KCC 2022

A Statistical Analysis of Stochastic Gradient Noises for GNNs

이정현, 정민찬, 허남규

KAIST AI
Graduate School of AI

Optimization and Statistical Inference LAB

한국정보과학회
Korean Institute of Information Scientists and Engineers

CONTENTS

1. Introduction

1.Introduction

Most of the optimization problems in ML/DL can be expressed as the following formulation:

$$\min_{w} F(w) \triangleq \min_{w} \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(w)$$

, where summands are the individual loss contributed by each data point or minibatch.

1.Introduction

- This optimization problem is usually solved via stochastic version of gradient descent method, called the stochastic gradient descent (SGD).

$$w_{t+1} = w_t - \eta \nabla \tilde{f}_k(w_t) = w_t - \eta \nabla F(w_t) + \eta U_t$$

$$\left(\nabla \tilde{f}_k(w) = (1/|\Omega_k|) \sum_{i \in \Omega_k} \nabla f^{(i)}(w) \, , \, U_t(w) = \nabla F(w) - \nabla \tilde{f}_k(w) \right)$$

- Computationally efficient
- The noises of SGD contribute towards better generalization capability of the resulting model.
  - Empirical studies: [Keskar et al., ICLR'16] [Smith et al., ICML'20] ...etc.
  - Theoretical studies: [Pesme et al., NeurIPS'21] [Damian et al., NeurIPS'21] ...etc.

1.Introduction

$$w_{t+1} = w_t - \eta \nabla \widetilde{f}_k(w_t) = w_t - \eta \nabla F(w_t) + \eta U_t$$

- Depending on the empirical observation and/or modeling assumption, one can either choose to model the stochastic gradient noise(SGN) U_t as normal or heavy-tailed.

- Precisely speaking, this distinction comes from whether that the second moment of U_t is finite or infinite.
  - Heuristically, with big enough batch size, we can invoke the (generalized) central limit theorem, depending on the assumption.

