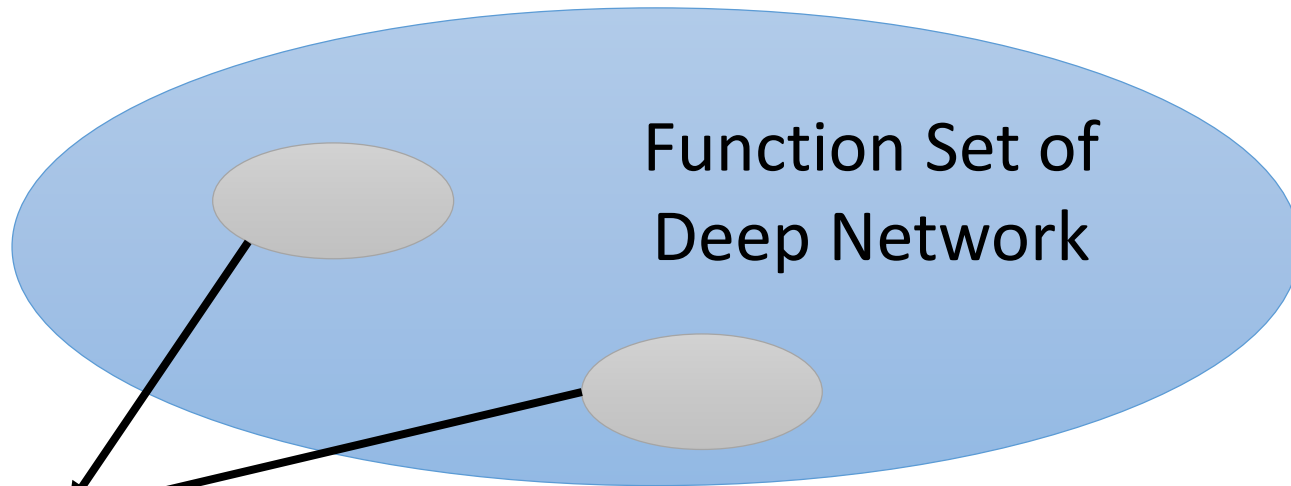
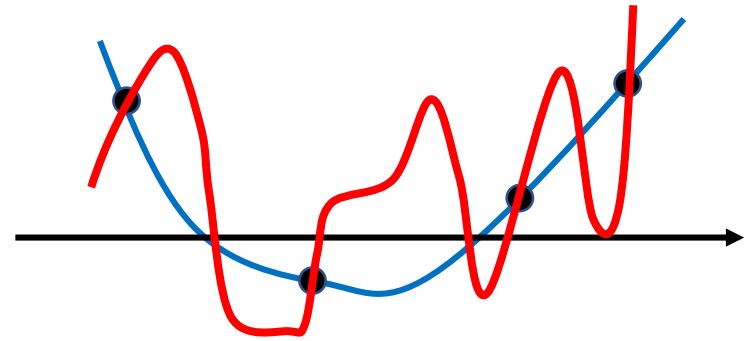


# Indicator of Generalization

# Introduction



Training zero is zero

- 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



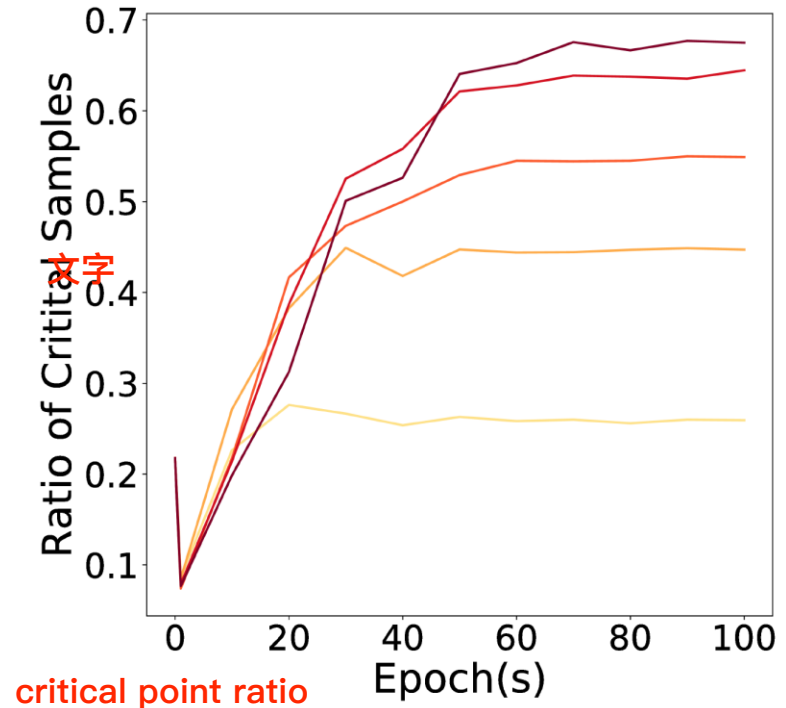
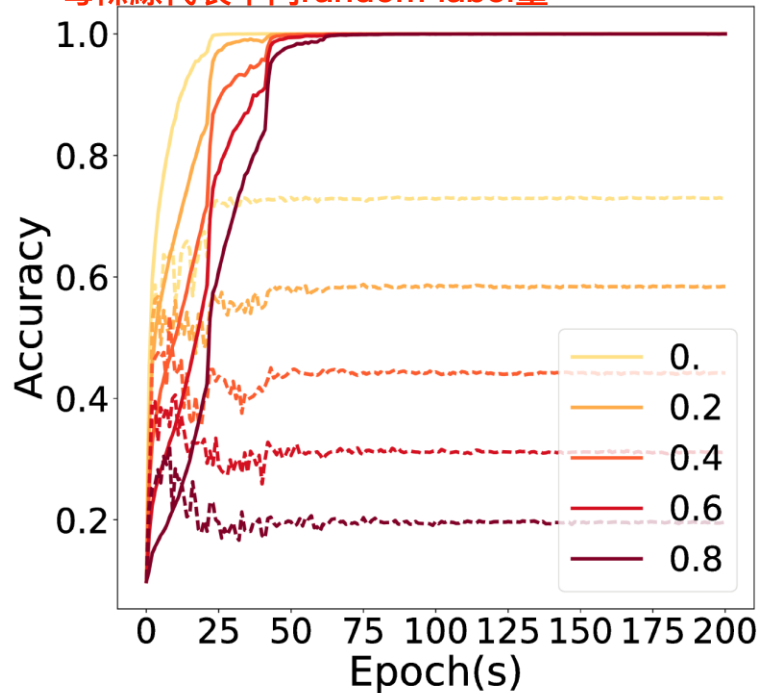
First layer of CIFAR-10

