# On the Dubious Relationship between Generalization and the Maximum Hessian Eigenvalue

Simran Kaur ● Jeremy Cohen ● Zachary C. Lipton | Carnegie Mellon University

## Introduction

- Training interventions, such as batch size and learning rate, impact generalization. *But, why?*

  Popular belief: these interventions boost generalization by guiding the training process towards “flat minima” (the opposite of which are sharp minima).

  Flat minima broadly refer to solutions with favorable geometric properties.

- Recent work proposed the sharpness-aware minimization (SAM) algorithm, which directly optimizes for flatness [1].
- Common flatness metric: the leading eigenvalue of the Hessian of the training loss ($\boldsymbol{\lambda}_{\mathbf{max}}$)
  Smaller $\boldsymbol{\lambda}_{\mathbf{max}}$ correspond to flatter minima.
- Dinh et al. [2] previously showed that one can make $\boldsymbol{\lambda}_{\mathbf{max}}$ arbitrarily large without harming generalization.

## Contributions

- We present further evidence that calls into question the influence of $\boldsymbol{\lambda}_{\mathbf{max}}$ on generalization

> We can control $\boldsymbol{\lambda}_{\mathbf{max}}$ and find that small $\boldsymbol{\lambda}_{\mathbf{max}}$ do not always improve generalization.

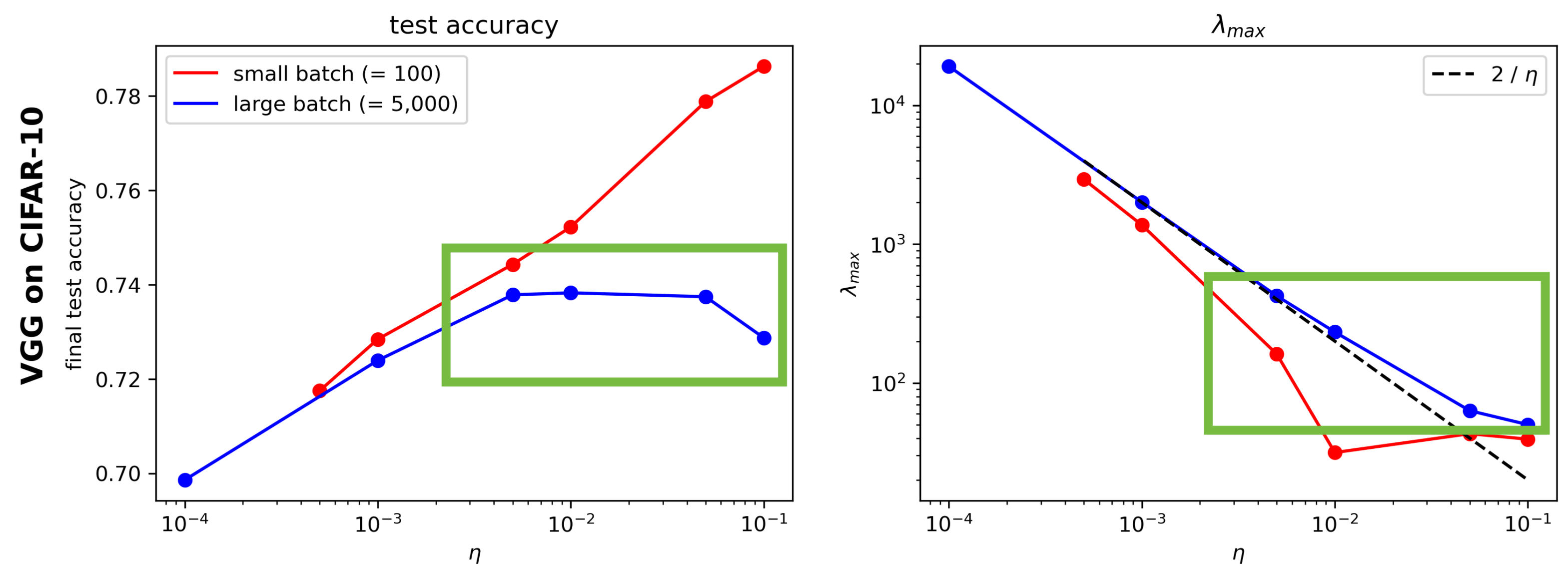
> We can boost generalization without promoting smaller $\boldsymbol{\lambda}_{\mathbf{max}}$.

