Studying Distribution of Stochastic Gradient Noise in Deep Neural Networks

Diyuan Wu, Manjot Singh, Shuqi Wang
Department of Computer and Communication Science, EPFL, Switzerland
diyuan.wu@epfl.ch, manjot.singh@epfl.ch, shuqi.wang@epfl.ch

Abstract—**By invoking Central Limit Theorem (CLT) and assuming large mini-batch size, the gradient noise (GN) in the stochastic gradient descent (SGD) algorithm is often considered to be Gaussian. We follow (Simsekli et al., 2019) in this work and by conducting experiments on different deep learning architectures and datasets, we empirically show that the GN is highly non-Gaussian and converges to heavy-tailed α stable random variable. We also perform additional experimental study where we measure the distribution of eigenvalues in the Hessian matrix and through that, we observe the impact of gradient noise on the convergence of SGD towards the wider minima.**

I. Introduction

Stochastic gradient descent (SGD) is one of the most popular optimization methods for training deep neural networks because of its computational efficiency as compared to full gradient methods. The variance of stochastic gradients (gradient noise) have been shown to play an important role in optimization and generalisation in the context of deep networks (Keskar et al., 2017). In this project, we first study the distribution of gradient noise and then based on the GN, we reflect on the behaviour of SGD in terms of choosing wider minima. We investigate above behaviour for fully-connected and convolutional models using negative log likelihood loss function on MNIST and CIFAR10 datasets. For each configuration, we monitor the tail index (α) of GN distribution and eigenvalues of the hessian matrix for different depths, widths of the network running on different datasets.

In section II, we describe definitions and literature on α-stable distribution and the way to estimate the tail-index for these distributions. Section III mentions all the details regarding the experimental setup. Section IV and V discuss the results obtained and future work respectively.

II. Methodology

In this section, we define the gradient noise induced by SGD and briefly discuss α- stable distribution and how to estimate tail-index for these distributions. Then we briefly discuss how the GN distribution affects transition of SGD between narrow and wider minima.

A. SGD Noise and α-stable distribution

Despite the faster convergence of SGD, one specific trait of SGD is inherent noise which originates from the way the training points are sampled. The stochastic gradient noise is defined as $U_k(\mathbf{w}) = \nabla\tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})$, where $\nabla\tilde{f}_k(\mathbf{w})$ denotes the stochastic gradient at iteration k and $\nabla f(\mathbf{w}) = \mathbb{E}[\nabla\tilde{f}_k(\mathbf{w})]$. The noise is often considered to be Gaussian with $U_k(\mathbf{w}) \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, where $\mathcal{N}$ denotes the multivariate (Gaussian) Normal distribution and $\mathbf{I}$ denotes the identity matrix of appropriate size. The rationale behind this is that, if the size of the minibatch b is large enough, then the distribution of U_k is approximately Gaussian according to Central Limit Theorem (CLT). Figure1 in (Simsekli et al., 2019) shows the dissimilarity between the gradient noise norm computed using a convolutional neural network (AlexNet) and norms of Gaussian random variables. The shape of the histogram clearly shows a heavy-tailed behaviour. (Simsekli et al., 2019) mentions the scenarios (and references therein) where the CLT assumption fails such as turbulent motions, oceanic fluid flows, finance, biological evolution, audio signals etc. One can prove a generalized CLT and show that the law of the sum of these independent and identical distributed variables with infinite variance still converges to a family of heavy-tailed distributions that is called the α-stable distribution. These distributions are parameterized by their tail-index $\alpha \in (0, 2]$ and they coincide with the Gaussian distribution when $\alpha = 2$. As α gets smaller, the distribution shows a heavier-tail.

B. Estimating Tail Index and Measuring GN

Following (Simsekli et al., 2019), we use a relatively recent estimator for α-stable distributions. We refer the reader to Theorem 3 in (Simsekli et al., 2019) and further references therein for more details on convergence of this estimator. For measuring GN, we compute the full gradient $\nabla f(\mathbf{w})$ and the stochastic gradient $\nabla\tilde{f}_i(\mathbf{w})$ for each minibatch i. Then, the stochastic gradient noise is computed as $U^i(\mathbf{w}) = \nabla\tilde{f}_i(\mathbf{w}) - \nabla f(\mathbf{w})$, vectorize each $U^i(\mathbf{w})$ and concatenate them to obtain a single vector. The reciprocal of this estimator gives the α tail index.

C. GN Distribution and type of minima

Recent studies related to metastability analysis of Brownian motion-driven SDEs (Bov, 2004) (direct consequence of Gaussian Noise assumption) demonstrates that for SGD to escape any minima will depend on the depth or height of that minima. The Brownian motion becomes inappropriate when the noise is non-Gaussian. So, authors in (Simsekli et al., 2019) study α-stable Levy motion whose increments have α-stable distribution. Theory behind Levy-driven SDEs implies that as α gets smaller, the process will escape narrow minima and will stay longer in a wider minima. In other words, probability for

TABLE I
THIS TABLE SHOWS DIFFERENT ARCHITECTURES, DIFFERENT DATASETS AND DIFFERENT CONFIGURATION ALONG DEPTH AND WIDTH DIRECTION.

