Studying the effect of training noise on tail behavior of weights and generalization in neural networks

Li Jipeng, Yuan Xueqiong

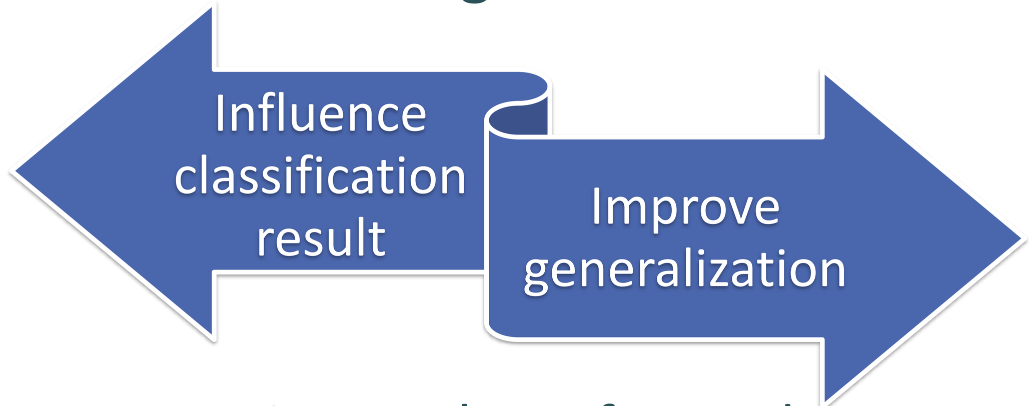
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China

ABSTRACT

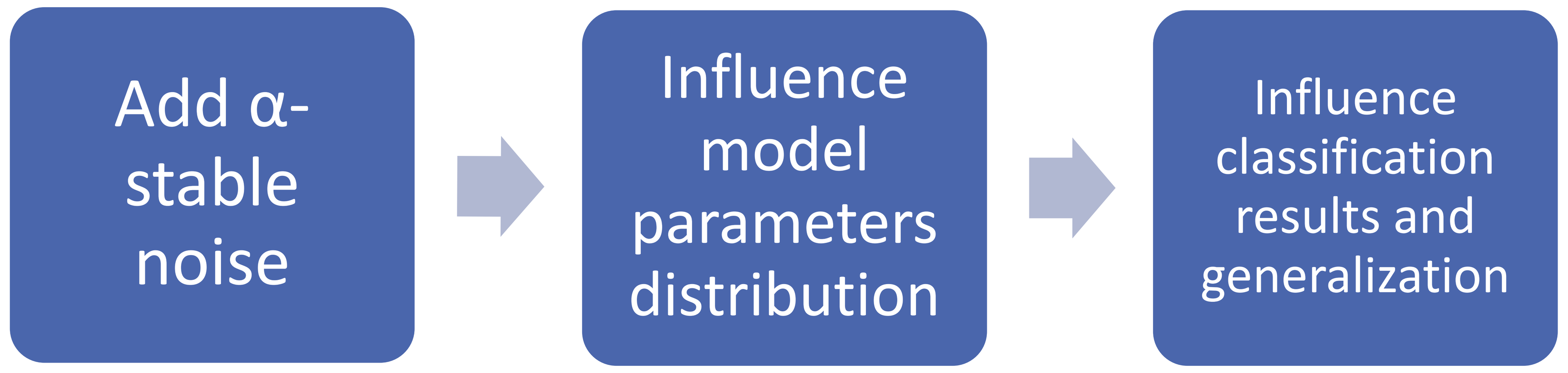
- Add alpha-stable noise with different parameters to two datasets and analyze their influences
- Find that adding noise leads to heavier tails to the weights of specific layers
- Find that Gaussian noise improves model generalization, but overfitting occurs when the noise becomes heavier-tailed

INTRODUCTION

- It is still an open question as to **how noise affects the training results of a model and what level of noise is beneficial for model training**.



- Most relevant experiments have focused on Gaussian noise, but in real life, alpha-stable noise is widely available.



METHOD

A. Alpha-stable noise

Alpha-stable noise is noise that conforms to the alpha-stable distribution and is defined by its characteristic function:

$$E(e^{itx}) = e^{-\sigma^{\alpha}|t|^{\alpha}} \quad (1)$$

The parameter α is the tail index, which controls the thickness of the distribution and ranges from 0 to 2. When α=2, it degenerates to Gaussian distribution, and when α=1, it degenerates to Cauchy. In this project, the value of α is chosen as 0.9,1,1.3,1.5,1.9 and 2.