# Brute-force Memorization ?

- Simple pattern first, then memorize exception

雖然最後train出來都可以達到1.0acc，但是其實是不同的network

每條線代表不同random label量

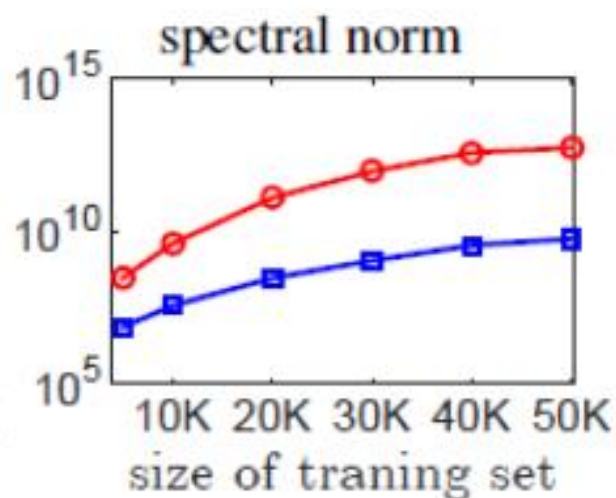
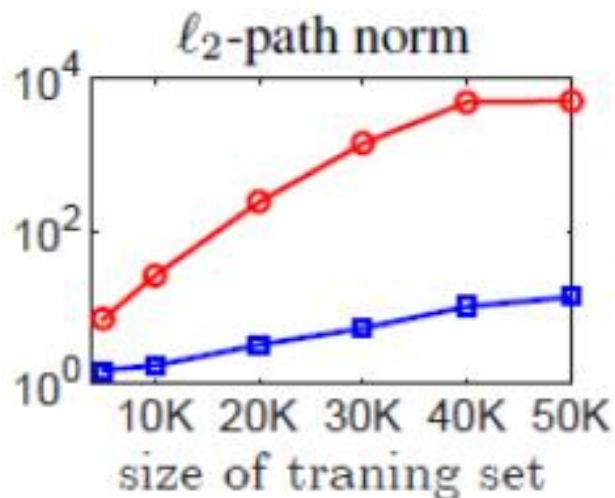
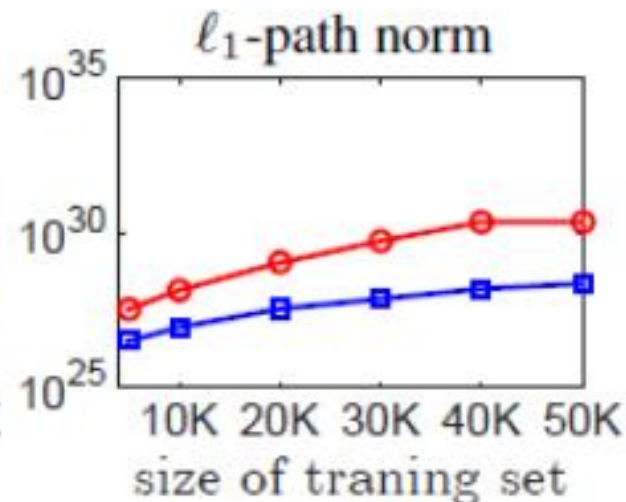
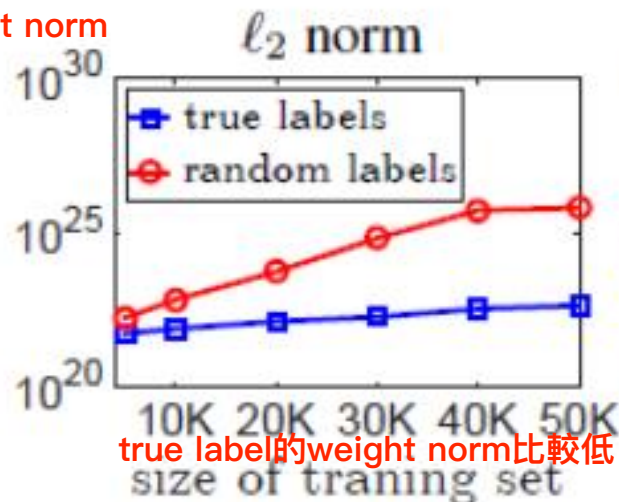


(b) Noise added on classification labels.

他將label data做PCA後呈現，如果發現這個point周圍擁有許多不同的label，表示他可能加入一點noise就會產生不同結果，意味著這是saddle point

# Brute-force Memorization ?

weight norm



# Sensitivity

Google Deepmind

# Jacobian Matrix

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

$$\frac{\partial y}{\partial x} = \underbrace{\hspace{10em}}_{\text{size of } x} \left\{ \begin{array}{l} \text{分別拿 } x_1/x_2/x_3 \text{ 對 } y_1/y_2 \text{ 做偏微分} \\ \text{size of } y \end{array} \right.$$

## Example

$$\begin{bmatrix} x_1 + x_2 x_3 \\ 2x_3 \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) \quad \frac{\partial y}{\partial x} = \begin{bmatrix} 1 & x_3 & x_2 \\ 0 & 0 & 2 \end{bmatrix}$$

# Sensitivity

要有data以及network才可以算sensitivity

透過sensitivity，我們可以計算這筆test data對整個network的變化影響多大，越大表示越容易判斷錯誤

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

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

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

input一個少量的變化對我們的輸出有什麼樣的影響  
( $x$ 有一點變化的時候對結果 $y$ 有什麼樣的變化)

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 minimzing sensitivity.

希望network的值越接近0越好

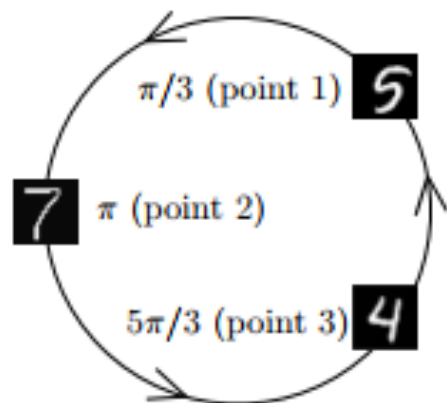


# Sensitivity – Empirical Results

- Sensitivity on and off the training data manifold

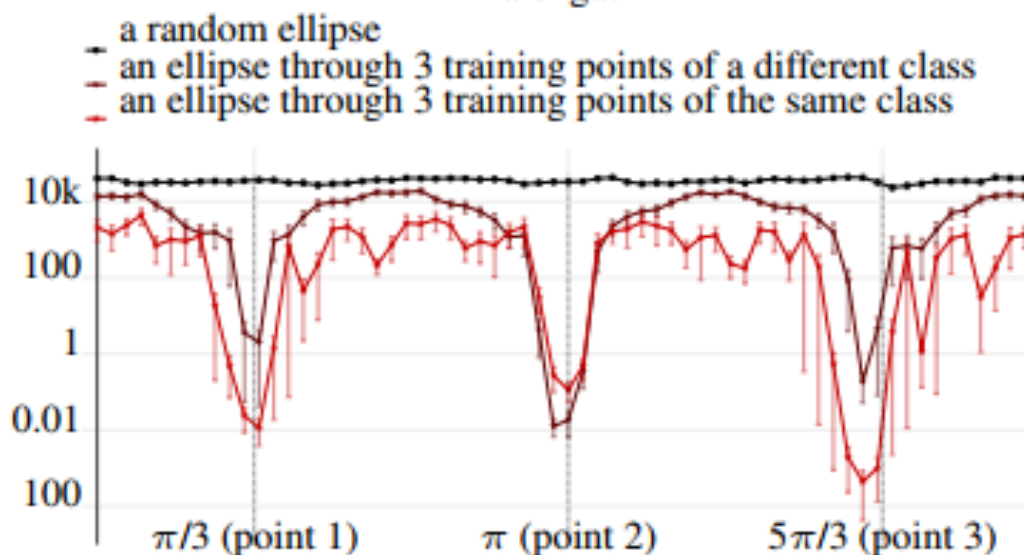
Trajectory

每個點都是彼此的interpolation



Mean Jacobian norm

along...

