff x(n), x(m) belong to class −1
δ^l(n, m) is active and independent of class information
Data points belong to class +1
Data points belong to class −1

Figure 6: Region of $W_{i,\bullet}^l$ and the corresponding behavior of $\delta_i^l(n,m)$.

Remark 12 *At the high block of the initialized MLP, the between-class angle is small and thus large part of the co-activation states $\{\delta_i^l(n,m)\}_{i=1}^D$ are independent of the class information (left in Fig. 6). Therefore, training changes all the angles in a similar way and the improvement of the distance ratio is small. On the other hand, the skip-connection and the BN preserve the between-class angles at high blocks and thus the co-activation states $\{\delta_i^l(n,m)\}_{i=1}^D$ still depend on the class information (right in Fig. 6). Therefore, training can change angles class-wisely and improves the distance ratio well.*

4. Numerical simulations

4.1. Dataset

We made a training set and a test set $\{x(n), y(n)\}_{n=1}^{1000}$ by subsampling data which belong to the class label 0 or 1 from the MNIST dataset or CIFAR10 dataset. Moreover, we applied PCA-whitening to these datasets to make the variance of each element of input vectors become one. The class label 0 was relabeled to -1.

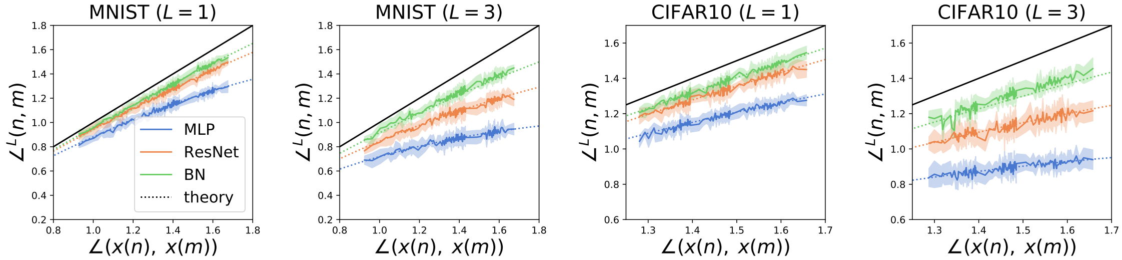


Figure 7: Recurrent relation of the angle by one block or three blocks of neural networks. We plotted the mean and the standard deviation of the angle over ten randomly initialized weights. The black lines show the identity.

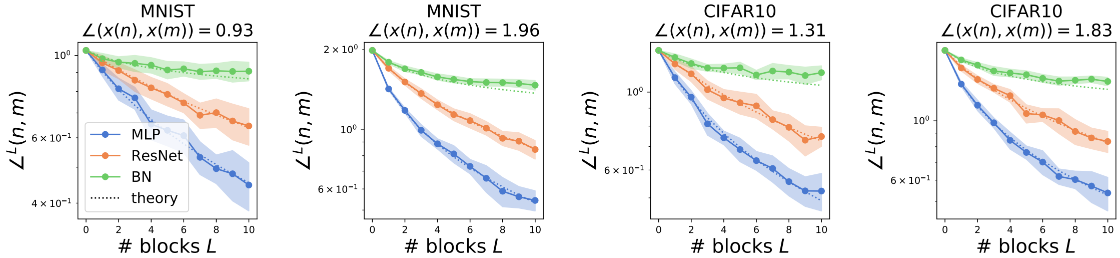


Figure 8: Dynamics of the angle through blocks of neural networks. We plotted the mean and the standard deviation of the angle over ten randomly initialized weights. The black lines show the identity.

4.2. Signal propagation through blocks

We calculated angles between feature vectors transformed by one block or three blocks of the neural networks (Fig. 7). The MLP strongly decreased the between-class (large) angle compared with the within-class (small) angle, the skip-connection and the BN relaxed this decrease, and these numerical values matched our theoretical values.

We chose two input vectors and calculated its angle at each block (Fig. 8). The MLP made this angle decrease quickly and the skip-connections and the BN relaxed this decrease, which agreed with our analysis.

