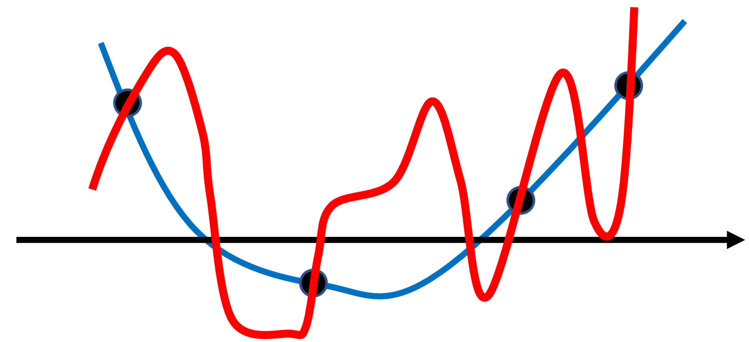
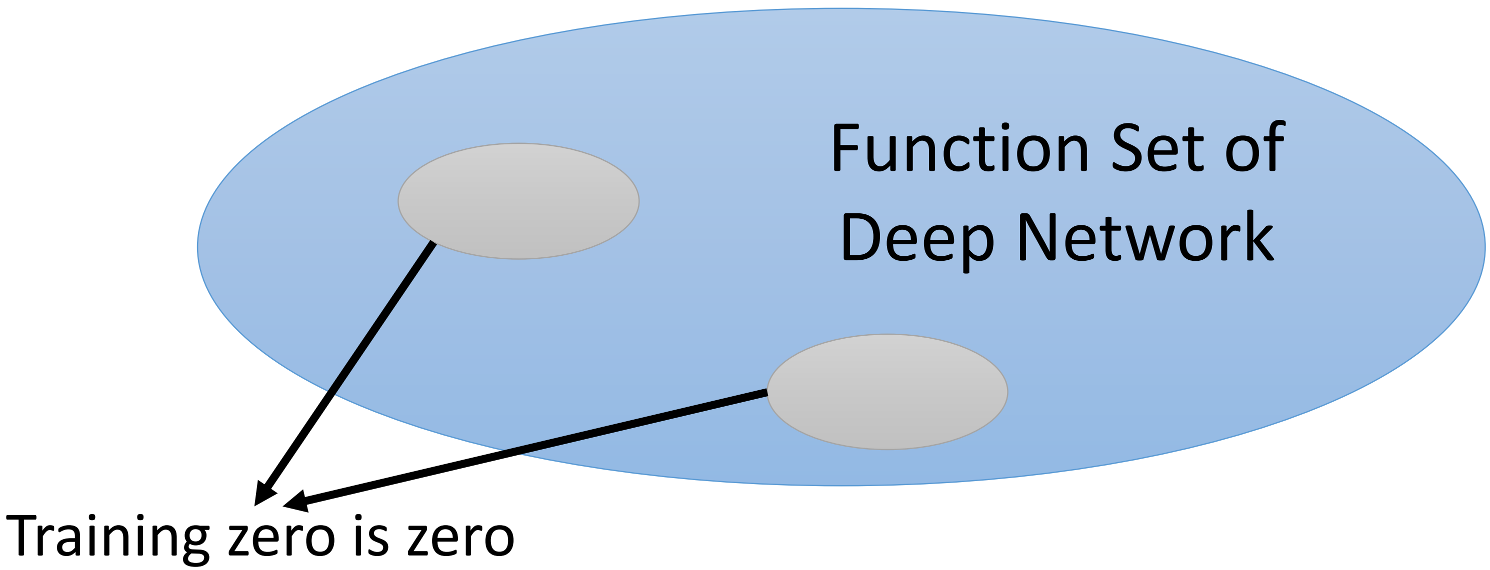
Indicator of Generalization

Introduction





- If many global optimums can zero training errors, which one can obtain generalized results?
- Use the indicator to find solution that generalizes well.
- ***Sharpness*** and ***Sensitivity***

Brute-force Memorization ?

