Adaptive Sharpness and Generalization

Hadi Hammoud, Orfeas Liossatos, Léo Nicollier
EPFL, Switzerland

***Abstract*—Recent research has emphasized the appeal of incorporating the sharpness of a training objective at its minimum. Flat minima in the loss surface tend to yield consistently low loss values despite minor disturbances from variations between training and testing data (1). While simple sharpness definitions have shown a positive correlation with test loss (negative correlation with generalization) (2), concerns exist regarding their sensitivity to reparametrizations (3). The effectiveness of adaptive sharpness in modern transformer models is now being questioned (4). In this study, we independently verify these results by measuring the adaptive sharpness on three diverse architectures trained from scratch: LeNet-5, a graph attention network, and a fully connected networks. In opposition to (3), we observe consistently negative correlation between adaptive sharpness and test loss, challenging the geometric intuition and suggesting that other factors may play a role in generalization beyond sharpness.**[1]

I. Introduction

The concept of incorporating the sharpness of a training objective at its minimum holds great intuitive appeal. It stems from the observation that when the loss surface encounters minor disturbances resulting from variations between training and testing data or the presence of out-of-distribution (OOD) inputs, networks characterized by flat minima are expected to exhibit consistently low loss values (1).

Empirical evidence has supported this intuition; mainly, it has been shown that, for simple definitions of sharpness, flat regions correlate positively with generalization (2). However, these definitions have come under scrutiny as they can be sensitive to reparametrizations (3). Using adaptive sharpness provably fixes the reparametrization problem, but a mounting body of evidence suggests that adaptive sharpness no longer correlates with generalization in modern transformer models (4). These counter-intuitive results motivate us to independently verify and gather empirical evidence from more settings.

To this end, our aim is to apply the adaptive sharpness definition to a diverse range of model architectures and sizes. Through this approach, our primary goal is to validate and corroborate the findings presented in previous research papers that delve into the question of whether a correlation exists between sharpness and generalization.

II. Background

A. Adaptive sharpness and its invariance

We begin by working towards a definition of sharpness. Data-label pairs come from a dataset $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, which is split into a training set $\mathcal{S}_{tr}$ and a test set $\mathcal{S}_{te}$. Given a model f with weights $\boldsymbol{w} \in \mathbb{R}^D$ and a data-label pair $\boldsymbol{z} \in \mathcal{Z}$, let $l(\boldsymbol{w}, \boldsymbol{z}) \in \mathbb{R}_+$ be a loss function such as cross-entropy or mean-squared error. Then if $\mathcal{S} \sim P^m$ is a batch of m pairs picked independently and uniformly from $\mathcal{S}_{tr}$, then $L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{m}\sum_{z\in\mathcal{S}} l(\boldsymbol{w}, \boldsymbol{z})$ is the loss on that batch. We define *adaptive worst-case m-sharpness* as the expected maximum difference of training-batch losses between two models that are at least ρ-close in p-norm.

$$S^{\rho}_{max}(\boldsymbol{w}, \boldsymbol{c}) := \mathbb{E}_{\mathcal{S}\sim P^m} \max_{||\boldsymbol{\delta}\odot\boldsymbol{c}^{-1}||_p \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w})$$

If we take the element-wise absolute value of the weights $\boldsymbol{c} = |\boldsymbol{w}|$, then this sharpness definition becomes invariant to positive multiplicative weight reparametrizations that preserve the model output. Indeed, for any $r \in \mathbb{R}^D_+$ such that $f(\boldsymbol{w} \odot \boldsymbol{r}) = f(\boldsymbol{w})$, we have