- $\boldsymbol{\lambda}_{\mathbf{max}}$ does not provide a scientific explanation for improvements in generalization
- We hope to inspire future efforts aimed at understanding the relationship between flatness and generalization

## Experiments

### 1. Small Batch vs. Large Batch SGD in DNNs

We train a VGG11 on CIFAR-10 using SGD with a fixed learning rate and with cross-entropy (CE) loss.

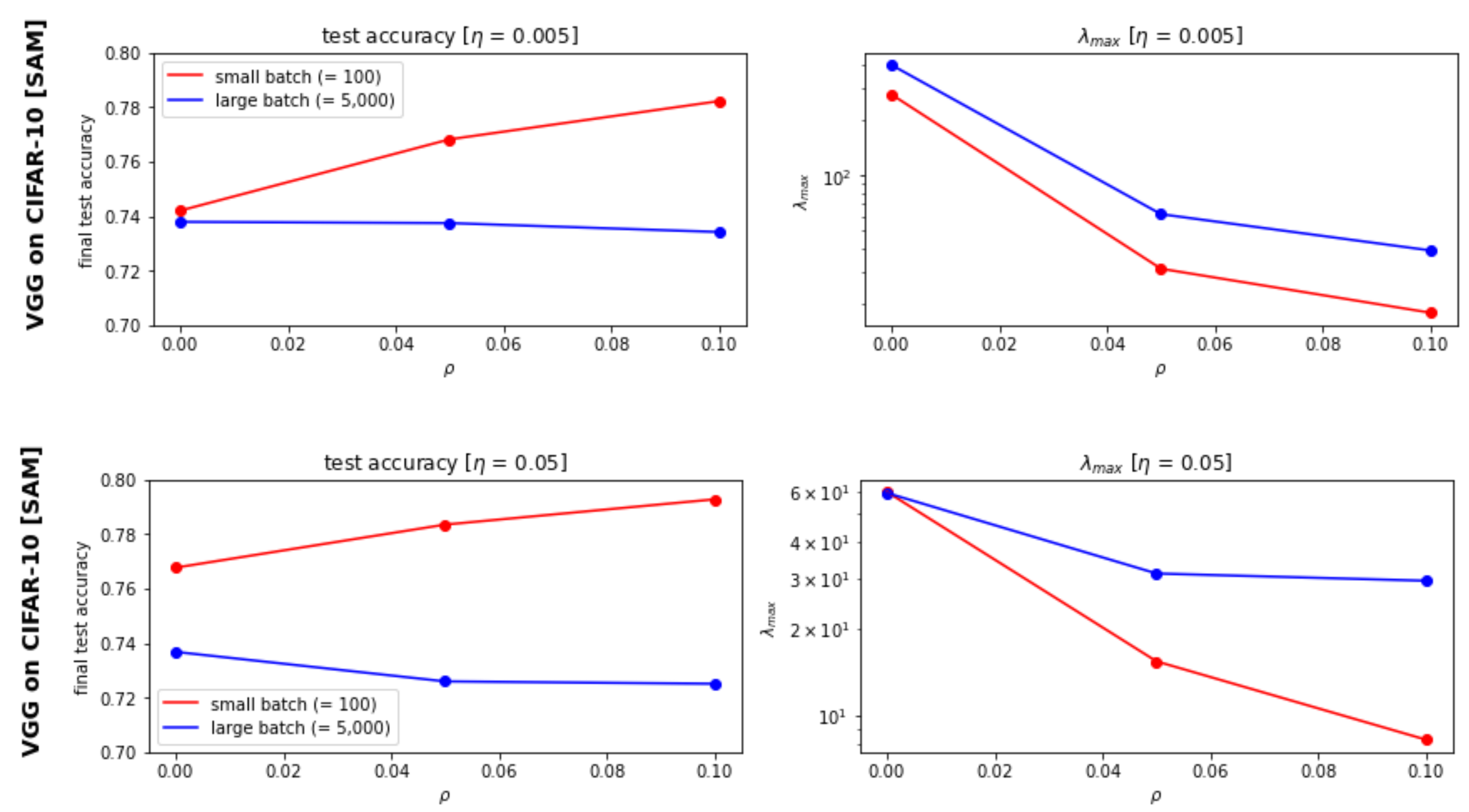



We observe that small batch SGD often exhibits generalization benefits from large learning rates, yet large batch SGD does not. This is consistent with previous work [3].