- Real labels v.s. random labels

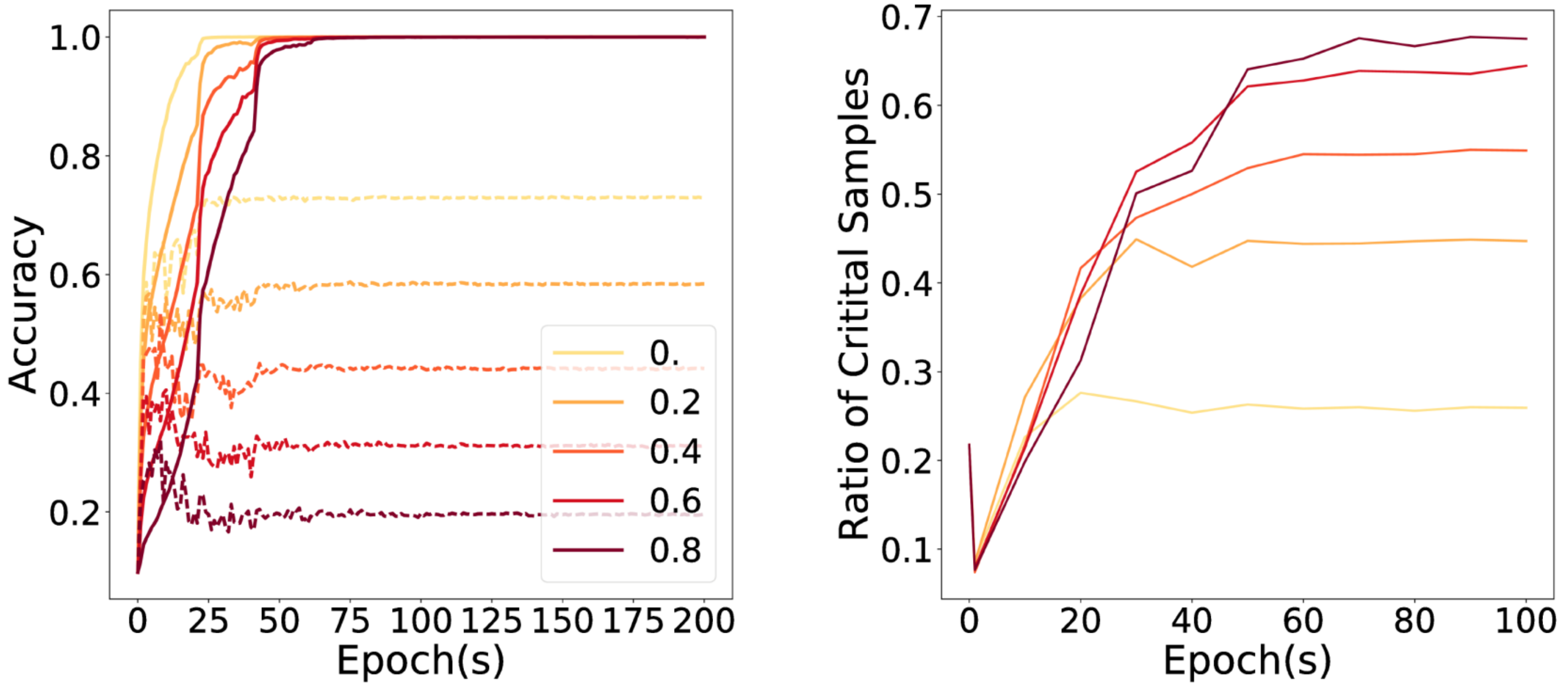
First layer of CIFAR-10

https://arxiv.org/pdf/1706.05394.pdf

https://arxiv.org/pdf/1706.05394.pdf

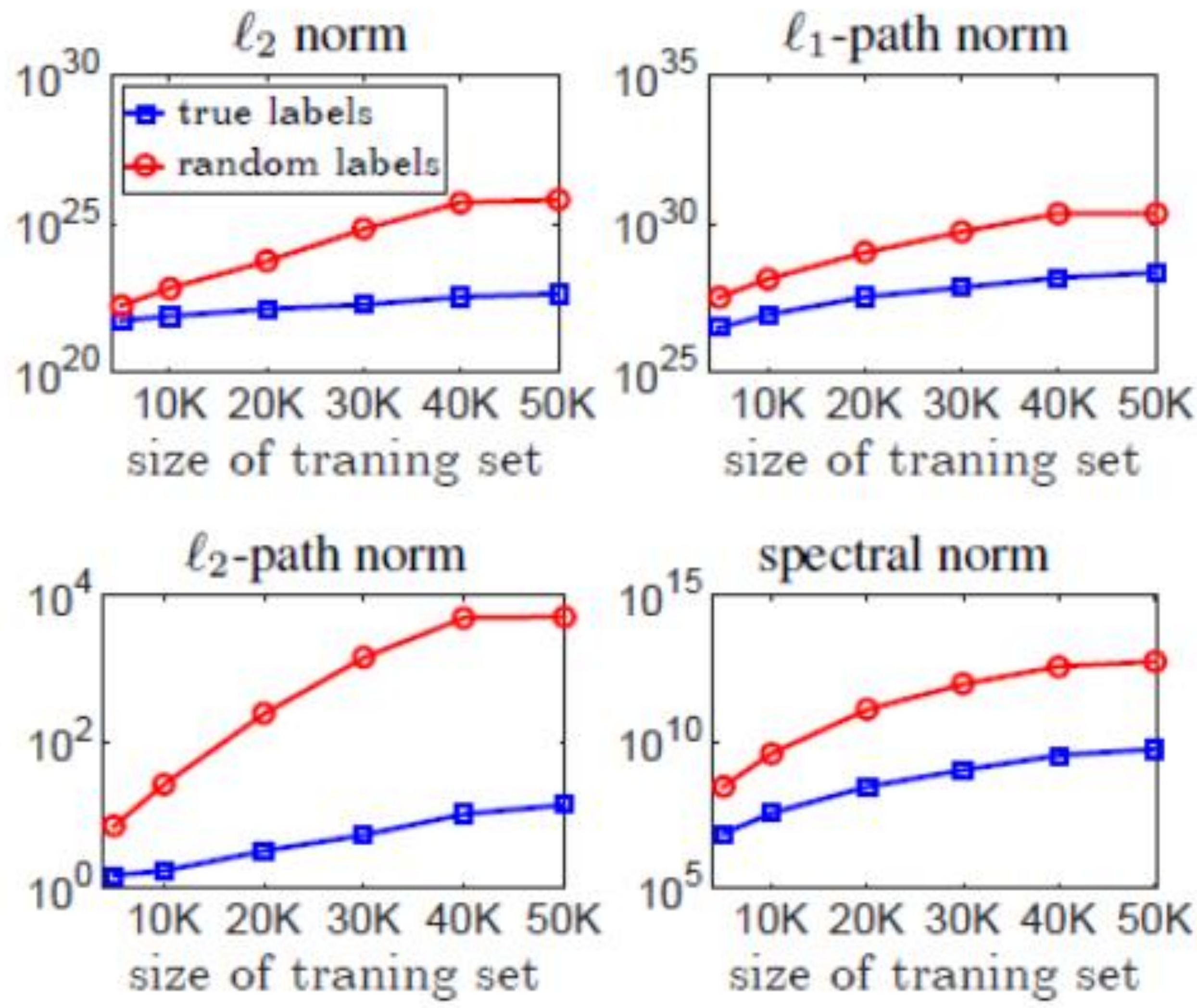
Brute-force Memorization ?

- Simple pattern first, then memorize exception



(b) Noise added on classification labels.

Brute-force Memorization ?



Sensitivity

Jacobian Matrix

$$y = f(x) \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\frac{\partial y}{\partial x} = \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{size of x}} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{size of y}$$

<u>Example</u>

$$\begin{bmatrix} x_1 + x_2 x_3 \\ 2x_3 \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) \qquad \frac{\partial y}{\partial x} = \begin{bmatrix} \qquad\qquad\qquad \end{bmatrix}$$

Sensitivity

- Given a network f, the sensitivity of a data point x is the Frobenius norm of the Jacobian

$$y = f(x) \qquad \frac{\partial y}{\partial x} = \begin{bmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \partial y_1/\partial x_3 \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 & \partial y_2/\partial x_3 \end{bmatrix}$$

$$\text{Sensitivity of x} = \sqrt{\sum_i \sum_j \left(\frac{\partial y_j}{\partial x_i}\right)^2}$$

By the sensitivity of a test data x, we can predict the performance.