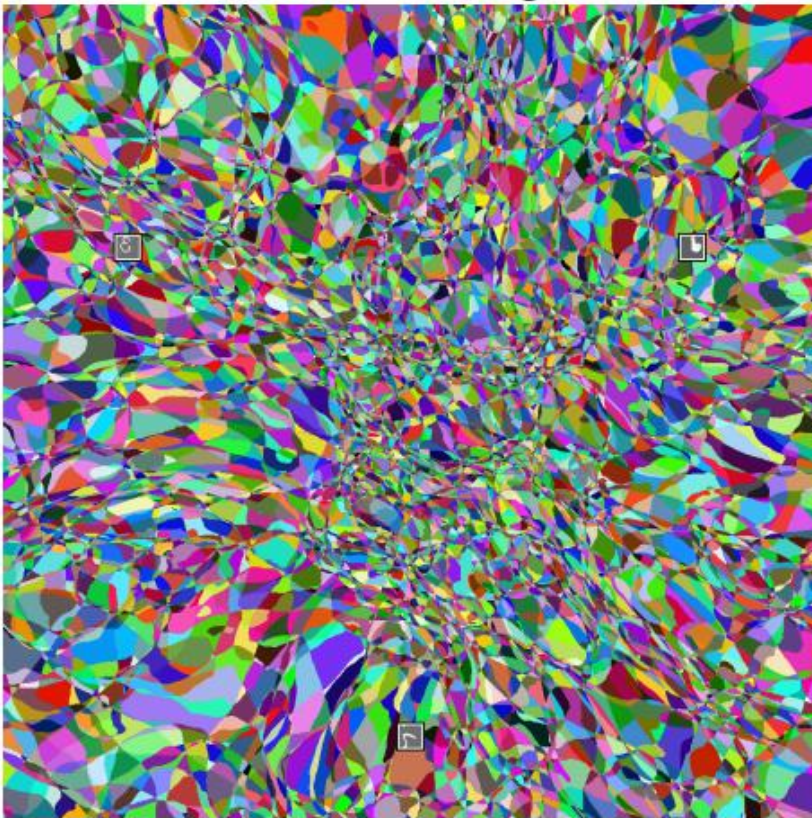


sample每個點所計算的jacobian norm

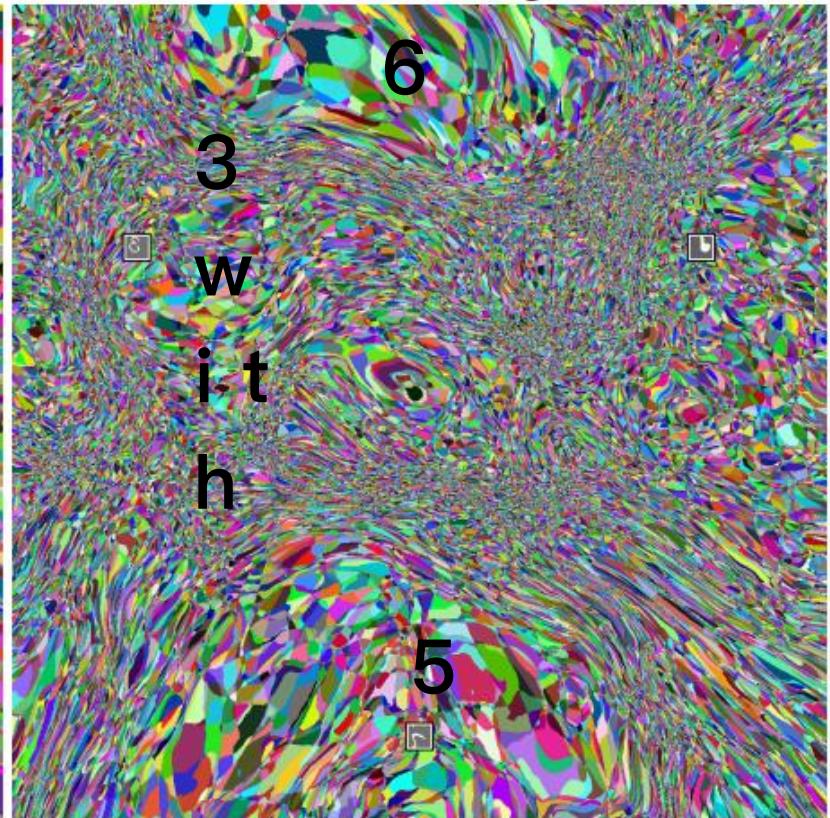
# Sensitivity – Empirical Results

- Sensitivity on and off the training data manifold

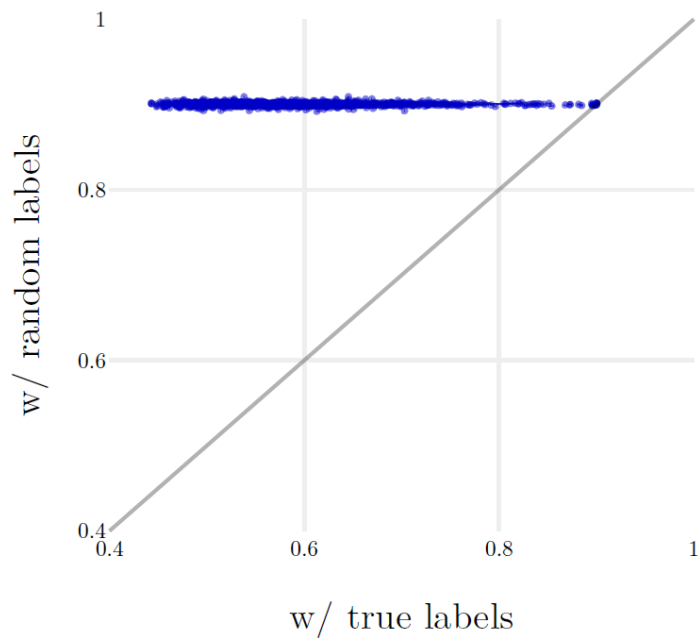
Before Training



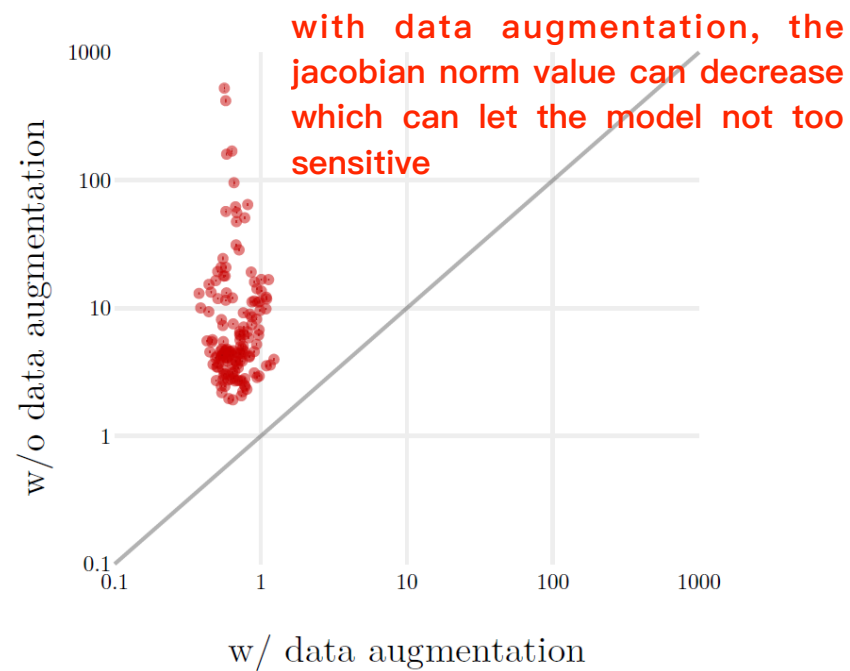
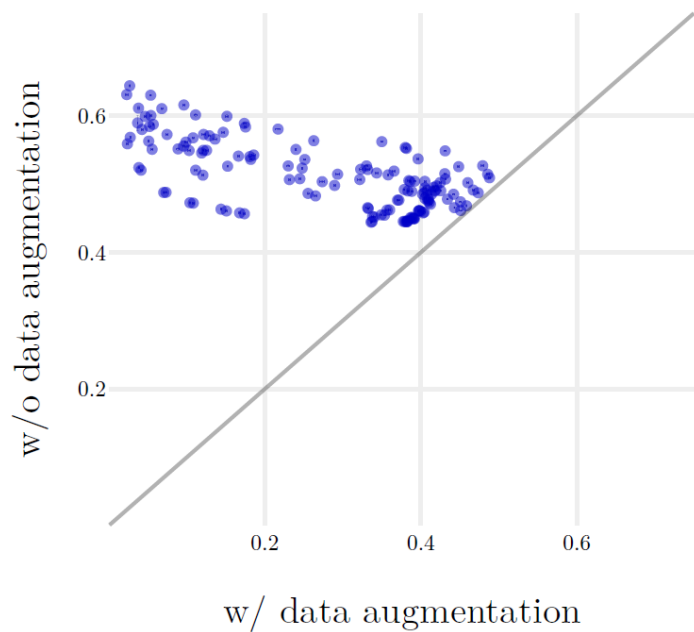
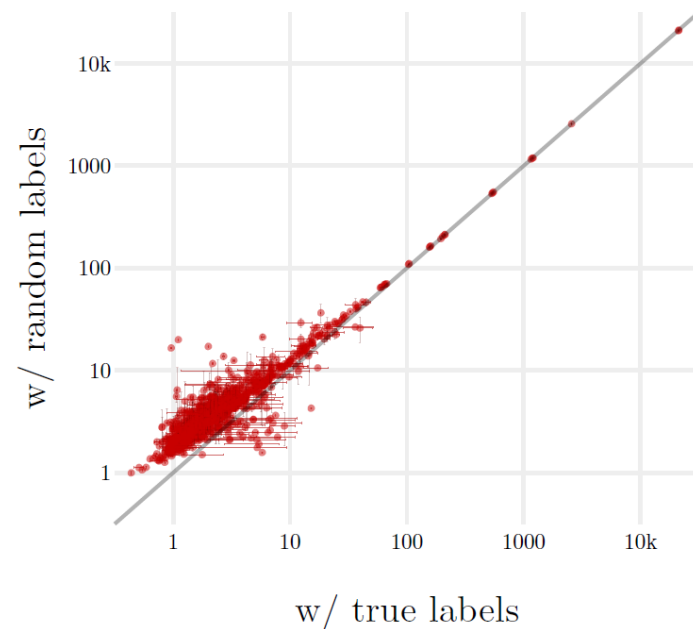
After Training