B. Dataset

MNIST and CIFAR10

C. Model

A 3-layers deep neural network was applied for both MNIST and CIFAR10 dataset, and the parameters are shown in the Table I.

TABLE I Parameters of the neural network

DEPTH	WIDTH	LEARNING RATE	BATCH SIZE
3	512	0.05	5

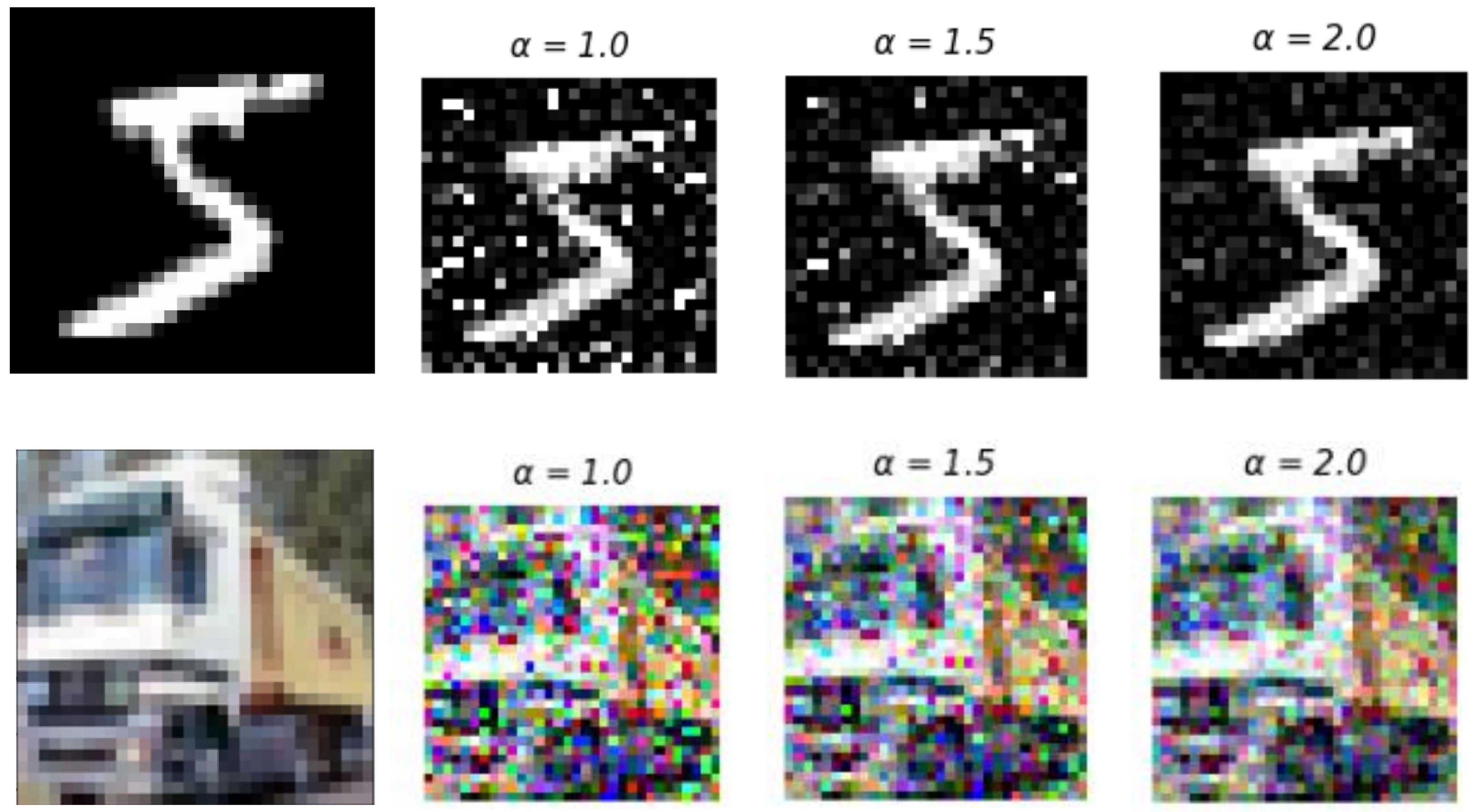


Figure 1. Examples of MNIST and CIFAR10 figure with noise

RESULT

- Part of the important results are shown in Figure 2 and Figure 3.