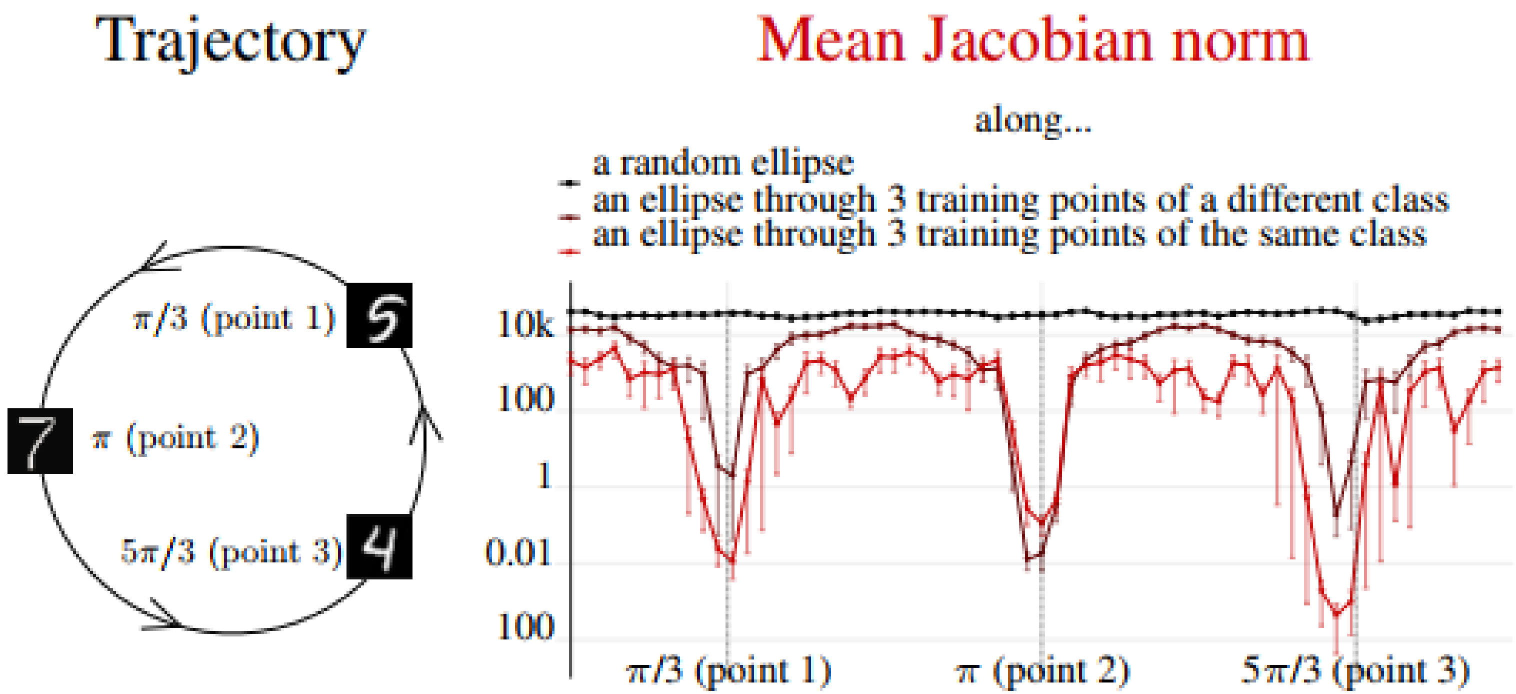
Without label

It is not surprise that sensitivity is related to generalization.

Regularization is kind of minimziing sensitivity.

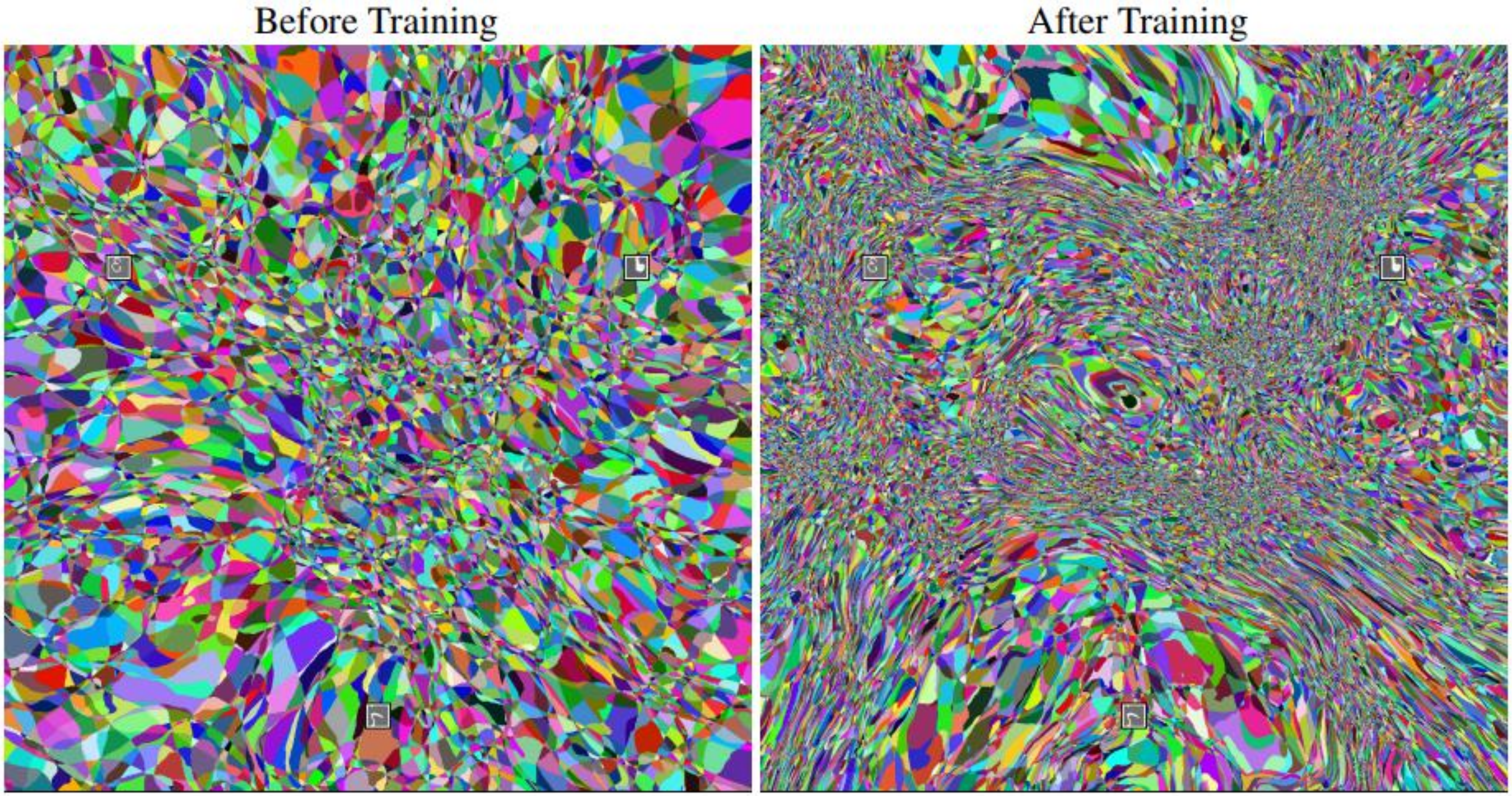
Sensitivity – Emprical Results

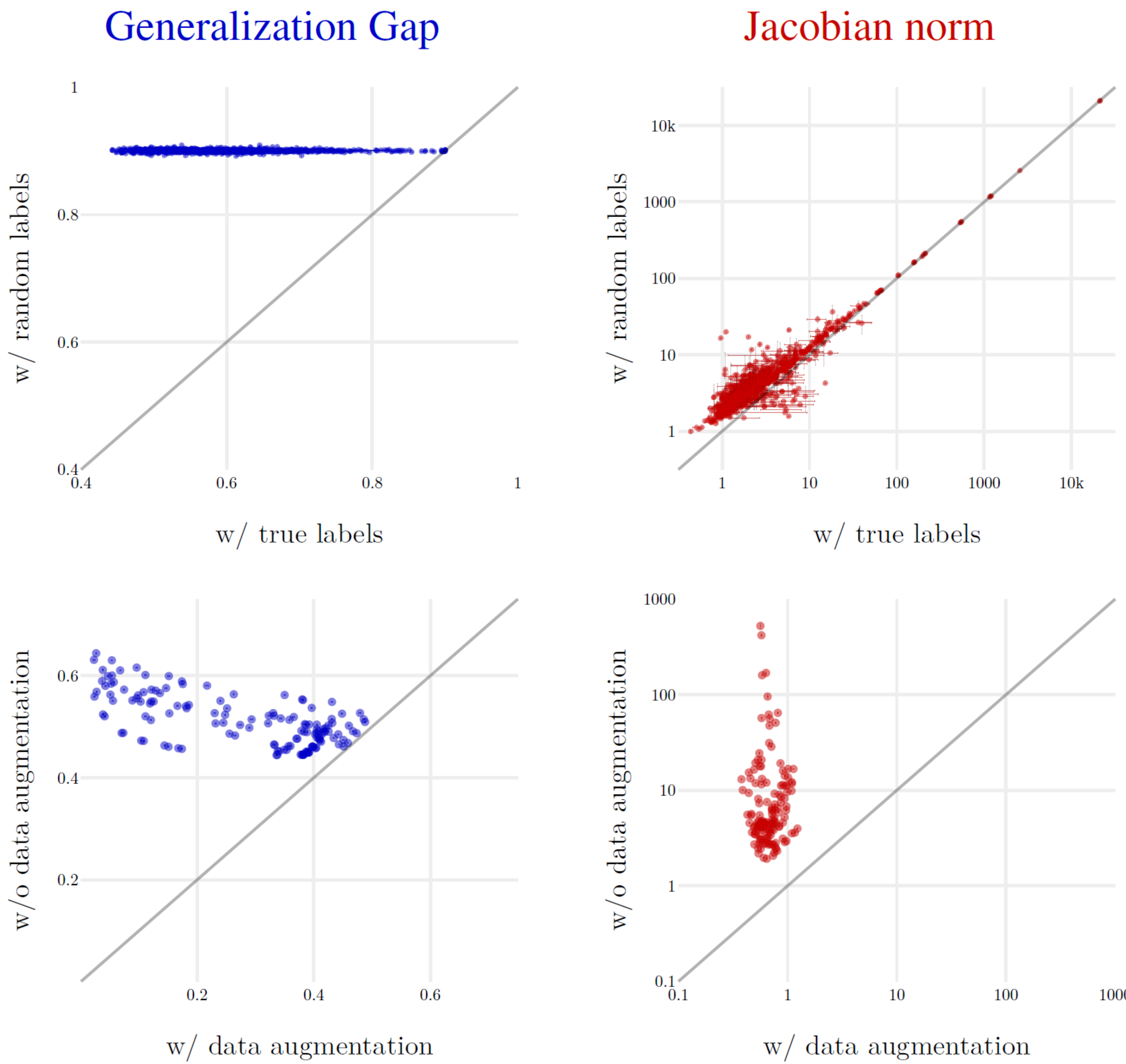
- Sensitivity on and off the training data manifold