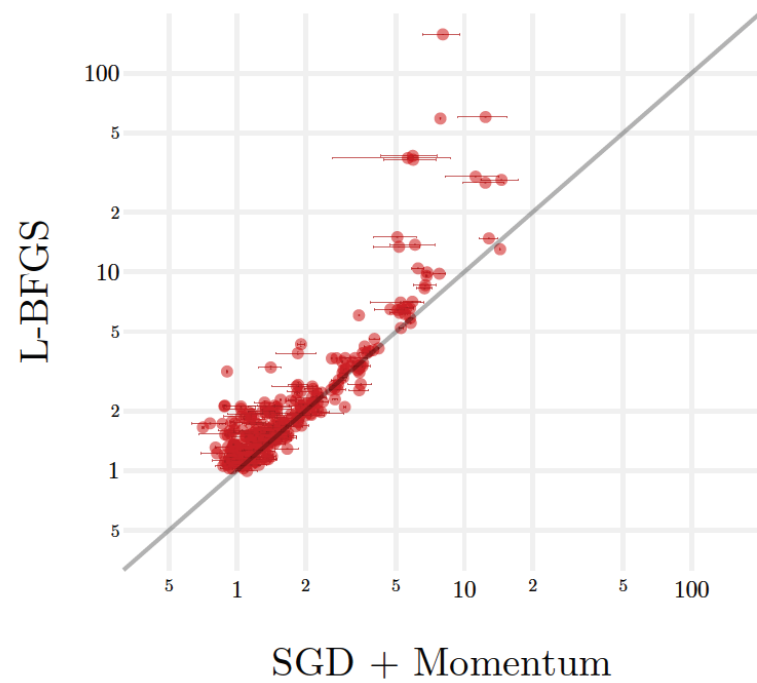
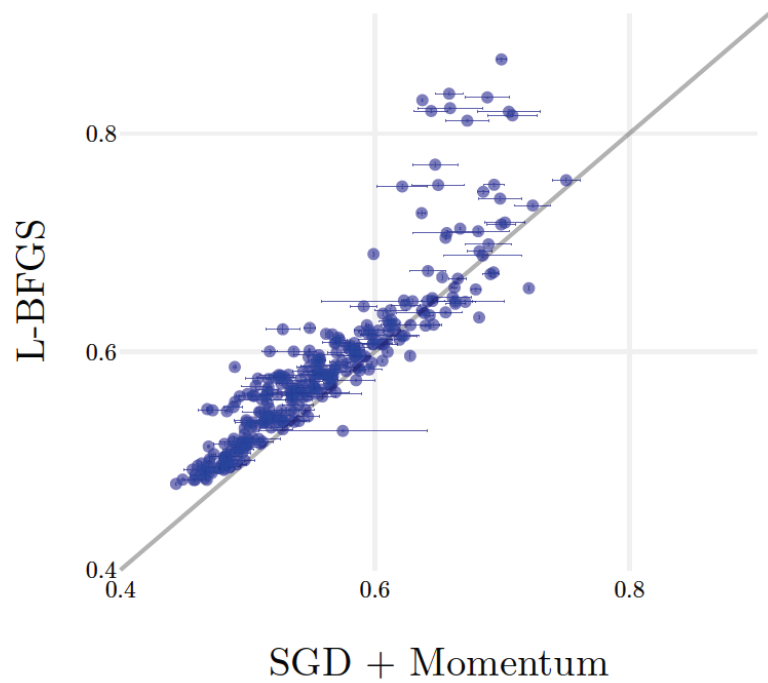
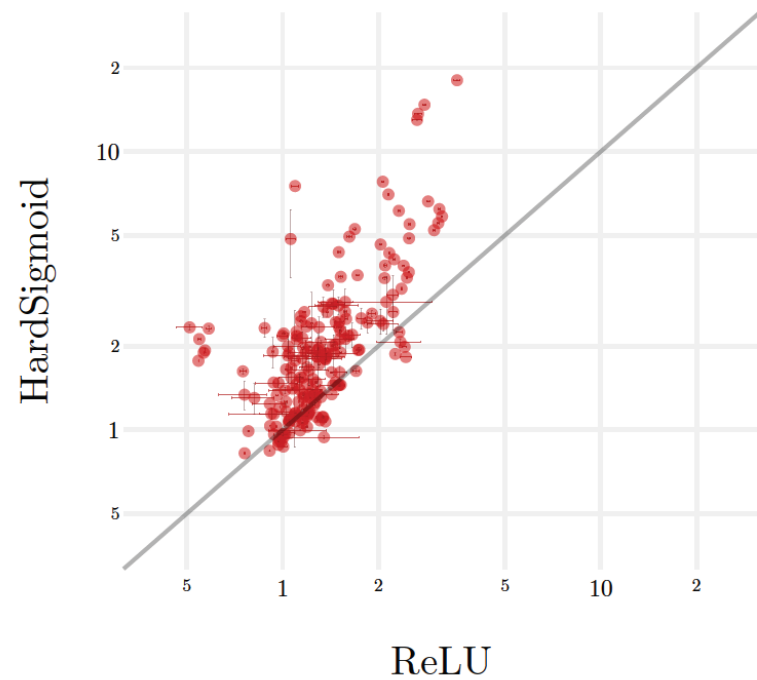
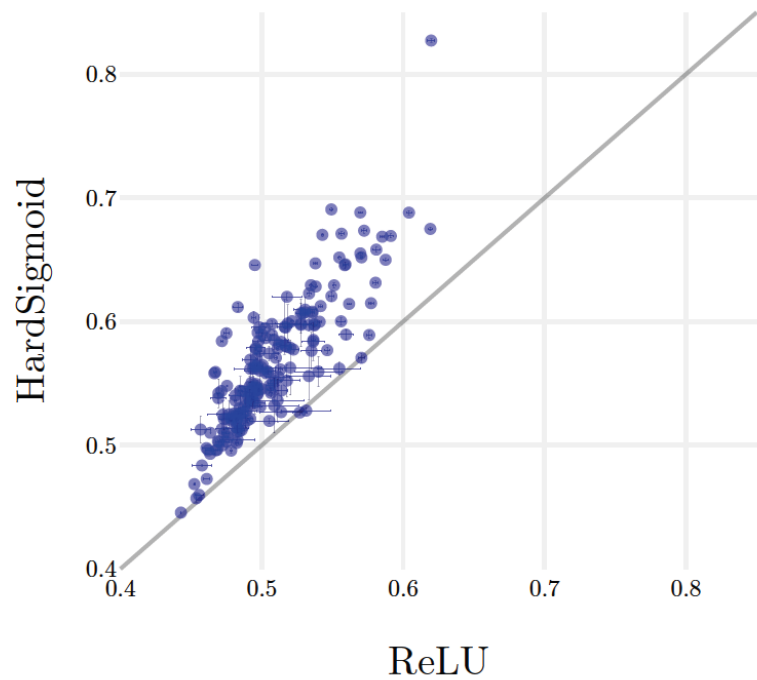