$$\begin{aligned}
&S^{\rho}_{max}(\boldsymbol{w}\odot\boldsymbol{r}, |\boldsymbol{w}\odot\boldsymbol{r}|)\\
&= \mathbb{E}_{\mathcal{S}} \max_{||\boldsymbol{\delta}\odot|\boldsymbol{w}\odot\boldsymbol{r}|^{-1}||_p \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{r}+\boldsymbol{\delta}) - L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{r})\\
&= \mathbb{E}_{\mathcal{S}} \max_{||\boldsymbol{\delta'}\odot|\boldsymbol{w}|^{-1}||_p \leq \rho} L_{\mathcal{S}}((\boldsymbol{w}+\boldsymbol{\delta'})\odot\boldsymbol{r}) - L_{\mathcal{S}}(\boldsymbol{w}\odot\boldsymbol{r})\\
&= \mathbb{E}_{\mathcal{S}} \max_{||\boldsymbol{\delta'}\odot|\boldsymbol{w}|^{-1}||_p \leq \rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta'}) - L_{\mathcal{S}}(\boldsymbol{w}) = S^{\rho}_{max}(\boldsymbol{w}, |\boldsymbol{w}|)
\end{aligned}$$

where $\boldsymbol{\delta'} := \boldsymbol{\delta} \odot \boldsymbol{r}^{-1}$. So we indeed take $\boldsymbol{c} = |\boldsymbol{w}|$ for our experiments. Alternatively, r could be negative and the invariance still holds for even p norms. Multiplicative weight reparametrizations are precisely those that could leave our models unchanged. For example, in LeNet-5, we could (1) scale convolutional parameters by any $r \in \mathbb{R}$ and know that the following batch normalization layer would produce the same output, or (2) introduce scaling before and after a ReLU layer since $\mathrm{ReLU}(rz)/r = \mathrm{ReLU}(z)$.

B. Estimating adaptive sharpness using PGA

In order to estimate the adaptive worst-case m-sharpness of the training loss around a model with weights $\boldsymbol{w}$, we randomly select a batch $\mathcal{S}$ of m training samples and use projected gradient ascent (PGA) to maximize the training loss while projecting the weights onto the p-norm ball of radius ρ after every gradient update. We perform 5 gradient updates and save the highest training loss $L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\delta})$ before restarting with another $\mathcal{S}$. We restart a total of 50 times in order to approximate the adaptive worst-case sharpness. We use the same learning rate as the one for training the models. Note that PGA will occasionally fail and produce a better model rather than a worse one.

[1]Python notebooks are available at https://github.com/HadiHammoud44/CS439_Sharpness

C. Kendall's rank correlation coefficient

We use Kendall's rank correlation coefficient τ to capture the relationship between the adaptive sharpness $s_1, \ldots, s_M$ and test loss $t_1, \ldots, t_M$ for a set of M models of the same architecture and dataset. Kendall's τ is given by

$$\tau = \frac{2}{M(M-1)} \sum_{i<j} \operatorname{sign}(t_i - t_j) \operatorname{sign}(s_i - s_j).$$

The advantage of this coefficient over covariance is that is can detect any monotonic relationship between two variables without assuming linearity, and it is robust to outliers. It is also just as easy to interpret, with $\tau = 1$ and $\tau = -1$ meaning positive and negative relationships respectively.

III. Models & Methods

We investigate the relationship between adaptive worst-case sharpness and test loss for three different model types and datasets by training the models from scratch. The particular treatment of each model and dataset before training is described later. For each model type, after normalizing its features, we train a pool of 50 models from Xavier-initialized weights with 64-batch SGD until the relative difference of training loss dips below 5% or until 10 epochs have been reached. The weights $\boldsymbol{w}$ after the last epoch are saved for computing sharpness. From there, we estimate the adaptive worse-case sharpness using PGA. For our experiments, we let $m = 64$ and $p = \infty$, following (4), while we try different values for ρ. The loss functions and learning rates for each model are summarized in Table I. Finally, we compute Kendall's rank correlation coefficient between the estimated sharpness and test loss.