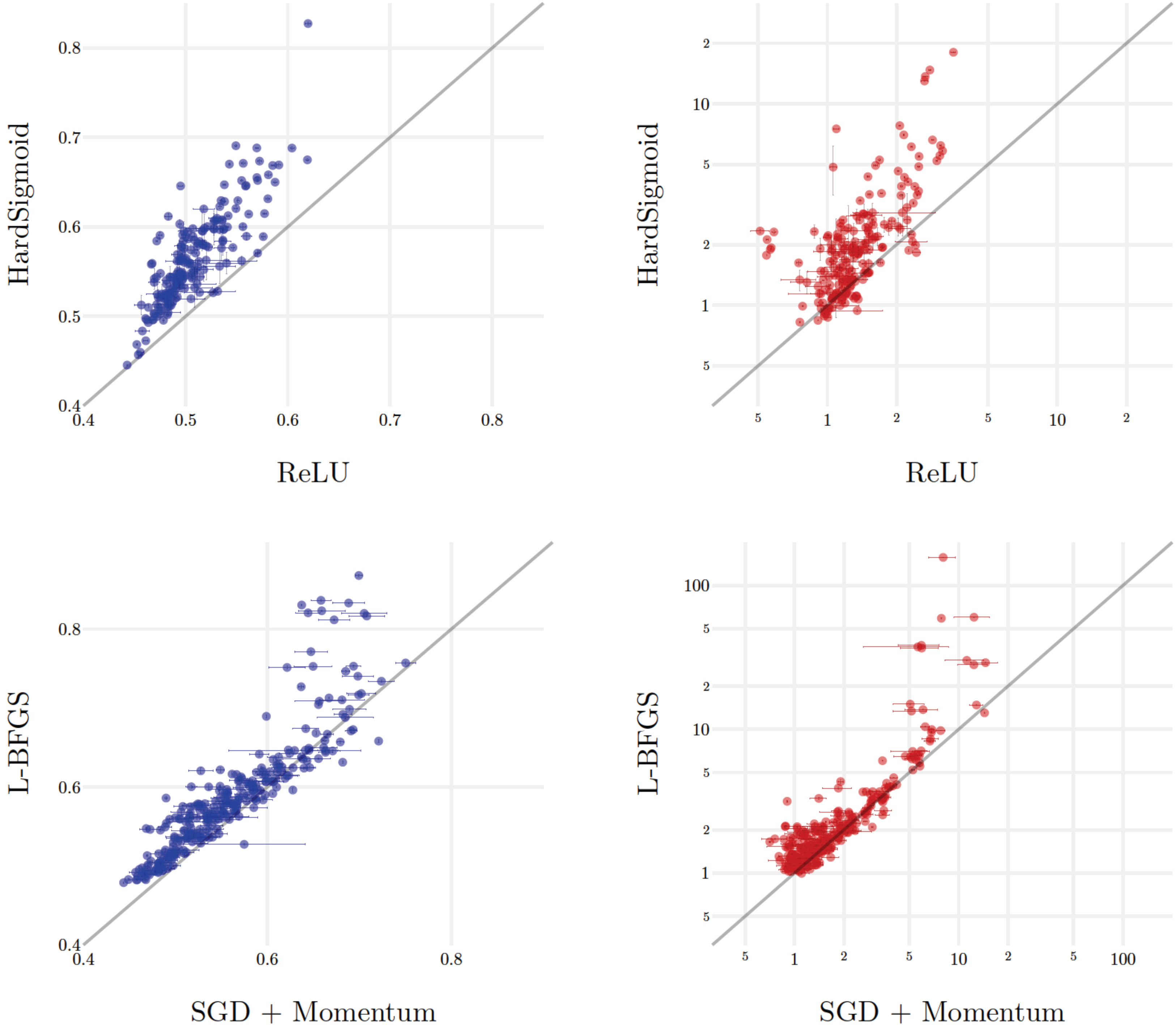
Sensitivity – Emprical Results

- Sensitivity on and off the training data manifold

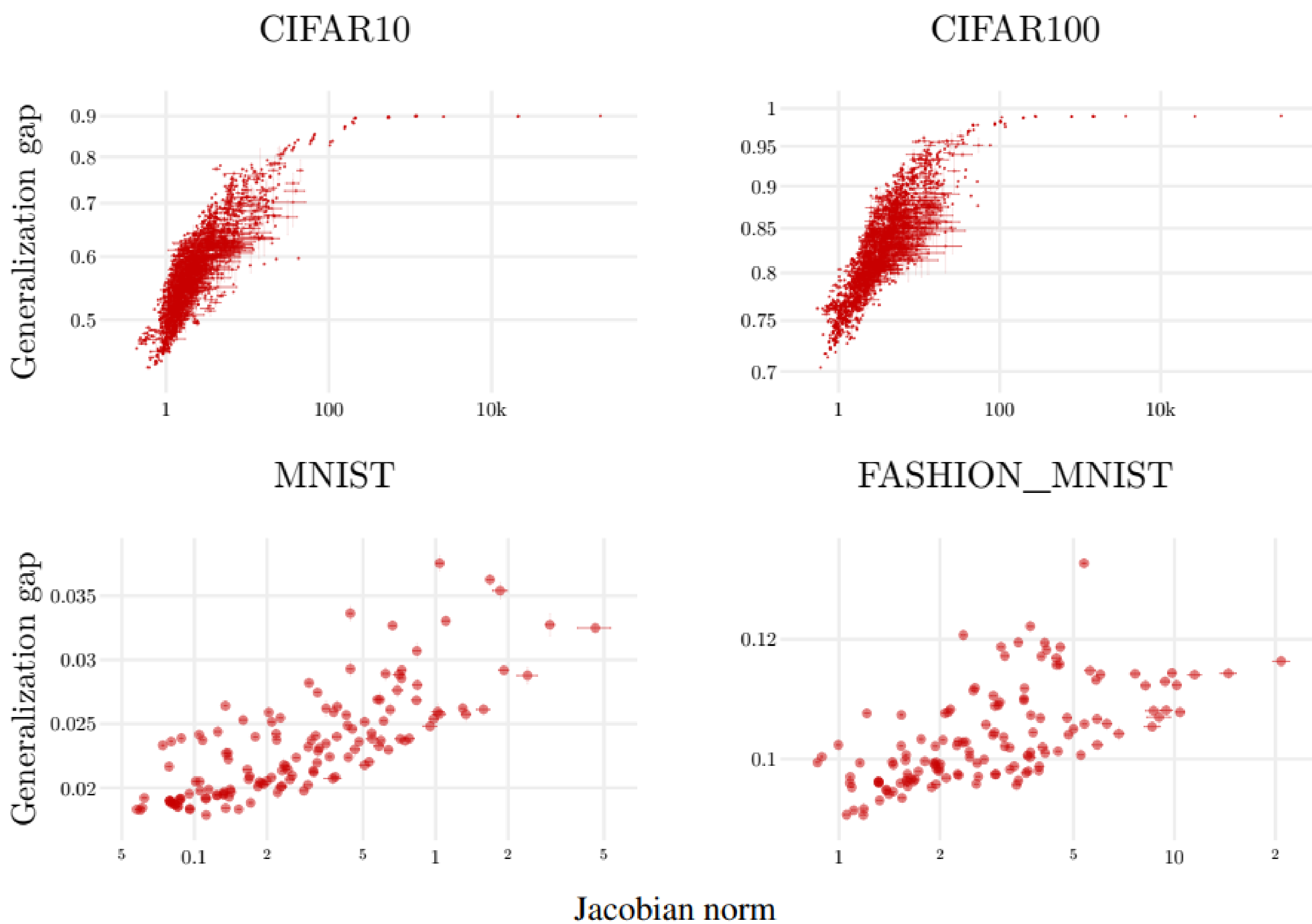




Generalization Gap
Jacobian norm
w/ random labels
w/ true labels
w/o data augmentation
w/ data augmentation
0.4
0.6
0.8
1
0.2
0.1
1
10
100
1000
10k

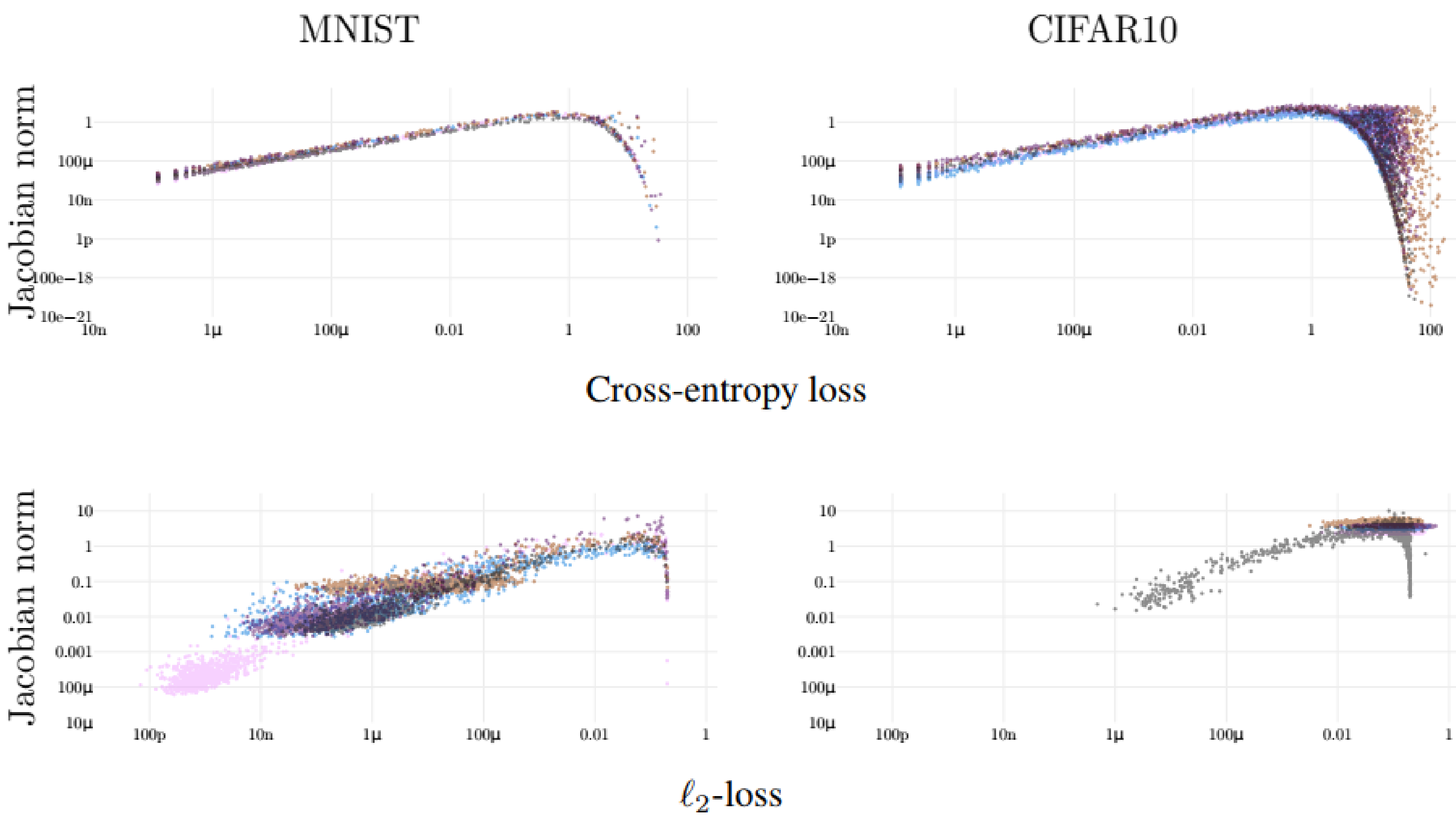

HardSigmoid
ReLU
0.4
0.5
0.6
0.7
0.8
HardSigmoid
ReLU
5
1
2
5
10
2
L-BFGS
SGD + Momentum
0.4
0.6
0.8
L-BFGS
SGD + Momentum
5
1
2
5
10
2
5
100