## Generalization Gap



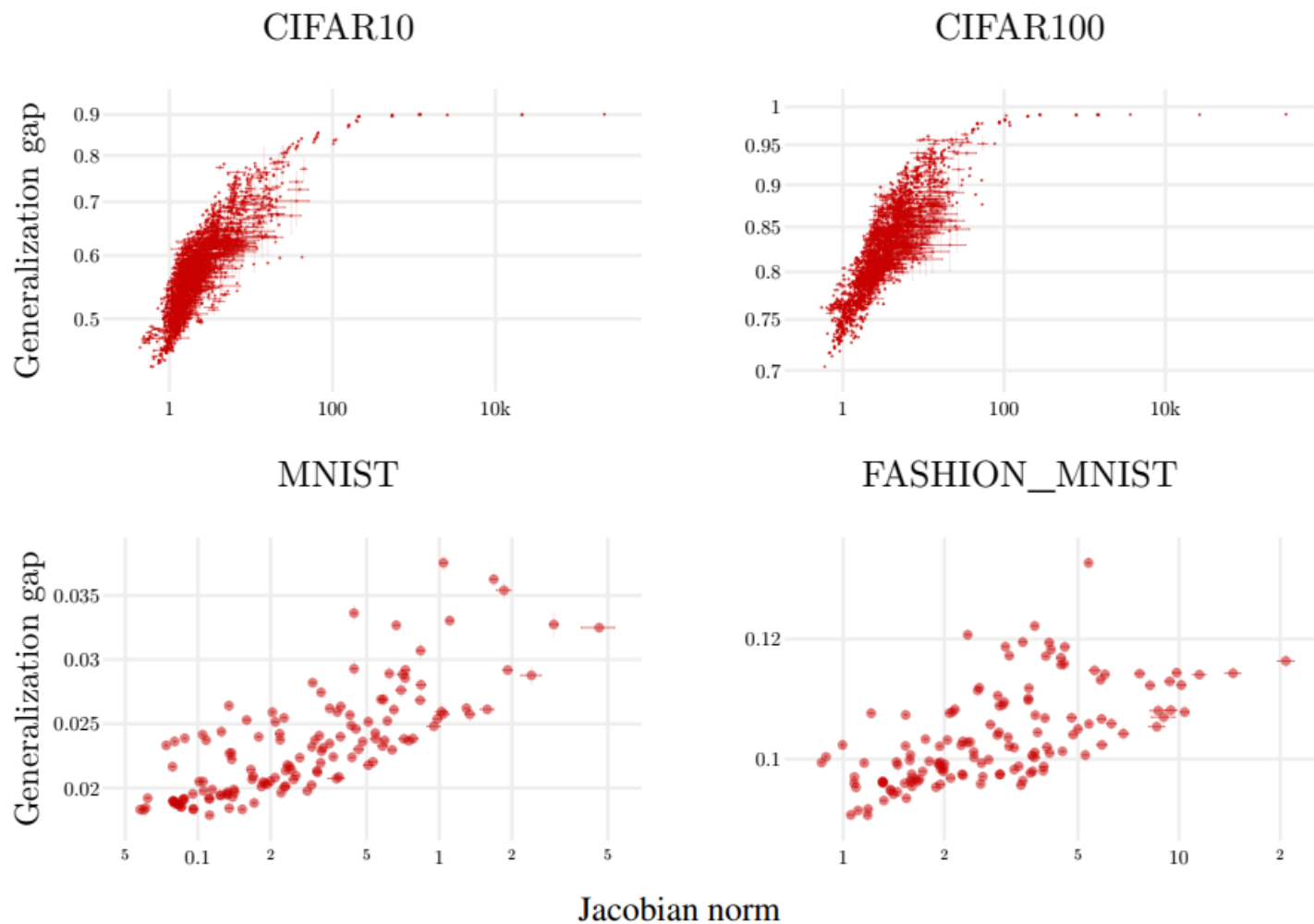
## Jacobian norm







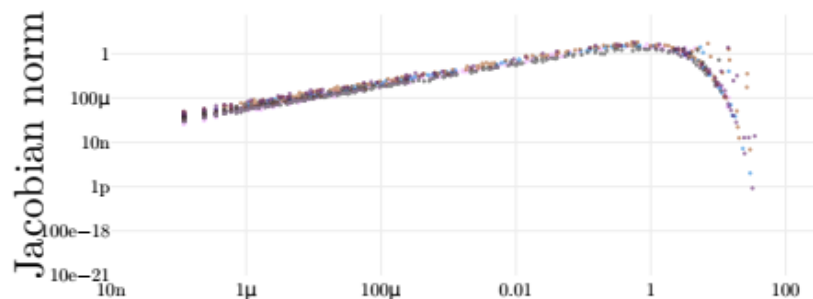
# Sensitivity v.s. Generalization



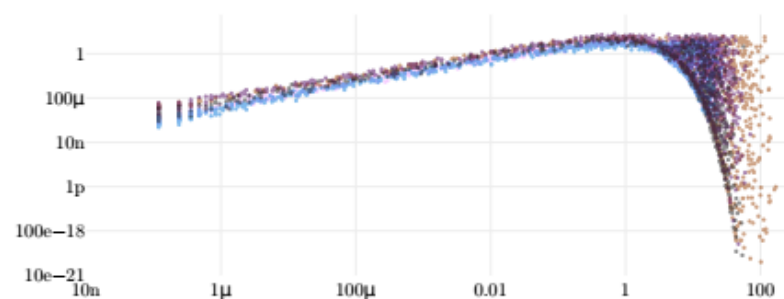
# Sensitivity v.s. Generalization

- individual points

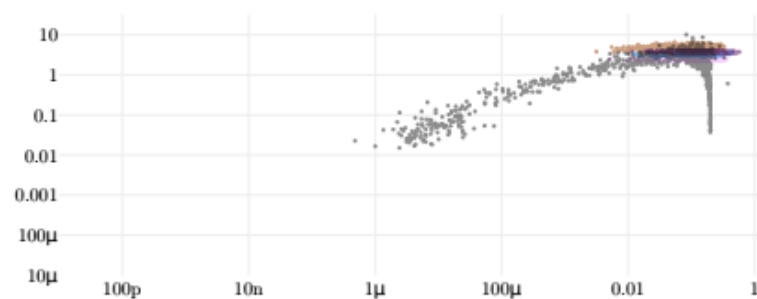
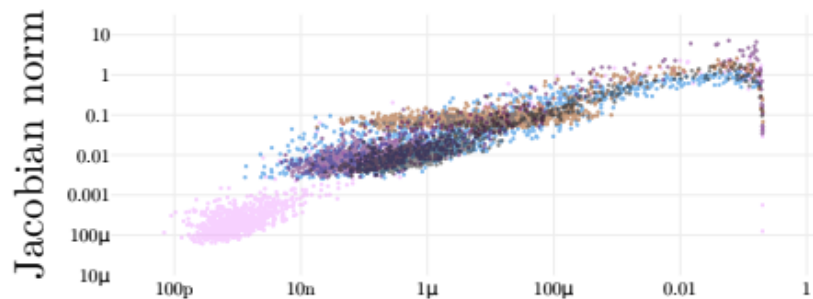
MNIST