Model Architecture	Depths	Widths	Dataset
FCN	2,4,8	64,128,256	MNIST
AlexNet	8	4,16,32	CIFAR10
Residual Net	8, 14, 20	width const	CIFAR10

the SGD to jump in a wide basin will increase. We argue that empirically, the process of optimization with SGD favors flatter landscape by showing that the spectrum of the Hessian has many near zero eigenvalues and few large eigenvalues at a near critical point with close to zero loss value.

D. The spectrum of Hessian matrix

We study the spectrum of the Hessian matrix of different neural network models. Note that when the training converges some point w_0, the norm of the gradient $\nabla f(w_0)$ is close to 0, then we have the following second order approximation for the objective function:

$$\begin{aligned} f(w) &= f(w_0) + \nabla f(w_0)(w - w_0) \\ &\quad + (w - w_0)^T \nabla H(w_0)(w - w_0) \\ &\approx f(w_0) + (w - w_0)^T \nabla H(w_0)(w - w_0) \end{aligned}$$

If all eigenvalues are positive, we know that the Hessian is positive semidefinite, then the $(w - w_0)^T \nabla H(w_0)(w - w_0)$ is always larger or equal to 0, which means w_0 is a local or global minima.

However, if the Hessian have both positive and negative eigenvalues, assume $\lambda < 0$, with the corresponding eigenvector v, then if we let $w = w_0 + v$, we have: $f(w) = f(w_0) + \lambda \|v\|^2 < f(w_0)$, which means we converge to a saddle point rather than a minima.

III. EXPERIMENTS

The goal of our experiments is to investigate the tail behaviour of the GN and measure the hessian eigenspectrum in the variety of scenarios. Table I mentions different model architectures, datasets and network configurations used for this study while table II indicates batch size, learning rate and number of training iterations used for each configuration.

The loss function used for each configuration is Negative Loss Likelihood (i.e. Cross-Entropy). The training algorithm is SGD with 0 momentum and 0 weight decay unless otherwise specified. The training runs until 100% training accuracy is achieved or until maximum number of iterations limit is reached (the latter limit is effective in the underparametrized models and maximum number of iterations are picked with performance and computational costs considered). At every 100th iteration, we log the full training and test accuracies, and the tail estimate of the gradients that are sampled using the corresponding mini-batch size.

In addition to above observations, we also compute the hessian spectrum at the end of training so as to check for wide critical points by inspecting the eigenvalues of Hessian matrix.

TABLE II
PARAMETERS FOR TRAINING

Parameter for training			
Model Architecture	FCN	Alexnet	Resnet
Batch size	100	100	100
Learning rate	0.1	0.1	0.1
Steps	5,000	25,000	25,000

Codebase - We run all our experiments on Google Colab. We used the source code provided by authors but modified it to our own requirements. The source code for this project is attached along in a zip file. The code for computing the Hessian information was taken from (Yao et al., 2020) [1].

IV. RESULTS AND DISCUSSION

In this section, we present the important and representative results.

A. Tail Index Results

Similar to (Simsekli et al., 2019), for our experiments, we compute the average of the tail-index measurements for the last $1,000$ iterations in the case of FCN (i.e. when α becomes stationary) so as to focus on the late stage dynamics whereas we do the average for the last 10,000 iterations in the case of AlexNet and ResNet.

Figure1 A) and B) shows tail-index measurements for FCN with varying widths and depths on MNIST dataset. It is observed that in all the cases, the tail index is far from 2 which shows that the distribution of the gradient noise is non-Gaussian. We observe that in general, for fully connected networks, the tail index decrease as the depth is increasing. For the case of depth = 8, the α decreases with increasing width.

Figure1 C) shows results for CNN (AlexNet) on CIFAR-10. The results show that α is very low for smaller width, increases with the width of the network but is still far from 2 indicating non Gaussian behavior of noise norm. We also observe that α behaves similarly for both trained and test sets.

The results for Resnet on CIFAR-10 are shown in Figure1 D). The value of α decreases as the depth of the network is increased from 8 to 20. In this case as well, α has similar values for both trained and test sets.

In conclusion, we observe that the:

- The α-value of each network architecture is bounded away from 2, which means the distribution of the gradient noise is non-Gaussian.
- These results show a strong interplay between the network architecture and the dataset used. For example, for FCN, the α-value decreases when the width of the networks is increasing, while for the Alexnet, it is the opposite.