TABLE I

Model	Learning Rate	Loss Function	ρ
FCNN	0.014	Mean Squared Error	0.025
LeNet-5	0.05	Cross-Entropy	0.0001
GAT	0.236	Cross-Entropy	0.02

A. Fully Connected Neural Network Model — Abalone

The Abalone dataset (5) is a dataset with 4177 data-points and 11 features. It consists of measurements of physical attributes of abalone, a type of marine snail. These attributes are male, female, infant, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight and number of rings. The target variable in this dataset is the age of the abalone, which is typically represented as the number of rings. The Abalone dataset was divided into a training set and a test set using a ratio of 0.8 for training and 0.2 for testing. For the model, a fully connected neural network (FCNN) architecture was employed, incorporating rectified linear unit (ReLU) activation functions. The network architecture consisted of four layers with node configurations of 10, 64, 64, and 1 respectively.

B. LeNet-5 — FashionMNIST

The FashionMNIST dataset (6) is a widely used benchmark that serves as an alternative to the classic MNIST handwritten digits dataset, offering a more challenging task of classifying 10 different categories of fashion items. The dataset consists of 60,000 training examples and 10,000 testing examples, with grayscale images of size 28x28 pixels.

As for training, we used the LeNet-5 model (7). LeNet-5 is a classic convolutional neural network architecture used for recognition tasks. It achieves the task by first applying two convolutional layers with 5x5 kernels, followed by batch normalization. Max pooling is then used to reduce the spatial dimensions of the output. The resulting feature maps are flattened and fed into fully connected layers. After the first two fully connected layers, dropout regularization with a rate of 0.5 is applied to prevent overfitting. The final fully connected layer produces logits for the 10 different digit classes. ReLU activation is employed throughout the model to introduce non-linearity.

C. Graph Attention Network (GAT) — Cora

The Cora dataset is a single graph with 2708 nodes representing scientific papers and 5728 edges representing citations (8). Each node is a binary vector of length 1433, indicating the presence or absence of a dictionary word in the paper. There are 7 node classes corresponding to different fields of research. We use a balanced set of 140 nodes for training, and evaluate on 1000 test nodes. We do not transform the data before training. As for the model, we use a Graph Attention Network (GAT) (9) with 8 hidden channels and 2 message passing layers with 0.6 dropout probability and ELU activations. Due to the slow start of training, the training data was traversed for a minimum of 10 epochs. Training was stopped with the relative difference of training loss criterion.

IV. Results

Figure 1 shows the scatter plots representing the test loss versus sharpness values obtained for all trained models. We observe negative τ coefficients of -0.311, -0.180 and -0.185 with small p-values, representing the probability of obtaining results at least as extreme as ours as under the hypothesis that sharpness and test loss are uncorrelated. This negative correlation suggests that higher adaptive sharpness could even correspond to improved generalization in these particular models, in line with the observations made in (4).

V. Conclusion

In conclusion, our study explored the relationship between sharpness and generalization in deep learning models by applying the adaptive sharpness definition to diverse models. Our findings have revealed a negative correlation pattern, which contradicts the previously established positive correlation mentioned in (3). Consequently, these results suggest that there may not be a significant correlation between the variables under investigation, and it is likely that other factors play a more influential role.