CIFAR10



Cross-entropy loss

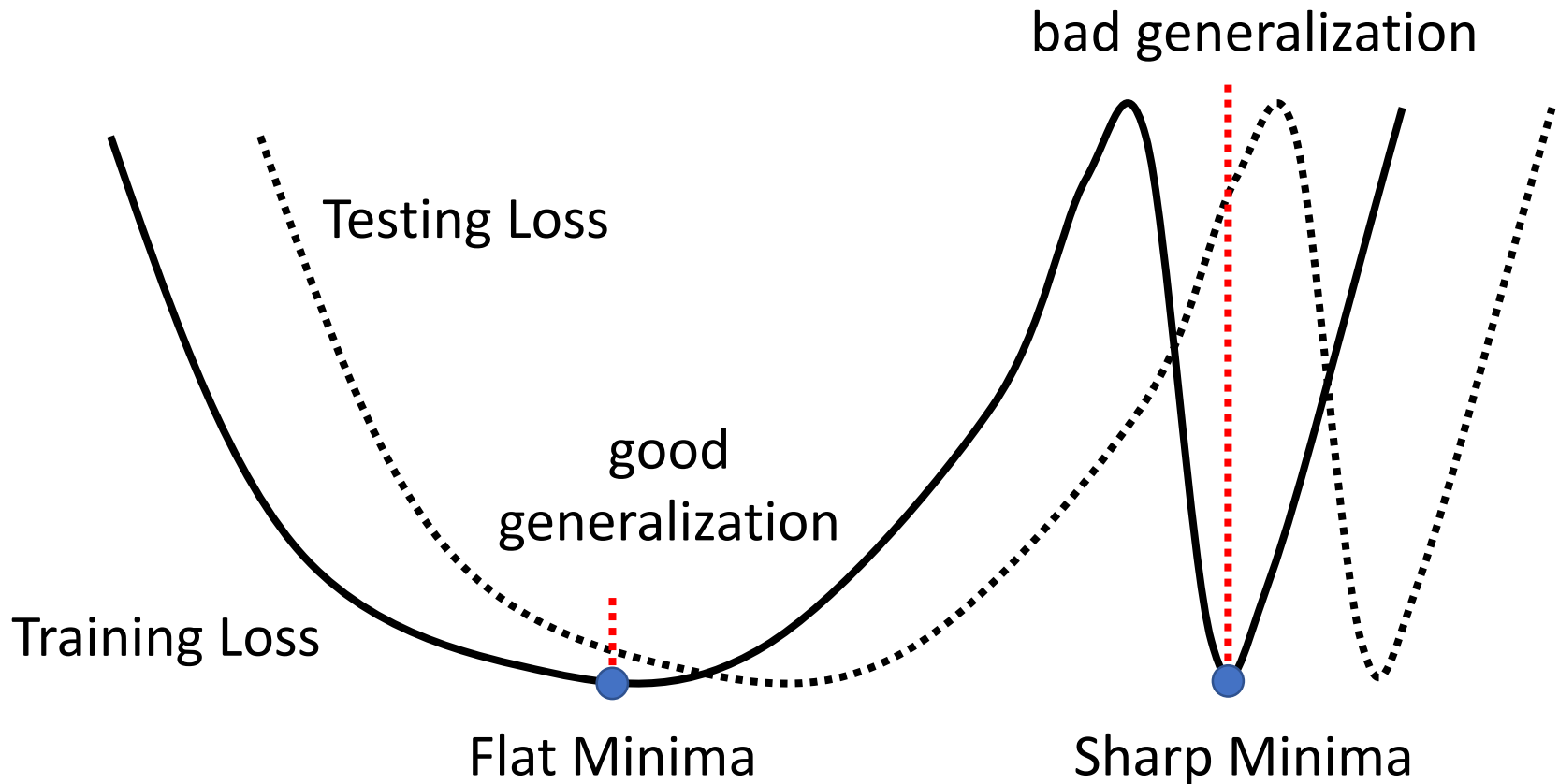


$\ell_2$ -loss

Sharpness

# Sharp Minima v.s Flat Minima

可能testing set的distribution與training set是有誤差的



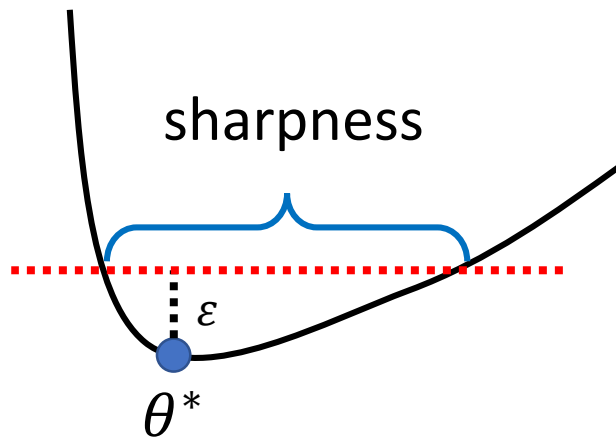
flat minima比較能夠對應testing set的mismatch

Explain self regularization?