[1] Code for Hessian Computation https://github.com/amirgholami/PyHessian

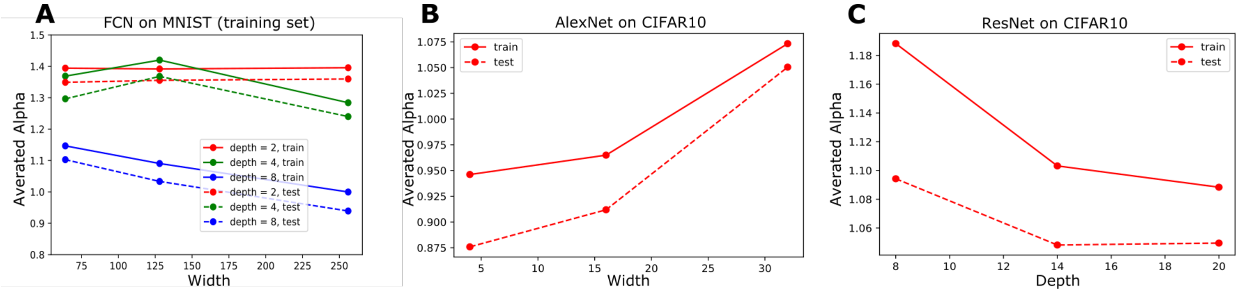

Fig. 1. **Strong interplay between the network architecture, dataset and the algorithm dynamics. A)** Estimation of α for varying width and depth of FCN on training set. **B)** Same as **A)** but on testing set. **C)** Same as **A)** but for AlexNet on both training and testing set. **D)** Same as **A)** but for ResNet on both training and testing set.

B. Hessian Results

Due to space constraint, we show the Hessian spectrum observed at the end of training only for a particular configuration of FCN, AlexNet and Resnet. Figure 2 show the corresponding plots respectively.

Figure2 A) shows the density of eigenvalues of the Hessian matrix for FCN with width 64 and depth 4. Note that for this experiment, we train the FCN for 20,000 steps instead of 5,000 in order to have better convergence. We observe that the eigenvalues are almost non-negative, which implies that the training converges to a wide minima (might be a local or global minima), rather than a saddle point.

Figure2 B) show the density of eigenvalues of the Hessian matrix for Alexnet with scale 32. We observe that the eigenvalues have both positive and negative value, which implies that the training converges to a saddle point. Figure2 C) shows results in the case of ResNet20. Similar to FCN, we observe that the eigenvalues are almost non-negative implying convergence to a wider minima.

From all the 3 plots, we observe that majority of the eigenvalues are close to 0 with only a few eigenvalues having positive and negative value. Therefore, according to our reasoning in subsection II-C, SGD ends up in a wider basin with high probability in the case of FCN and ResNet whereas for AlexNet, convergence to critical point doesn't yield conclusive results on the type of the minima.

This viewpoint also shows that the choice of the dataset and architecture also have a significant impact on landscape of the problem.

V. Conclusion and Future Work

In our study, we investigated the tail behavior of the gradient noise in deep neural networks and empirically showed that the gradient noise is highly non-Gaussian. We also showed that SGD ends up in a wide basin on the basis that most of the eigenvalues are 0. We also study the spectrum of Hessian and we observe the properties of the convergent point (whether it is a minima or a saddle point) from it. This also poses interesting questions. For instance, the dependence of GN on algorithm parameters (i.e. step-size, momentum, batch size) is yet unclear and needs to be investigated.

References

Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 006(4):399–424, 2004.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. 2017.

U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning, (ICML)*, 2019.

Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney. Pyhessian: Neural networks through the lens of the hessian, 2020.

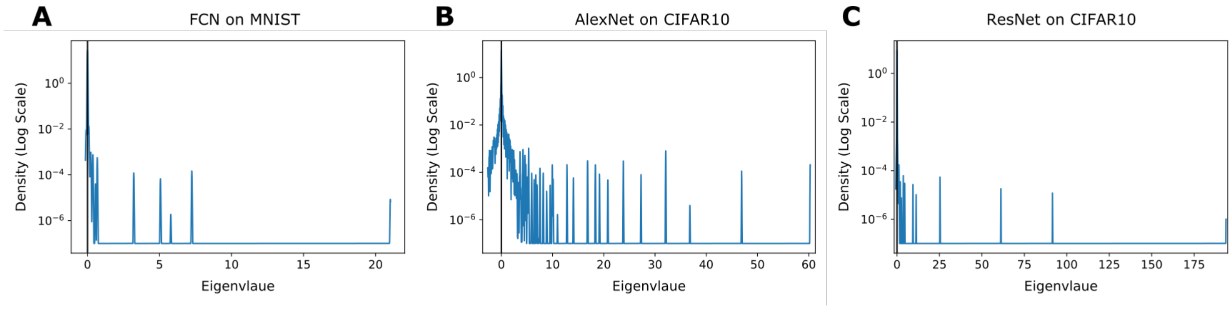

Fig. 2. **Eigenvalue spectrum density plot. A)** Results of FCN (width=64, depth=4) trained on MNIST. **B)** Results of AlexNet (width=32) trained on CIFAR10. **C)** Results of ResNet (depth=20) trained on CIFAR10.