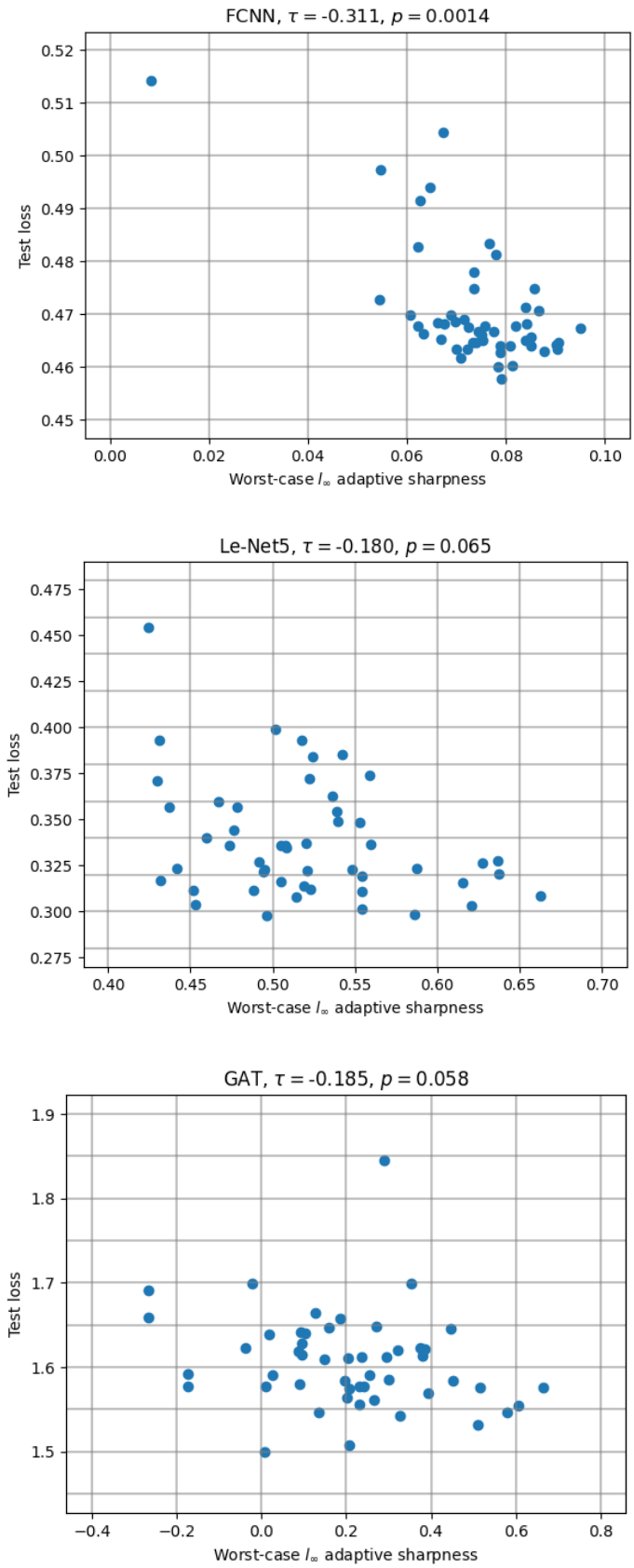

Fig. 1: For 50 diverse instances each of FCNN, LeNet-5, and GAT, we compute the Kendall coefficient τ and significance p measured between test loss and worst-case l_∞ adaptive sharpness at ρ values from Table I.

Overall, these findings invite us to rethink the response of the loss surface to variations between training and testing data and the conventional geometric intuition behind it. Further research is warranted to uncover why sharpness can appear to correlate positively or negatively with test loss between different settings.

References

[1] S. Hochreiter and J. Schmidhuber, "Simplifying neural nets by discovering flat minima," in *Conference on Neural Information Processing Systems (NeurIPS)*, 1995, pp. 529–536.

[2] 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," in *International Conference on Learning Representations (ICLR)*, 2016.

[3] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning (ICML)*, 2021.

[4] M. Andriushchenko, F. Croce, M. Müller, M. Hein, and N. Flammarion, "A modern look at the relationship between sharpness and generalization," *arXiv preprint arXiv:2302.07011*, 2023.

[5] S. T. T. S. C. A. Nash, Warwick and W. Ford, "Abalone," 1995.

[6] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the construction of internet portals with machine learning," *Information Retrieval*, vol. 3, no. 2, pp. 127–163, 2000. [Online]. Available: https://doi.org/10.1023/a:1009953814988

[9] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017. [Online]. Available: https://arxiv.org/abs/1710.10903