Sensitivity v.s. Generalization



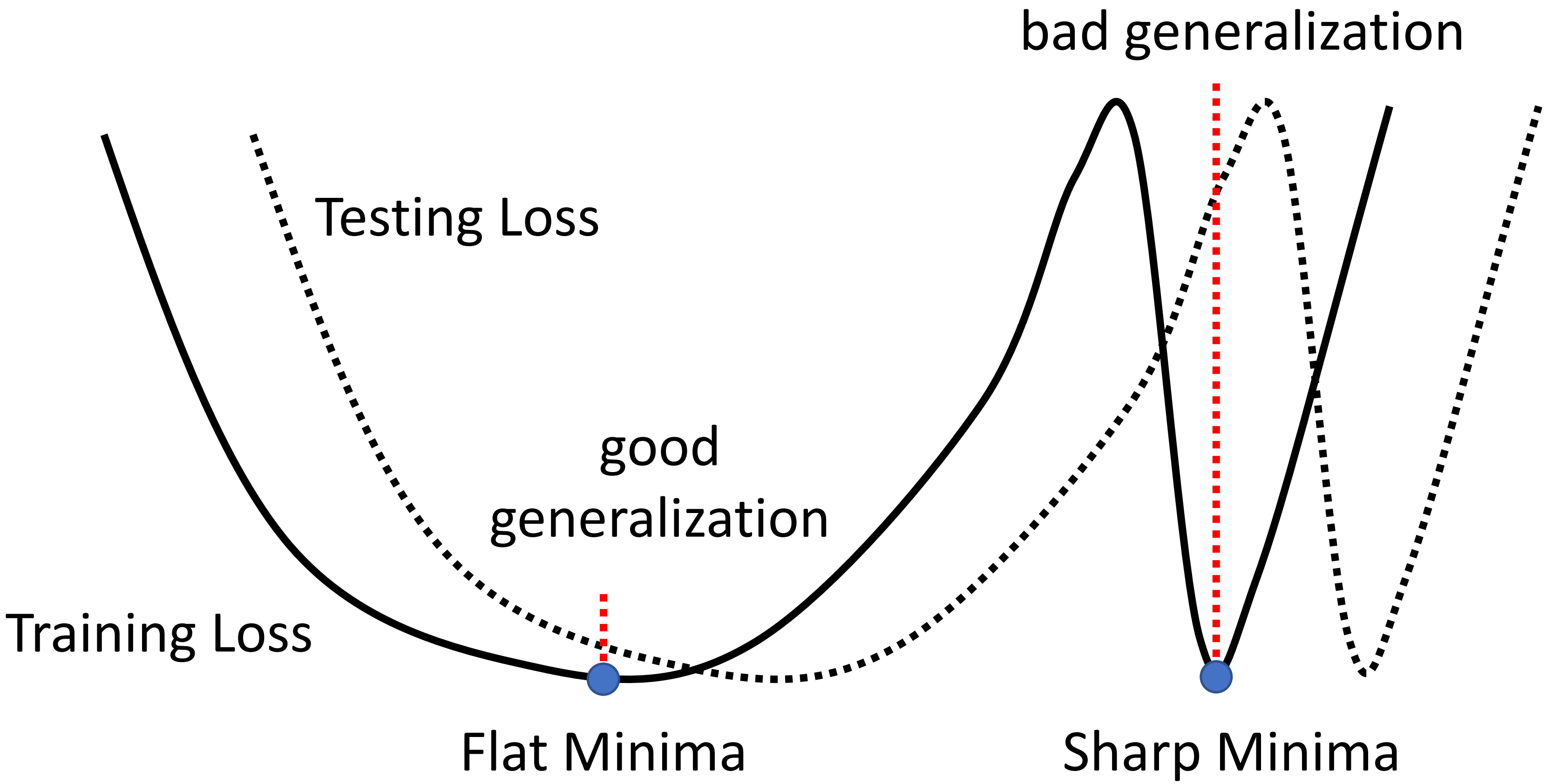
Sensitivity v.s. Generalization

- individual points



Sharpness

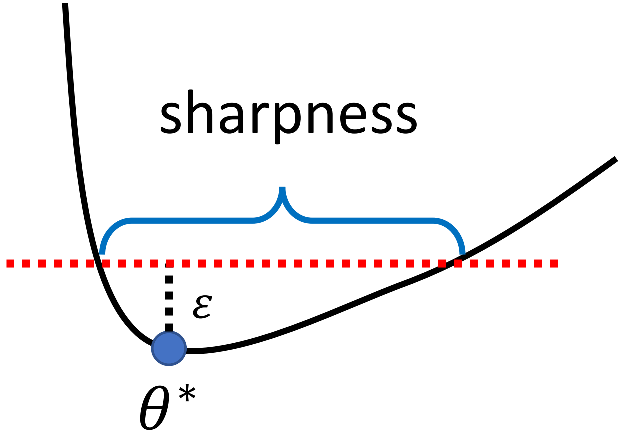
Sharp Minima v.s Flat Minima



Explain self regularization?