4.3. Co-activation rate through blocks

We calculated the co-activation rates between feature vectors transformed by three blocks of the neural networks (Fig. 9). A small angle encouraged the co-activation and these numerical values matched our theoretical values.

We chose two input vectors and calculated its activation rate at each block (Fig. 10). The activation rate of the MLP quickly increased through the blocks and the skip-connection and the BN suppressed this increase, which agreed with our analysis.

4.4. Effect of training on the distance ratio and relation to the error

We stacked a softmax layer on top of the L blocks neural networks and trained these models with the stochastic gradient descent by minimizing the cross-entropy loss. We calculated error rates and the ratios of the between-class distance to the within-class distance in Theorem 1,

$$\frac{v^T(m_{+1} - m_{-1})}{\sqrt{v^T S_{+1} v} + \sqrt{v^T S_{-1} v}} \qquad \text{where } \; v = (m_{+1} - m_{-1}), \tag{15}$$

on the test set. We plotted distance ratios before and after training (Fig. 11) and the relationship between the distance ratio and the classification error (Fig. 12). We applied

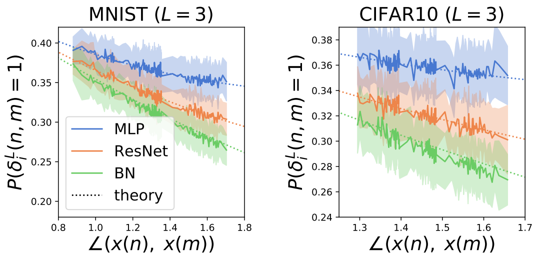


Figure 9: Co-activation rates. We plotted the mean and the standard deviation of the co-activation rate over ten randomly initialized weights.

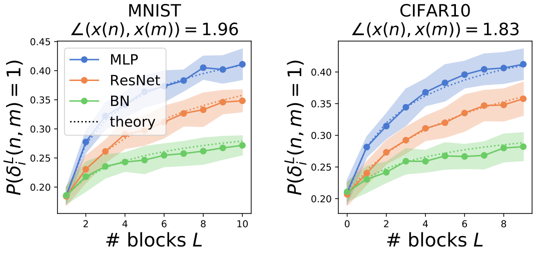


Figure 10: Co-activation rates through blocks. We plotted the mean and the standard deviation of the co-activation rate over ten randomly initialized weights.

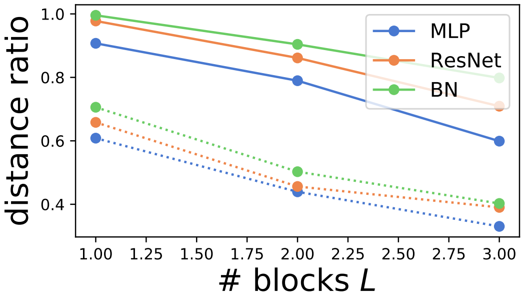


Figure 11: Distance ratio of the neural networks before (dot lines) and after training (bold lines).

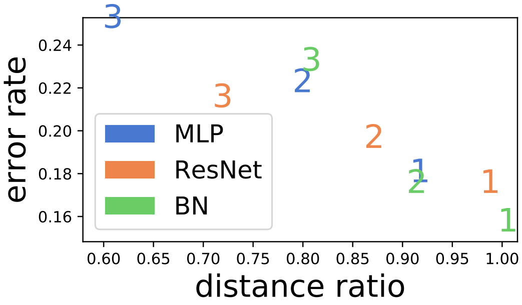


Figure 12: Error rate vs. the distance ratio. Each number means the number of blocks of the neural networks.

this simulation only on the binary class CIFAR10 dataset because the binary class MNIST dataset was so easy that the all DNN perfectly classified data even on the test set.

Fig. 11 shows that the skip-connection and the BN improve the distance ratio compared with the MLP although deeper neural networks degraded the distance ratio. Fig. 12 shows that the error rate is negatively correlated to the distance ratio (the correlation coefficient is -0.91). These results supported our claim.