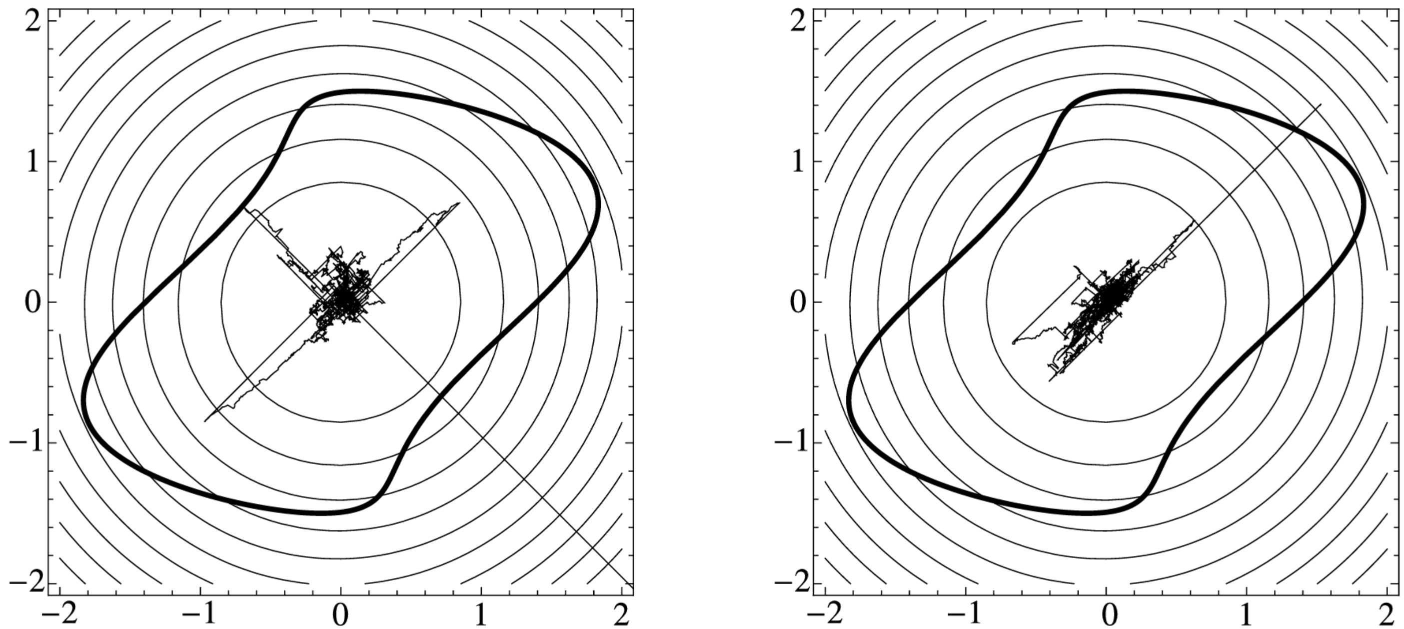
1.Introduction

- The importance of such assumption is highlighted when we analyze SGD via its counterpart SDE.
  - Under appropriate limit (vanishing learning rate, big enough batch size), we can analyze the SGD in the continuous regime.
  - Close connection with the stochastic gradient Langevin dynamics. [Welling & Teh, ICML'11] [Mandt et al., JMLR 2017]

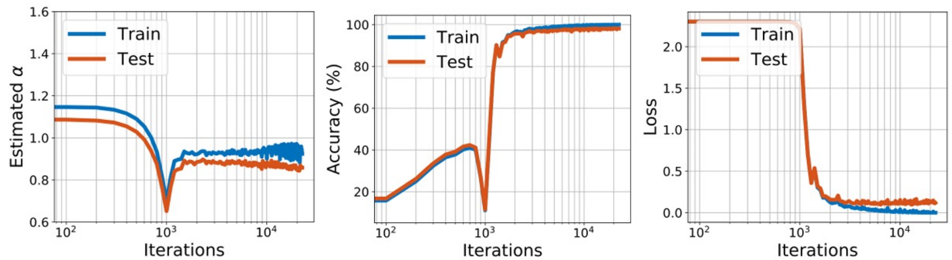
- The stochastic process driving that SDE thus depends on the assumption!
  - SGD as Brownian-driven SDE: [Li et al., JMLR 2019] [Li et al., NeurIPS'21]
  - SGD as Levy-driven SDE: [Simsekli et al., ICML'19] [Zhou et al., NeurIPS'20]

1.Introduction

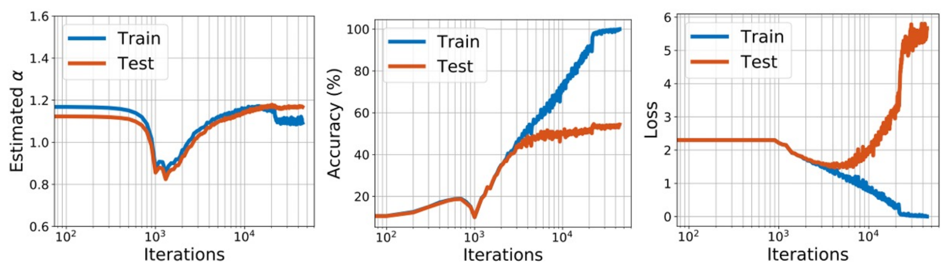
- The behaviors of these SDEs are completely different (see [Simsekli et al., ICML'19] and references therein for more details)



- Therefore, by empirically measuring the tail property of SGN, we can expect the characteristic of training process.
- This was first proposed in [Simsekli et al., ICML'19]:



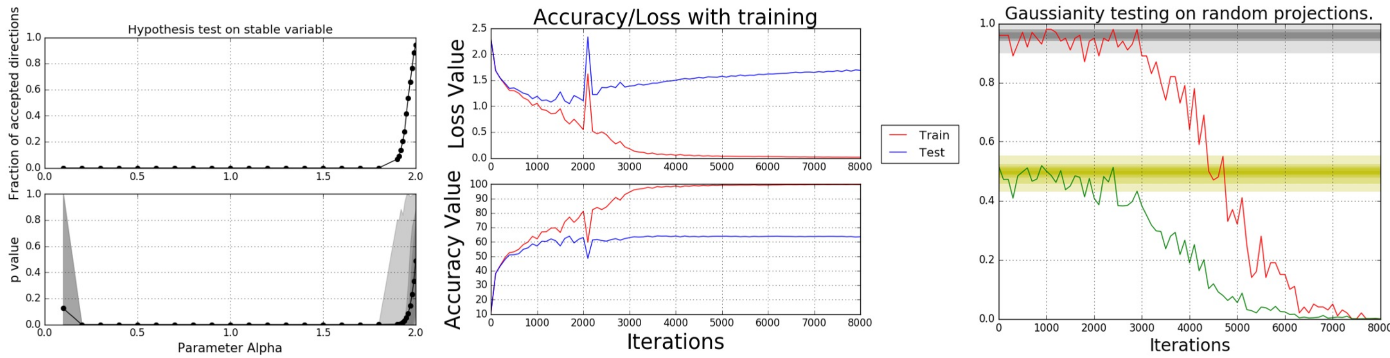
(a) MNIST



(b) CIFAR10

Figure 7. The iteration-wise behavior of of α for the FCN.

- Preliminary statistical tests [Panigrahi et al., NeurIPSW'19] showed that the heavy-tailedness depends heavily on the hyperparameters



(a) (b) (c)

- A more sophisticated statistical analysis [Wang et al., ICLR'22] showed that actually, the SGN often displays the behavior of lognormal distribution

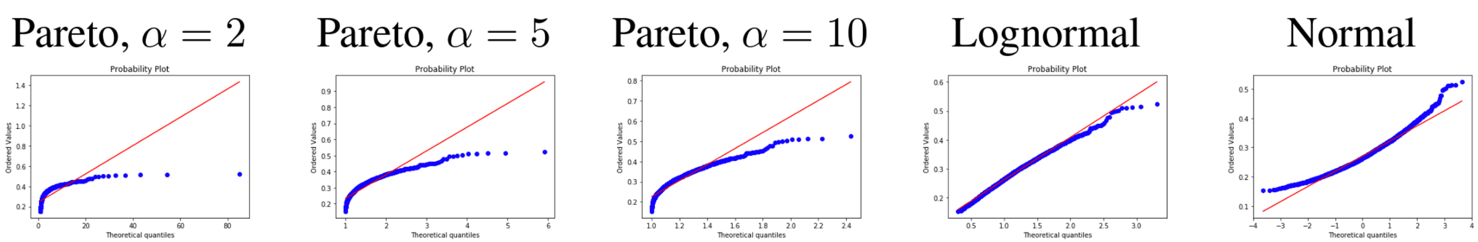


Figure A.7: Ablation Study, Corrupted FMNIST & LeNet: At the beginning

1.Introduction

- Despite the abundance of literature in analyzing stochastic optimization, not much work has been done that analyzes stochastic training process on the Graph Neural Network (GNN).

- Therefore, as the first step, we would like to tackle the following question:

What are the statistical properties of SGNs when we perform stochastic training of GNNs?

2. Problem Settings

2.Problem Settings

- Graph Neural Network (GNN) transforms node feature from the original graph. We can use transformed feature to various machine learning tasks.