Large learning rates induce smaller $\boldsymbol{\lambda}_{\mathbf{max}}$, regardless of batch size.

### 2. Sharpness-Aware Minimization (SAM)

We train a VGG11 on CIFAR-10 using SGD with a fixed learning rate, CE loss, and the SAM training objective, which directly optimizes for flatness (according to a certain definition of flatness).

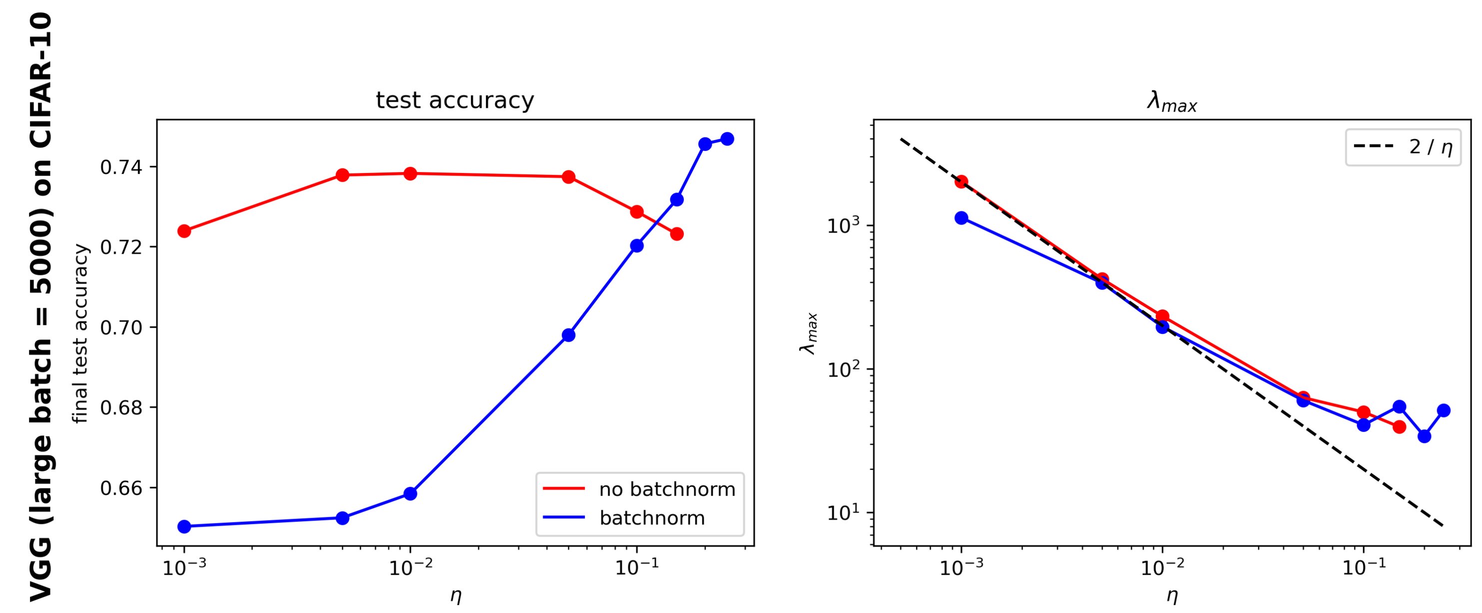



We observe that a higher $\rho$ (sharpness penalty) causes the test accuracy to be higher in the small batch setting and lower in the large batch setting.

In both cases, a higher $\rho$ induces smaller $\boldsymbol{\lambda}_{\mathbf{max}}$.