5. Conclusion

To clear the success of the ResNet and the BN, we analyzed the change of the distance between feature vectors through blocks of the MLP, the ResNet, and the ResNet with BN because a large ratio of the between-class distance to the within-class distance of feature vectors at the last block induces small classification error. Our results show that the randomly initialized MLP degrades this ratio because of the strong decrease of the between-class angle. The skip-connection and the BN relax this degradation thanks to the preservation of the between-class angle. Our analysis also implies that this preservation of the angle at initialization encourages training to improve the distance ratio and thus the skip-connection and the BN induce high performance. Numerical simulations supported these results.

Acknowledgments

This work was supported in part by JSPS-KAKENHI numbers, 18J15055 and 18K19821, and the NAIST Big Data Project.

References

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Raja Giryes, Guillermo Sapiro, and Alexander M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Trans. Signal Processing*, 64(13):3444–3457, 2016.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Talha Cihad Gulcu and Alper Gungor. Comments on" deep neural networks with random gaussian weights: A universal classification strategy?". *arXiv preprint arXiv:1901.02182*, 2019.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *International Conference on Learning Representations*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning*, pages 2847–2854, 2017.

Matus Telgarsky. benefits of depth in neural networks. In *Conference on Learning Theory*, pages 1517–1539, 2016.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

Lior Wolf and Amnon Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4(Oct):913–931, 2003.

Osamu Yamaguchi, Kazuhiro Fukui, and K-i Maeda. Face recognition using temporal image sequence. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 318–323. IEEE, 1998.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pages 7103–7114, 2017.

Appendix A. Proof of the recurrence relation (Theorem 3,5,6, and 11)

Before we prove the recurrence relation, we present the following lemma which plays an important role in our analysis. The strategy of its proof is the same way as Giryes et al. (2016); Gulcu and Gungor (2019), that is, we use the fact that the random Gaussian vector is uniformly distributed on the sphere. Theorem 11 can be proved in the same way.

Lemma 13 *Let $w \in \mathbb{R}^D$ be a random vector which follows the Gaussian distribution $\mathcal{N}\left(0, \sigma^2 I\right)$. Then, for $x(n), x(m) \in \mathbb{R}^D$,*

$$\mathbb{E}\left[\phi(w^T x(n)) \cdot \phi(w^T x(m))\right] = \frac{\sigma^2}{2}\|x(n)\|\|x(m)\| \cdot \psi(\angle(x(n), x(m))) \tag{16}$$

where $\psi(\theta) = \frac{1}{\pi}\{\sin\theta + (\pi - \theta)\cos\theta\}$.

Proof of Lemma 13 Without loss of generality, we can take $x(n)$ to lie along the w_1-axis and $x(m)$ to lie on the $w_1 w_2$-plane. Integrate out the $D-2$ orthogonal coordinates of the random weight vectors and represent $x(n), x(m)$ by the remaining two-dimensional Cartesian coordinate system of $w_1 w_2$-plane. Consider integration over $w_1 w_2$-plane by the change of variables $(w_1, w_2) = (r\cos\theta, r\sin\theta)$. Let $u(\theta) = (\cos\theta, \sin\theta)^T$. Then,

$$\mathbb{E}\left[\phi(w^T x(n)) \cdot \phi(w^T x(m))\right] = \int_0^\infty \int_0^{2\pi} p(ru(\theta))\phi(ru(\theta)^T x(n))\phi(ru(\theta)^T x(m)) \; J d\theta dr \tag{17}$$

where $p(ru(\theta)) = \frac{1}{2\pi\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right)$ is the two dimensional Gaussian density function and $J = \left|\det\left[\frac{\partial(w_1,w_2)^T}{\partial r}, \frac{\partial(w_1,w_2)^T}{\partial\theta}\right]\right| = r$ is the Jacobian. Notice that $\phi(u(\theta)^T x(n)) \cdot \phi(u(\theta)^T x(m))$ is non-zero iff $u(\theta)$ is the same direction as input vectors $x(n), x(m)$ (Fig. 11).