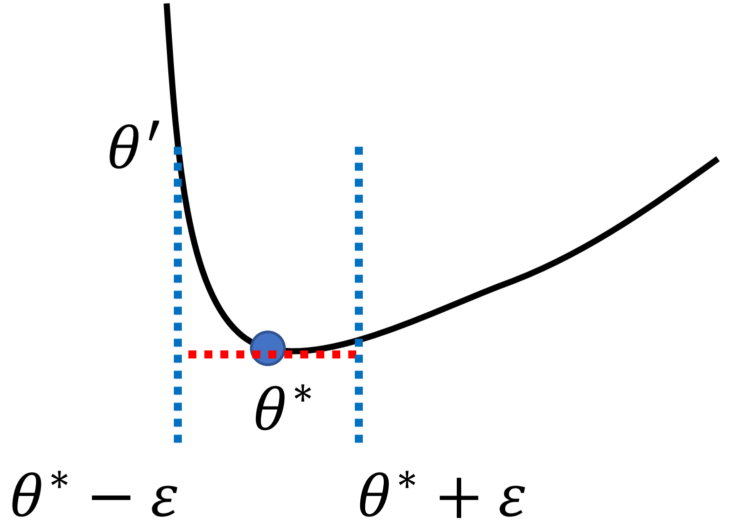
Definition of Sharpness

<u>Definition 1</u>



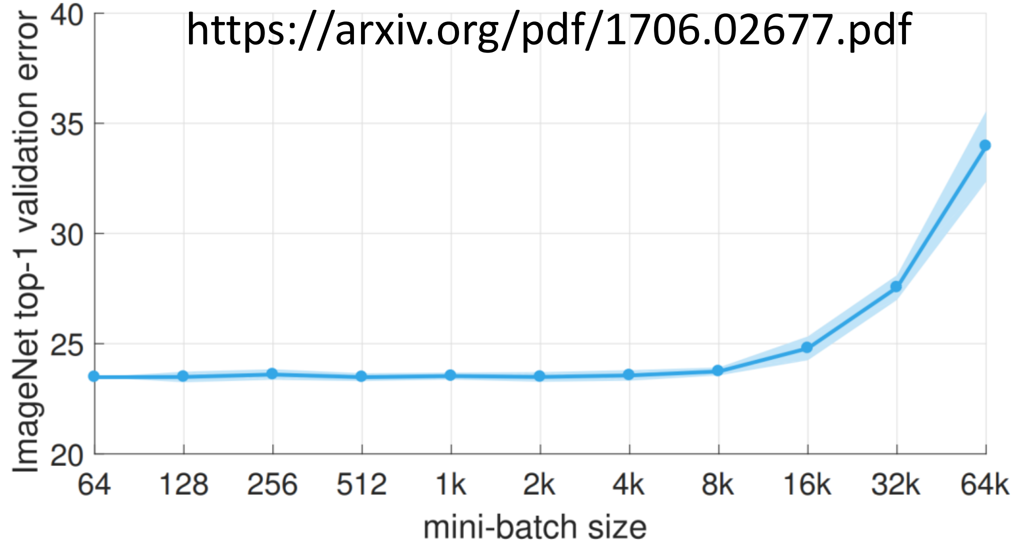
<u>Definition 2</u>

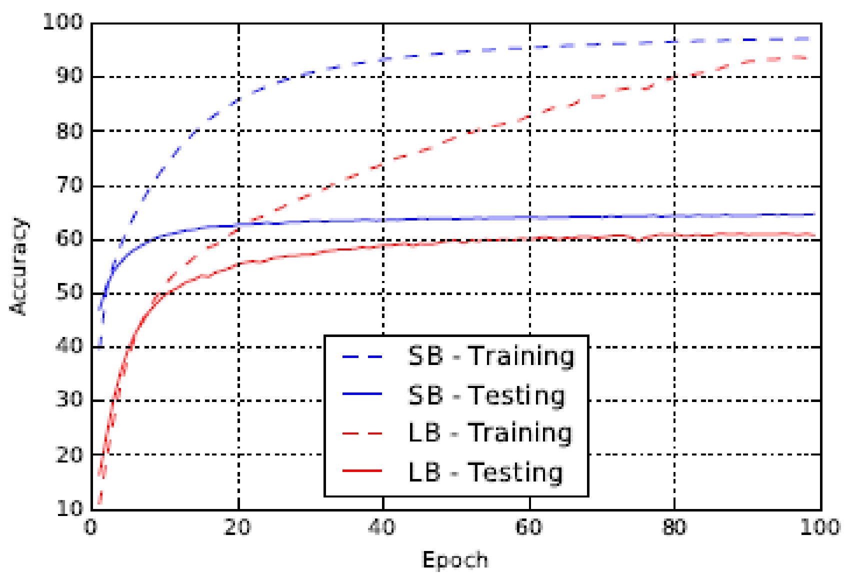
$$\text{Sharpness} = L(\theta') - L(\theta^*)$$



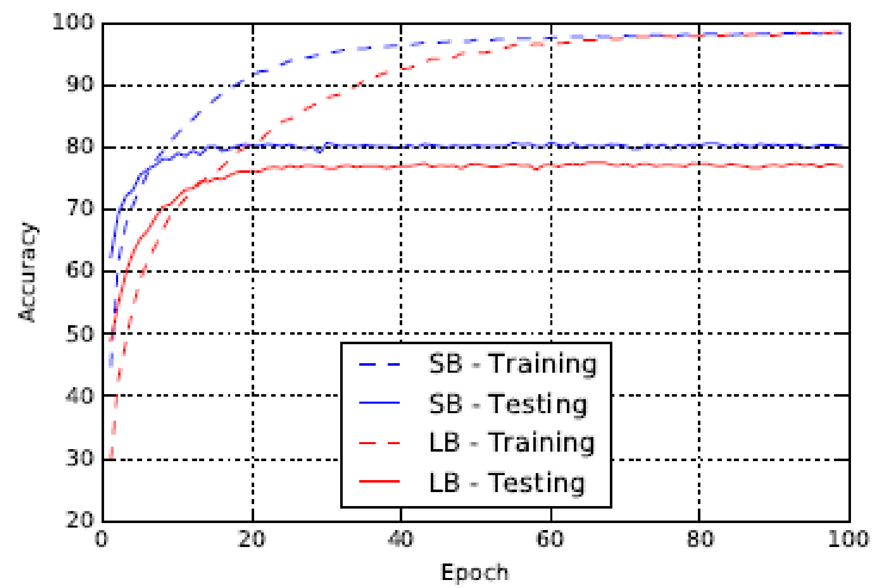
Batch Size

https://arxiv.org/pdf/1706.02677.pdf





(a) Network F_2



(b) Network C_1

Batch Size v.s. Sharpness

Name	Network Type	Data set
F_1	Fully Connected	MNIST (LeCun et al., 1998a)
F_2	Fully Connected	TIMIT (Garofolo et al., 1993)
C_1	(Shallow) Convolutional	CIFAR-10 (Krizhevsky & Hinton, 2009)
C_2	(Deep) Convolutional	CIFAR-10
C_3	(Shallow) Convolutional	CIFAR-100 (Krizhevsky & Hinton, 2009)