### 3. Batch Normalization in DNNs

We train a VGG11 (+ BN) on CIFAR-10 using SGD with a fixed learning rate and CE loss.

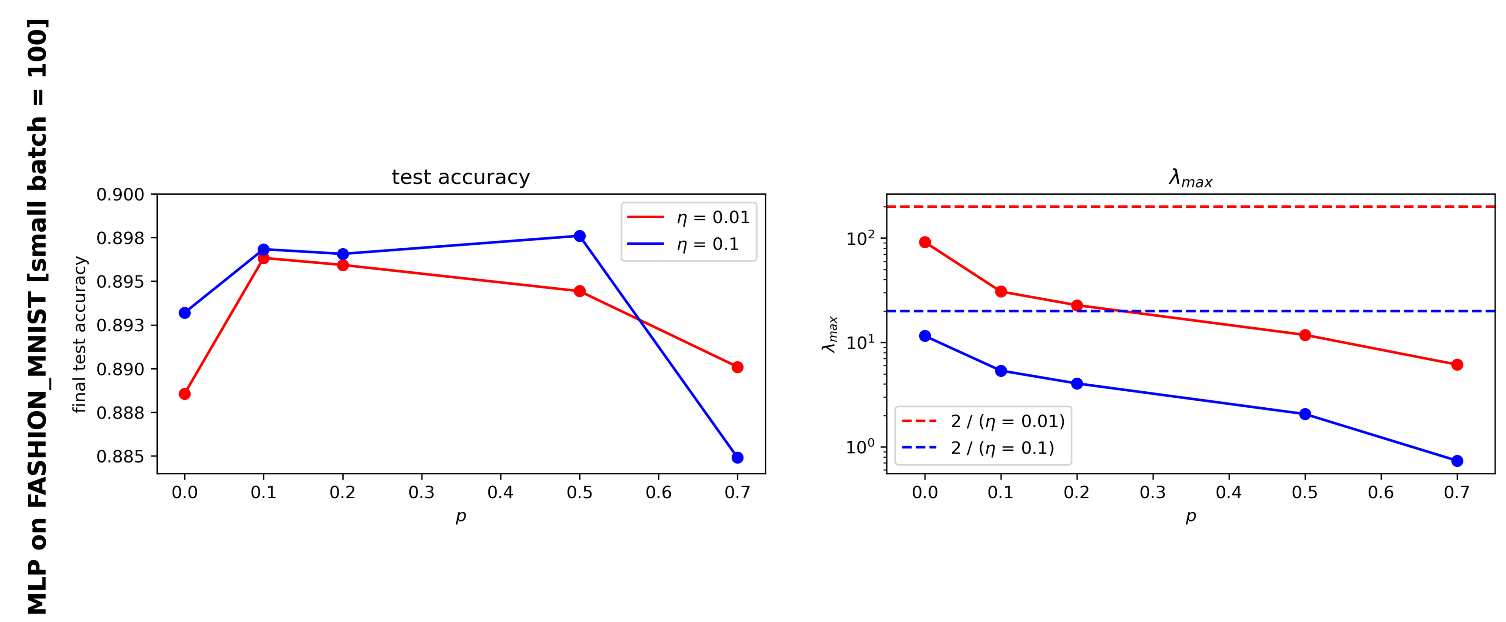



At large learning rates, models exhibit generalization benefits from BN.

For a fixed learning, $\boldsymbol{\lambda}_{\mathbf{max}}$ found by models with and without BN are comparable in the large batch regime.

### 4. Dropout in DNNs

We train an MLP (with 2 hidden layers and dropout) on FASHION_MNIST with a fixed learning rate and CE loss.




For a fixed learning rate and batch size, models with some dropout generalize better.

Higher dropout probabilities promote flatter solutions.

Yet, excessively high dropout probabilities do not generalize better.


#### References

[1] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *In International Conference on Learning Representations*, 2021.

[2] Dinh, L., Pascanu, R., Bengio, S., Bengio, Y. Sharp minima can generalize for deep nets, 2017.

[3] F. He, T. Liu, and D. Tao, “Control batch size and learning rate to generalize well: Theoretical and empirical evidence,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1141–1150.