$$\begin{aligned}\mathbb{E}\left[\phi(w^T x(n)) \cdot \phi(w^T x(m))\right] &= \int_0^\infty p(ru(\theta))\; r^3 \; dr \cdot \int_0^{2\pi} \phi(u(\theta)^T x(n)) \; \phi(u(\theta)^T x(m)) \; d\theta \\ &= \frac{\sigma^2}{\pi}\int_{-(\frac{\pi}{2} - \angle(x(n),x(m)))}^{\frac{\pi}{2}} \|x(n)\|\cos\theta \cdot \|x(m)\|\cos(\theta - \angle(x(n), x(m)))\; d\theta \\ &= \frac{\sigma^2}{2}\|x(n)\|\|x(m)\| \cdot \psi(\angle(x(n), x(m))).\end{aligned} \tag{18}$$

Table 2: Recurrence relation. Let $\psi(\theta) = \frac{1}{\pi}\{\sin\theta + (\pi - \theta)\cos\theta\}$

Model	Length $q^{l+1}(n)$	Inner product $q^{l+1}(n,m)$
MLP	$\lVert h^l(n)\rVert^2$	$\lVert h^l(n)\rVert \lVert h^l(m)\rVert \psi(\angle(h^l(n), h^l(m)))$
ResNet	$2\lVert h^l(n)\rVert^2$	$\lVert h^l(n)\rVert \lVert h^l(m)\rVert \left(\psi(\angle(h^l(n), h^l(m))) + \cos\angle(h^l(n), h^l(m))\right)$
BN	$\frac{l+3}{l+2}\lVert h^l(n)\rVert^2$	$\lVert h^l(n)\rVert \lVert h^l(m)\rVert \left(\frac{1}{l+3}\psi(\angle(h^l(n), h^l(m))) + \frac{l+2}{l+3}\cos\angle(h^l(n), h^l(m))\right)$

A.1. Recurrence relation of MLP

Let $W^l_{i,\bullet} = \left(W^l_{i,1}, ..., W^l_{i,D}\right)$ be a row vector.
Length:

$$\begin{aligned} q^{l+1}(n) &= \mathbb{E}\left[\lVert h^{l+1}(n)\rVert^2\right] = \sum_{i=1}^{D} \mathbb{E}\left[\phi(W^l_{i,\bullet} h^l(n))^2\right] \\ &= \sum_{i=1}^{D} \frac{1}{2}\mathbb{E}\left[\sum_{j=1}^{D} {W^l_{i,j}}^2 h^l_j(n)^2\right] = \lVert h^l(n)\rVert^2. \end{aligned} \tag{19}$$

Inner-product:

$$\begin{aligned} q^{l+1}(n,m) &= \mathbb{E}\left[h^{l+1}(n)^T h^{l+1}(m)\right] \\ &= \mathbb{E}\left[\phi(W^l h^l(n))^T \phi(W^l h^l(m))\right] \\ &= \sum_{i=1}^{D} \mathbb{E}\left[\phi(W^l_{i,\bullet} h^l(n))^T \phi(W^l_{i,\bullet} h^l(m))\right]. \end{aligned}$$

Note that $W^l_{i,\bullet}$ is the random Gaussian vector. Apply Lemma 11 into the above equation.

$$q^{l+1}(n,m) = \lVert h^l(n)\rVert \lVert h^l(m)\rVert \cdot \psi\left(\angle(h^l(n), h^l(m))\right). \tag{20}$$

Then, we can calculate the recurrence relation of the angle by applying these into Eq.9.

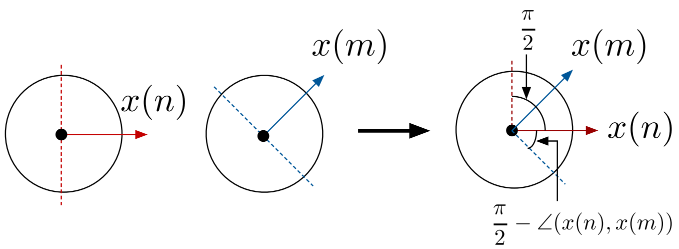


Figure 13: Condition of $\phi(u(\theta)^T x(n)) \cdot \phi(u(\theta)^T x(m)) > 0$.