CONCLUSION

- Adding noise makes the alpha of the weights between the first and second hidden layers **smaller** than that of clean data
- Compared with clean data, adding noise can **improve the training accuracy.**
- When the **noise is close to Gaussian (alpha=2 and 1.9),** the test accuracy is significantly improved—**improve generalization**.
- When **the alpha of noise becomes smaller**, the test accuracy becomes smaller, and the gap between test accuracy and training accuracy becomes larger—**worse generalization.**

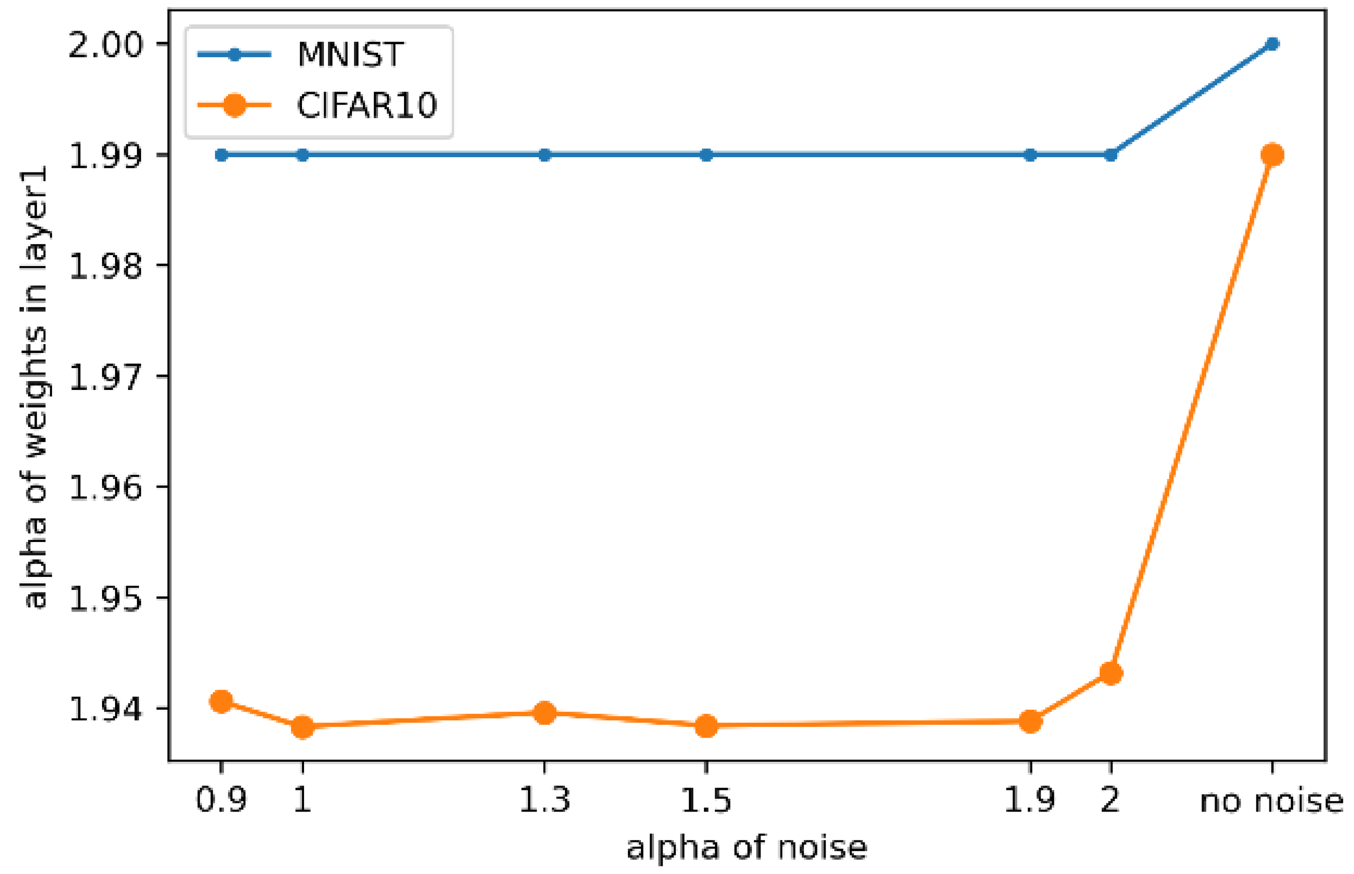


Figure 2. Alpha of the weights between the first and second hidden layers with different alpha of noise.