C_4	(Deep) Convolutional	CIFAR-100

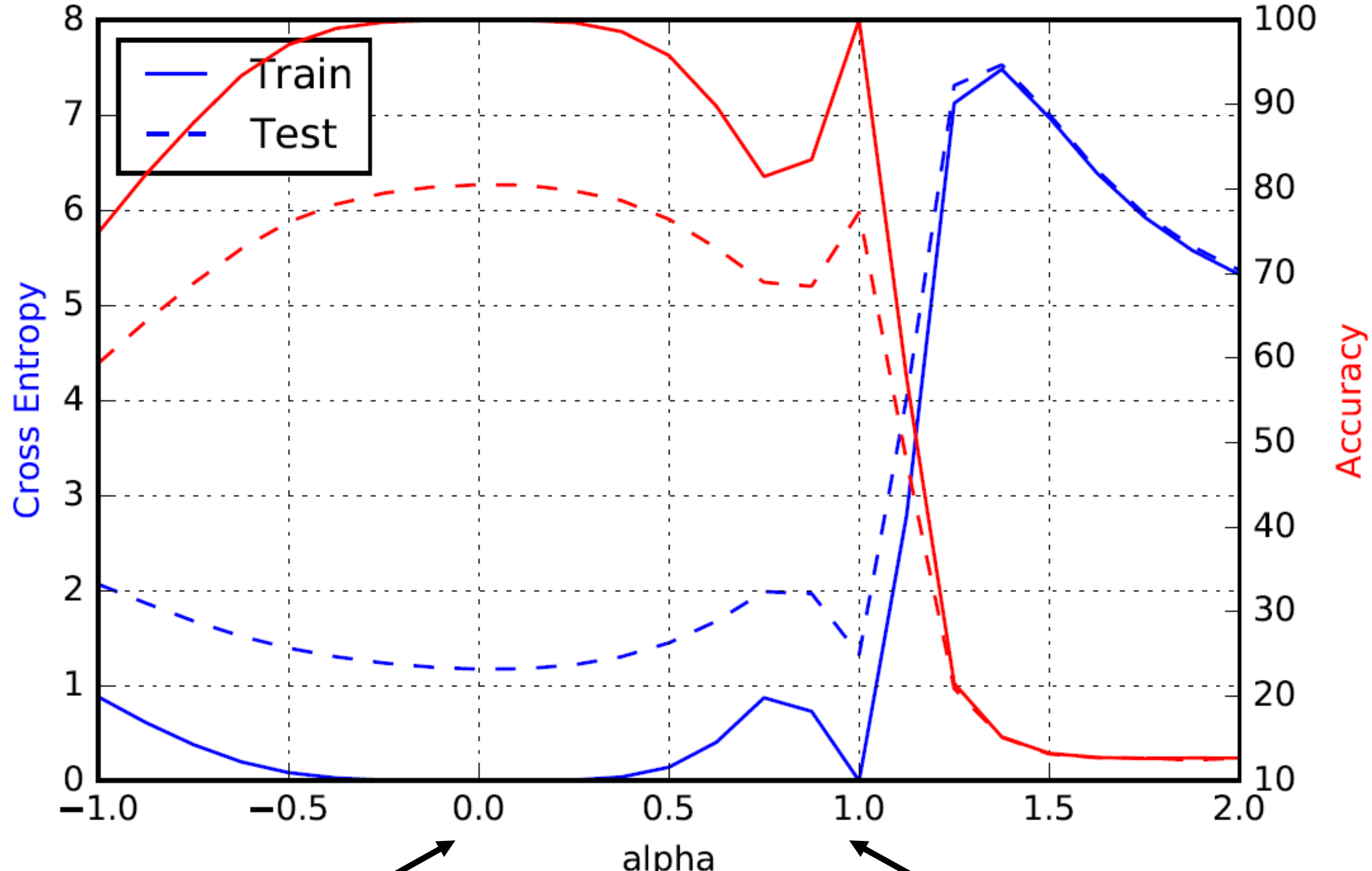
	Training Accuracy		Testing Accuracy	
Name	SB	LB	SB	LB
F_1	$99.66\% \pm 0.05\%$	$99.92\% \pm 0.01\%$	$98.03\% \pm 0.07\%$	$97.81\% \pm 0.07\%$
F_2	$99.99\% \pm 0.03\%$	$98.35\% \pm 2.08\%$	$64.02\% \pm 0.2\%$	$59.45\% \pm 1.05\%$
C_1	$99.89\% \pm 0.02\%$	$99.66\% \pm 0.2\%$	$80.04\% \pm 0.12\%$	$77.26\% \pm 0.42\%$
C_2	$99.99\% \pm 0.04\%$	$99.99\% \pm 0.01\%$	$89.24\% \pm 0.12\%$	$87.26\% \pm 0.07\%$
C_3	$99.56\% \pm 0.44\%$	$99.88\% \pm 0.30\%$	$49.58\% \pm 0.39\%$	$46.45\% \pm 0.43\%$
C_4	$99.10\% \pm 1.23\%$	$99.57\% \pm 1.84\%$	$63.08\% \pm 0.5\%$	$57.81\% \pm 0.17\%$