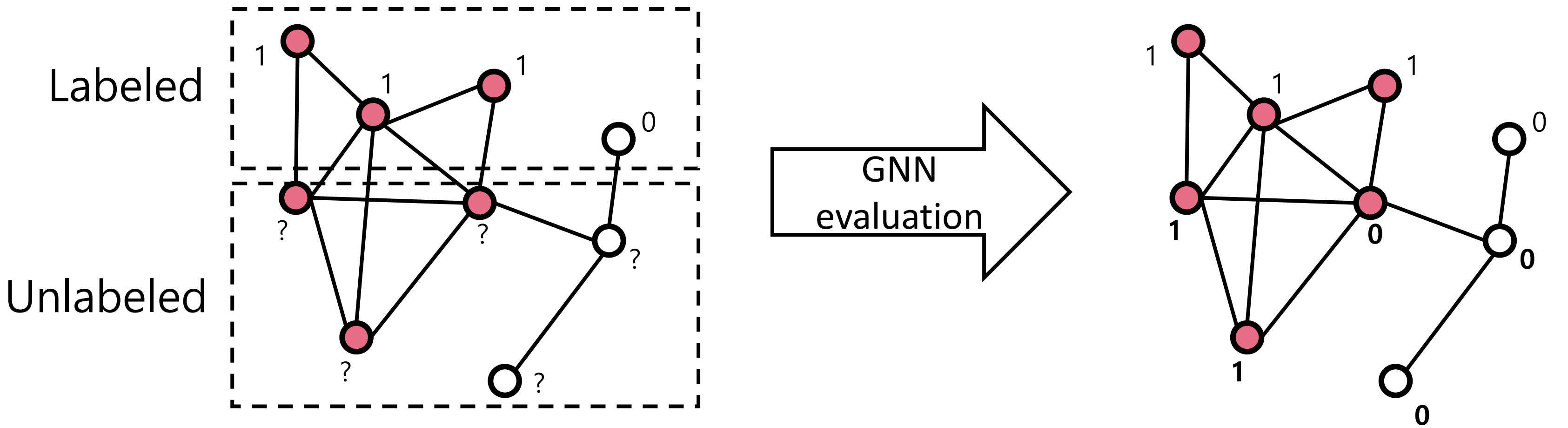
- Several aggregation schemes have proposed, including the famous two methods:

GCN: $$h_v' = RELU\left(\sum_{u \in N(v)} W \frac{h_u}{|N(v)|} + h_v \right)$$

GIN: $$h_v' = MLP_{\Phi}\left((1+\epsilon) \cdot MLP_f(h_v) + \sum_{u \in N(v)} MLP_f(h_u) \right)$$

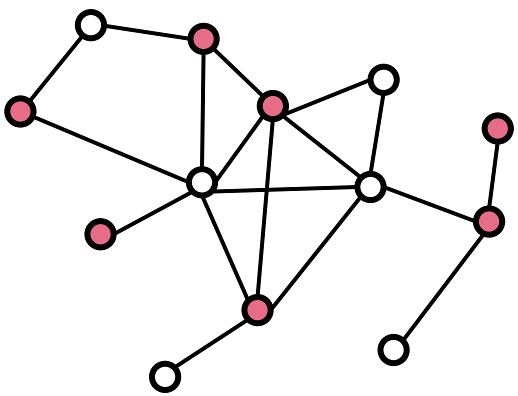
2.Problem Settings

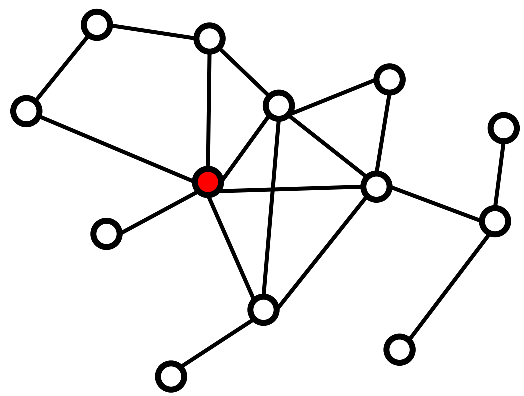
- We used the semi-supervised node classification setting: Given partial labeled vertices, we should predict the label of the left unlabeled vertices after training.

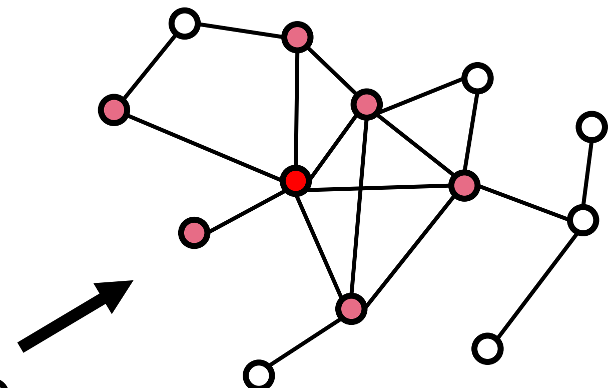


2.Problem Settings

- There are two methods of sampling nodes: (1) node batching and (2) neighborhood sampling.
- We consider uniformly random node batching *without* neighborhood sampling.
  - This is to isolate such additional effect