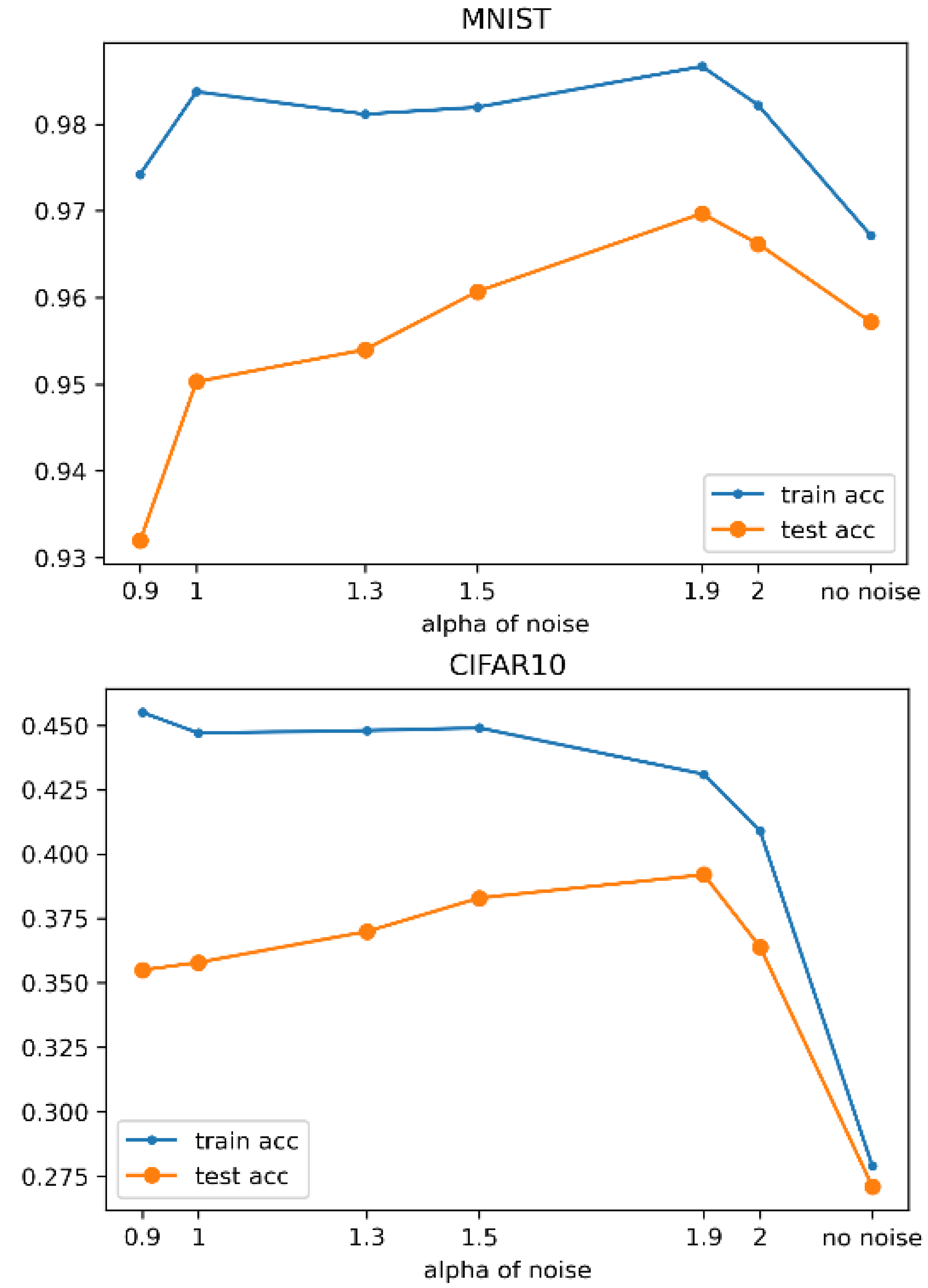


Figure 3. Train accuracy and test accuracy with different alpha of noise.