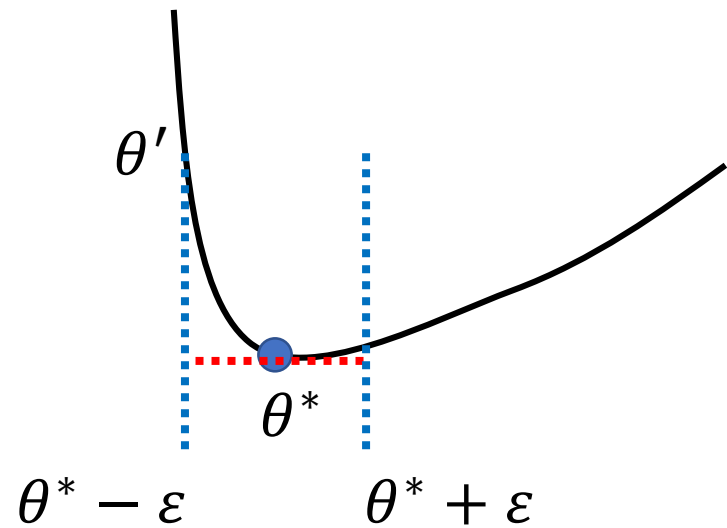
# Definition of Sharpness

## Definition 1



## Definition 2

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



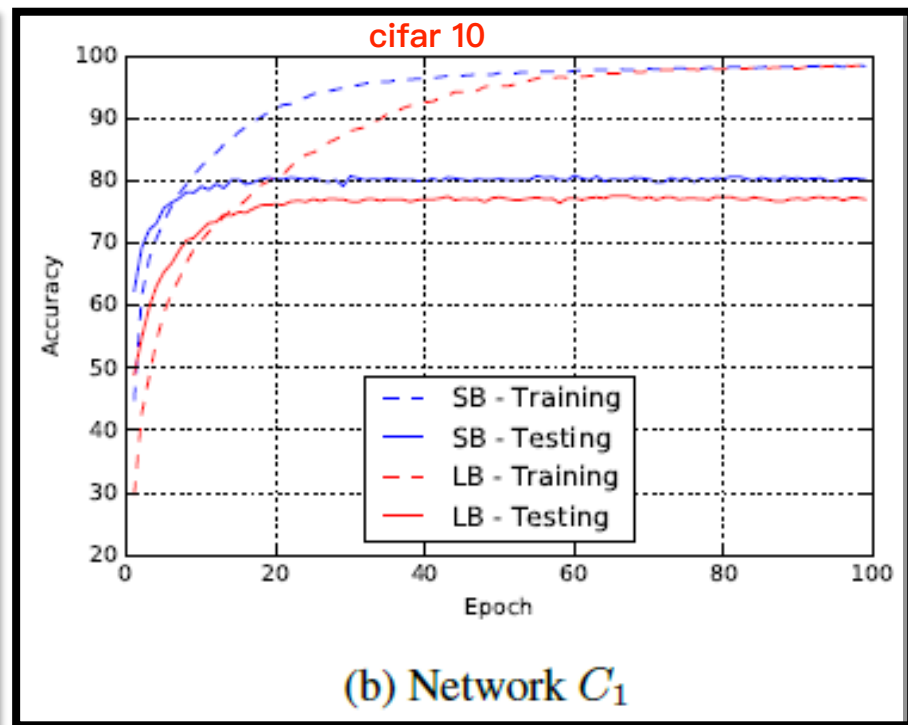
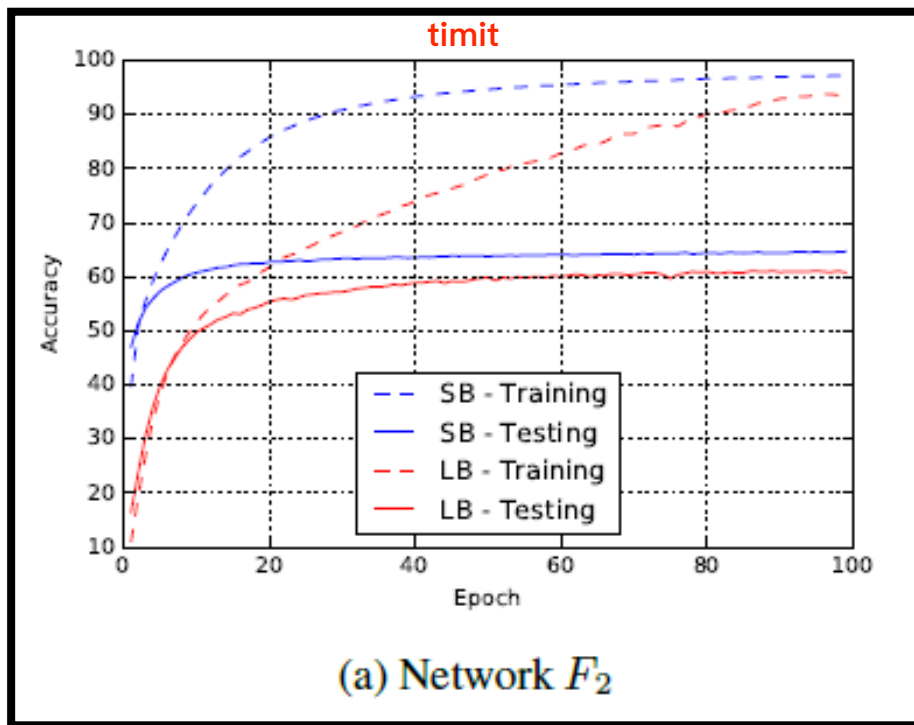
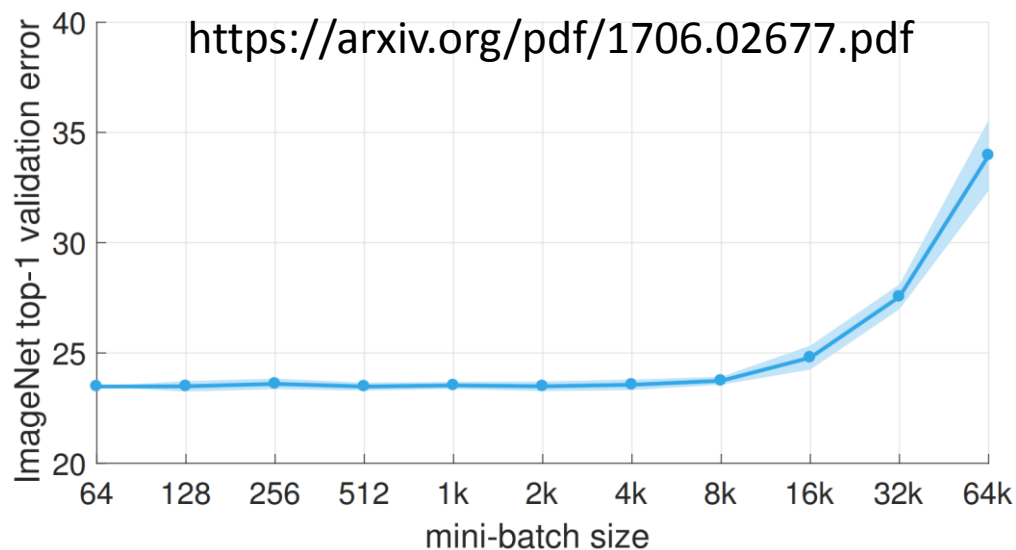
# Batch Size