(1) Randomly select N

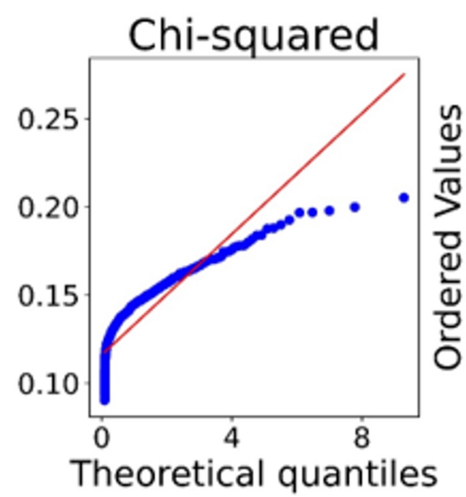
(2-1) Randomly select one

(2-2) Select neighbors

3. Experimental Settings

2.Problem Settings

- At prescribed epochs, we measured the norms of the SGNs, which are then formed from 1000 random batches.
  - We consider three epochs: beginning, middle, and end (see the paper for more details)
- We visualize the SGN norm distribution to some predefined distribution (e.g., normal, log normal, Pareto) using **QQ-plots**. [Wang et al., ICLR'22]
- We use the Cora dataset.



4. Results

4.Results (GCN)

Optimization and
Statistical Inference LAB

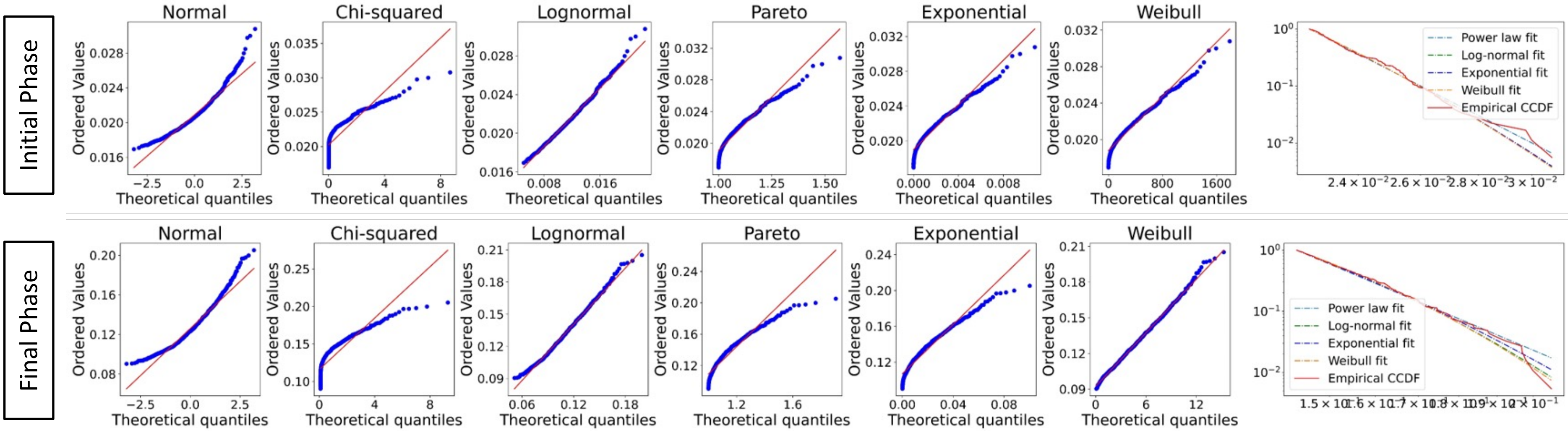


그림 1 (GCN)

- Normal and Pareto(Power law) distributions do *not* fit well, while Log Normal distribution has good fit. → This is in line with the observations for vision tasks [Wang et al., ICLR'22].
- Chi-squared distribution does not fit well.