SB = 256

LB = 0.1 x data set

	$\epsilon = 10^{-3}$		$\epsilon = 5 \cdot 10^{-4}$	
	SB	LB	SB	LB
F_1	1.23 ± 0.83	205.14 ± 69.52	0.61 ± 0.27	42.90 ± 17.14
F_2	1.39 ± 0.02	310.64 ± 38.46	0.90 ± 0.05	93.15 ± 6.81
C_1	28.58 ± 3.13	707.23 ± 43.04	7.08 ± 0.88	227.31 ± 23.23
C_2	8.68 ± 1.32	925.32 ± 38.29	2.07 ± 0.86	175.31 ± 18.28
C_3	29.85 ± 5.98	258.75 ± 8.96	8.56 ± 0.99	105.11 ± 13.22
C_4	12.83 ± 3.84	421.84 ± 36.97	4.07 ± 0.87	109.35 ± 16.57

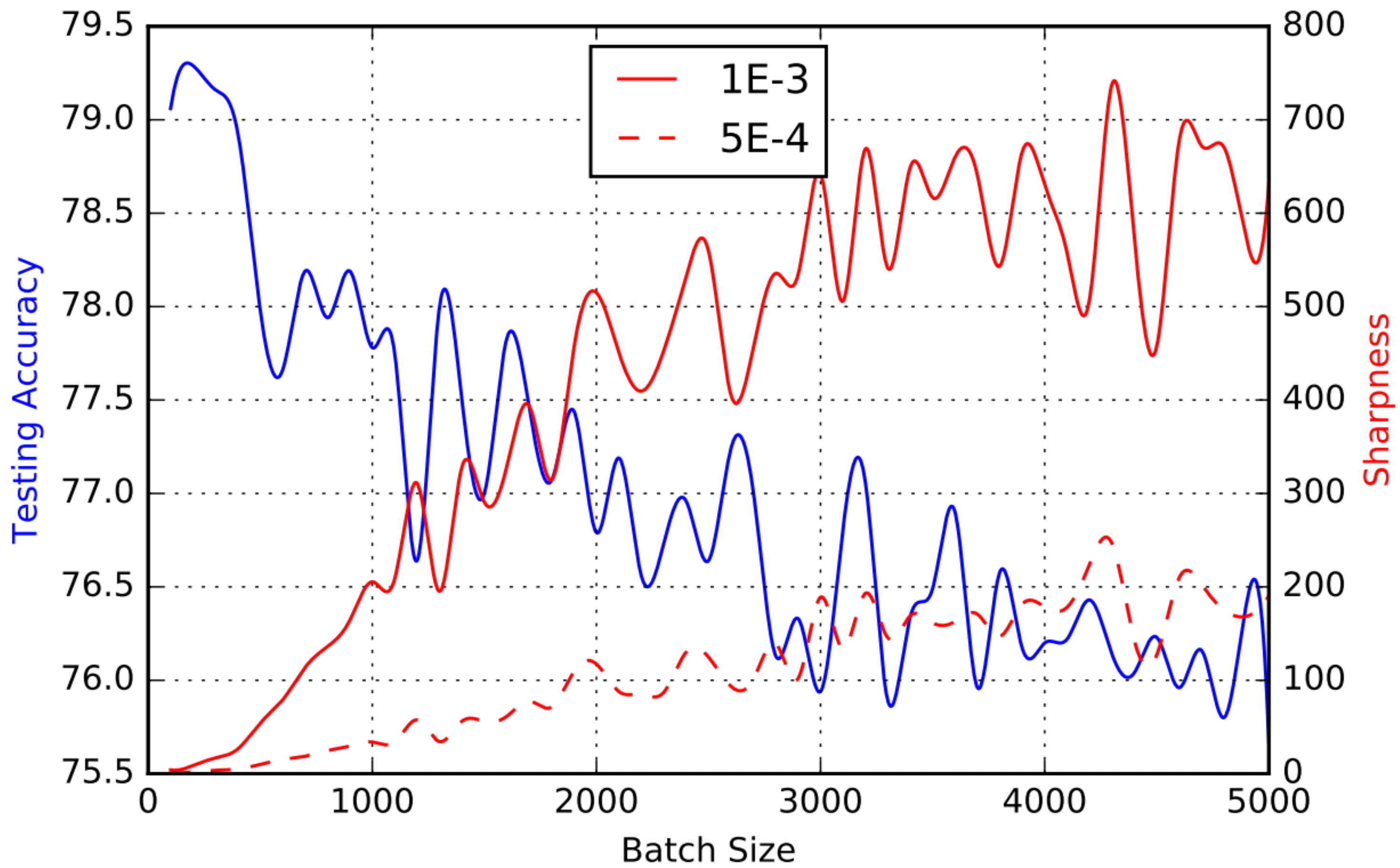
Batch Size v.s. Sharpness



Small Batch

Large Batch

Batch Size v.s. Sharpness



Concluding Remarks

Summary

- Good generalization are associated with sensitivity
- Good generalization are associated with flatness (?)
- Understaning the indicator for generalization helps us develop algorithm in the future

Reference

- Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien, “A Closer Look at Memorization in Deep Networks”, ICML, 2017
- Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang, “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima”, ICLR, 2017
- Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, Riccardo Zecchina, “Entropy-SGD: Biasing Gradient Descent Into Wide Valleys”, ICLR, 2017
- Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, Nathan Srebro, Exploring Generalization in Deep Learning, NIPS, 2017
- Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio, Sharp Minima Can Generalize For Deep Nets, PMLR, 2017
- Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, Jascha Sohl-Dickstein, Sensitivity and Generalization in Neural Networks: an Empirical Study, ICLR, 2018