A.2. Recurrence relation of ResNet

Length:

$$\begin{aligned} q^{l+1}(n) &= \mathbb{E}\left[\|W^{l,2}\phi(W^{l,1}h^l(n)) + h^l(n)\|^2\right] \\ &= \mathbb{E}\left[\|W^{l,2}\phi(W^{l,1}h^l(n)\|^2\right] + 2\mathbb{E}\left[h(n)^{l^T} W^{l,2}\phi(W^{l,1}h^l(n))\right] + \|h^l\|^2 \\ &= 2\|h^l(n)\|^2. \end{aligned} \tag{21}$$

Inner-product:

$$\begin{aligned} q^{l+1}(n,m) &= \mathbb{E}\left[\left\{W^{l,2}\phi(W^{l,1}h^l(n) + h^l(n)\right\}^T \left\{W^{l,2}\phi(W^{l,1}h^l(m) + h^l(m)\right\}\right] \\ &= \mathbb{E}\left[\left\{W^{l,2}\phi(W^{l,1}h^l(n))\right\}^T \left\{W^{l,2}\phi(W^{l,1}h^l(m))\right\}\right] + h^l(n)^T\mathbb{E}\left[W^{l,2}\phi(W^{l,1}h^l(m)\right] \\ &\qquad + h^l(m)^T\mathbb{E}\left[W^{l,2}\phi(W^{l,1}h^l(n)\right] + h^l(n)^T h^l(m). \end{aligned} \tag{22}$$

The first term can be calculated using Lemma 11.

$$\begin{aligned} &\mathbb{E}\left[\left\{W^{l,2}\phi(W^{l,1}h^l(n))\right\}^T \left\{W^{l,2}\phi(W^{l,1}h^l(m))\right\}\right] = \sum_{i=1}^{D}\mathbb{E}\Big[\sum_{j=1}^{D} {W_{i,j}^{l,2}}^2 \phi(W_{j,\bullet}^{l,1}h^l(n)) \cdot \phi(W_{j,\bullet}^{l,1}h^l(m))\Big] \\ &= \sum_{i=1}^{D}\frac{1}{D}\mathbb{E}\Big[\sum_{j=1}^{D}\phi(W_{j,\bullet}^{l,1}h^l(n)) \cdot \phi(W_{j,\bullet}^{l,1}h^l(m))\Big] = \|h^l(n)\|\|h^l(m)\| \cdot \psi(\angle(h^l(n), h^l(m))). \end{aligned} \tag{23}$$

The second term and the third term are zero. Then,

$$q^{l+1}(n,m) = \|h^l(n)\|\|h^l(m)\|\big(\psi(\angle(h^l(n), h^l(m))) + \cos\angle(h^l(n), h^l(m))\big). \tag{24}$$

Thus, we can calculate the recurrence relation of the angle by applying these into Eq.9.

A.3. Recurrence relation of ResNet with BN

We first calculate the statistics of BN.

Mean:

$$\mathbb{E}\left[u_i^{l+1}\right] = \mathbb{E}\left[W_{i,\bullet}^{l,1}h^l\right] = 0. \tag{25}$$