4.Results (GIN)

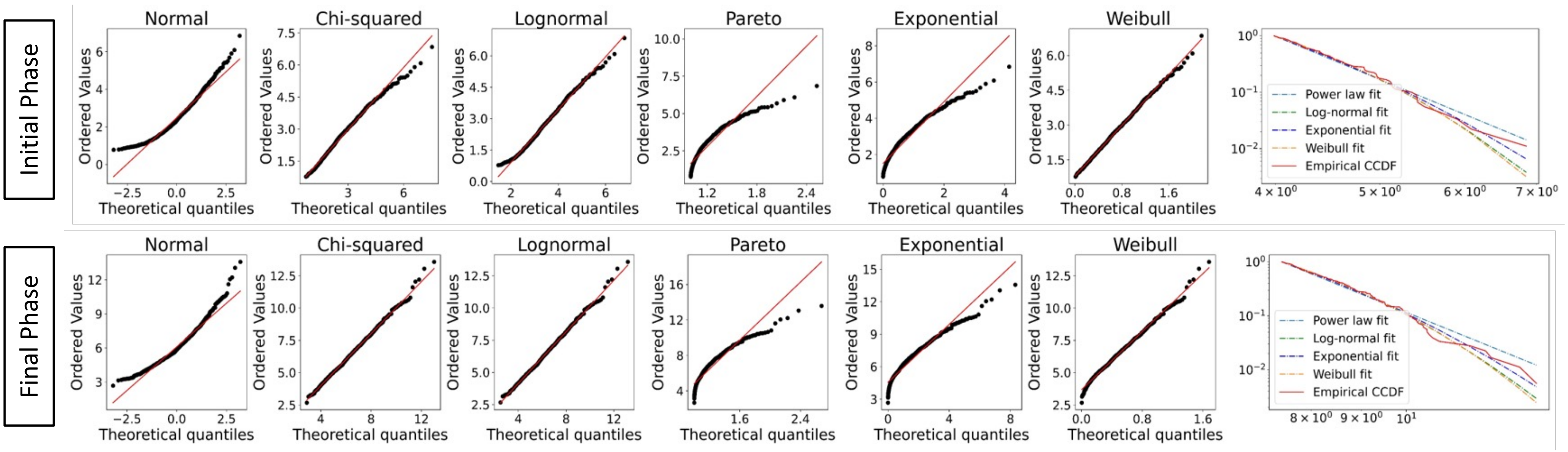


그림 2 (GIN)

- Similar as GCN i.e. normal and Pareto provide poor fits, while lognormal provides somewhat good fit.

Chi-squared distribution does fit well.

→ Chi-squared fitting distinguish the GCN and GIN.

Takeaways/Future Works

- We provide a preliminary statistical analysis of SGN of GNNs, following [Wang et al., ICLR'22].
- The statistical behaviors of SGD for GNNs are similar with that for the common vision tasks.
- According to chi-squared distribution "test", tail properties of GCN and GIN differ.
  - Is this reliable conclusion? Their behaviors are the same for normal, Pareto (in that those two do not fit well), and lognormal (in that this provides good fit)
- The most interesting future direction would be to see whether specific graph properties can be incorporated into the dynamics of SGD (e.g. degree distribution, graph topology…etc)
  - This has been recently done for distributed learning setting [Gurbuzbalaban et al., arXiv'22]

Thank you

Optimization and Statistical Inference LAB
yunseyoung@kaist.ac.kr
대전 유성구 대학로291 한국과학기술원
www.osi.kaist.ac.kr

References

1. Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *ICLR 2017*.
2. Smith, S., Elsen, E., and De, S. On the Generalization Benefit of Noise in Stochastic Gradient Descent. In *ICML 2020*.
3. Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit Bias of SGD for Diagonal Linear Network. In *NeurIPS 2021*.
4. Damian, A., Ma, T., and Lee., J. D. Label Noise SGD Provably Prefers Flat Global Minimizers. In *NeurIPS 2021*.
5. Welling, M., and Teh, Y. W. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *ICML 2011.*
6. Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic Gradient Descent as Approximate Bayesian Inference. In *Journal of Machine Learning Research* 18 (2017)
7. Simsekli, U., Sagun, L., and Gurbuzbalaban, M. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *ICML 2019*.
8. Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S., and E, W. Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning. In *NeurIPS 2020.*
9. Wang, X., Oh, S., and Rhee, C-H. Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise. In *ICLR 2022.*