Variance:

$$\begin{aligned}
\mathrm{Var}\left(u_i^{l+1}\right) &= \mathrm{Var}\Big(\sum_{j=1}^{D} W_{i,j}^{l,1} h_j^l\Big) = \frac{2}{D}\sum_{j=1}^{D}\mathrm{Var}\left(h_j^l\right) \\
&= \frac{2}{D}\sum_{j=1}^{D}\mathrm{Var}\left(W_{j,\bullet}^{l-1,2}\phi(\mathrm{BN}(W^{l-1,1}h^{l-1})) + h_j^{l-1}\right) \\
&= \frac{2}{D}\sum_{j=1}^{D}\Big\{\sum_{k=1}^{D}\frac{1}{D}\frac{1}{2} + \mathrm{Var}\left(h_j^{l-1}\right)\Big\} \qquad (26) \\
&= \frac{2}{D}\sum_{j=1}^{D}\Big\{\frac{1}{2} + \mathrm{Var}\left(h_j^{l-1}\right)\Big\} \\
&= \frac{2}{D}\sum_{j=1}^{D}\left(\frac{l}{2} + \mathrm{Var}\left(x_j\right)\right) = l + 2.
\end{aligned}$$

By applying these statistics into Eq.3, we can derive the recurrence relations in the same way as those of the ResNet.

Appendix B. Proof of the angle dynamics through layers (Table 1)

B.1. Angle dynamics through layers of MLP

If θ is small, $\arccos(\psi(\theta))$ can be well approximated by linear function, $a \cdot \angle(h^l(n), h^l(m))$, where $a < 1$ is a positive constant. Therefore,

$$\angle^1(n,m) \simeq a \cdot \angle(x(n), x(m)) \qquad (27)$$

Apply this procedure L times, we obtain

$$\angle^L(n,m) \simeq a^L \cdot \angle(x(n), x(m)). \qquad (28)$$

B.2. Angle dynamics through layers of ResNet and ResNet with BN

Notice that $\psi(\theta) + \cos\theta$ is positive in $\theta \in [0, 2]$ and $\arccos\theta$ is concave in positive θ. Therefore, we can obtain a lower bound of the angle dynamics of the ResNet by using the Jensen inequality.

$$\begin{aligned}
\angle^1(n,m) &= \arccos\left(\frac{1}{2}\psi(\angle(x(n),x(m))) + \frac{1}{2}\cos\angle(x(n),x(m))\right) \\
&> \frac{1}{2}\arccos\left(\psi(\angle(x(n),x(m)))\right) + \frac{1}{2}\cdot\angle(x(n),x(m)) \qquad (29) \\
&\simeq \frac{1+a}{2}\cdot\angle(x(n),x(m))
\end{aligned}$$

Apply this procedure L times, we obtain

$$\angle^L(n,m) \gtrapprox \left(\frac{1+a}{2}\right)^L \cdot \angle(x(n), x(m)). \qquad (30)$$

We can obtain a lower bound of the angle dynamics of the ResNet with BN in the same way as that of the ResNet.

Appendix C. Proof of Theorem 9

Cosine similarity can be written as follows.

$$\begin{aligned}
&\cos\angle(h^{l+1}(n), h^{l+1}(m)) = \frac{h^{l+1}(n)^T h^{l+1}(m)}{\|h^{l+1}(n)\|\|h^{l+1}(m)\|} \\
&= \frac{1}{\|h^{l+1}(n)\|\|h^{l+1}(m)\|}\sum_{i=1}^{D}\phi(W^l_{i,\bullet}h^l(n))\cdot\phi(W^l_{i,\bullet}h^l(m)) \\
&= \frac{\|h^l(n)\|\|h^l(m)\|}{\|h^{l+1}(n)\|\|h^{l+1}(m)\|}\sum_{i=1}^{D}\delta^l_i(n,m)\cdot\|W^l_{i,\bullet}\|^2\cdot\cos\angle(W^l_{i,\bullet}, h^l(n))\cos\angle(W^l_{i,\bullet}, h^l(m)) \\
&\propto \sum_{i=1}^{D}\delta^l_i(n,m)\cdot\|W^l_{i,\bullet}\|^2\cdot\cos\angle(W^l_{i,\bullet}, h^l(n))\cos\big(\angle(W^l_{i,\bullet}, h^l(n)) - \angle(h^l(n), h^l(m))\big)
\end{aligned} \tag{31}$$

where $\delta^l_i(n,m) = 1[W^l_{i,\bullet}h^l(n) > 0]\cdot 1[W^l_{i,\bullet}h^l(m) > 0]$.