batch size跟sharpness是有關係的

一般來說batch\_size越大 validation error越大

batch\_size變化對testing acc有影響

如果epoch夠多 training acc都可以接近



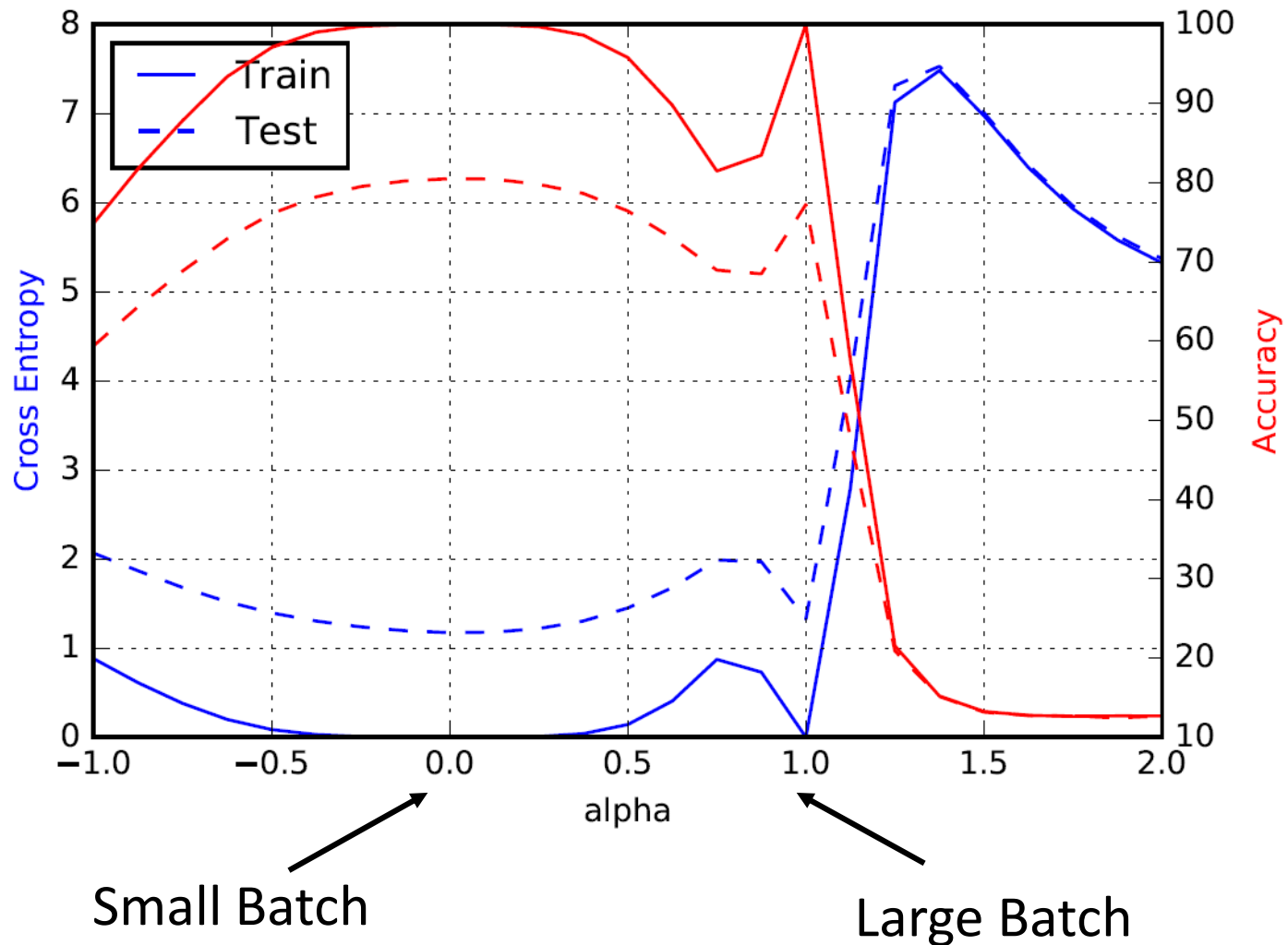
# Batch Size v.s. Sharpness

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

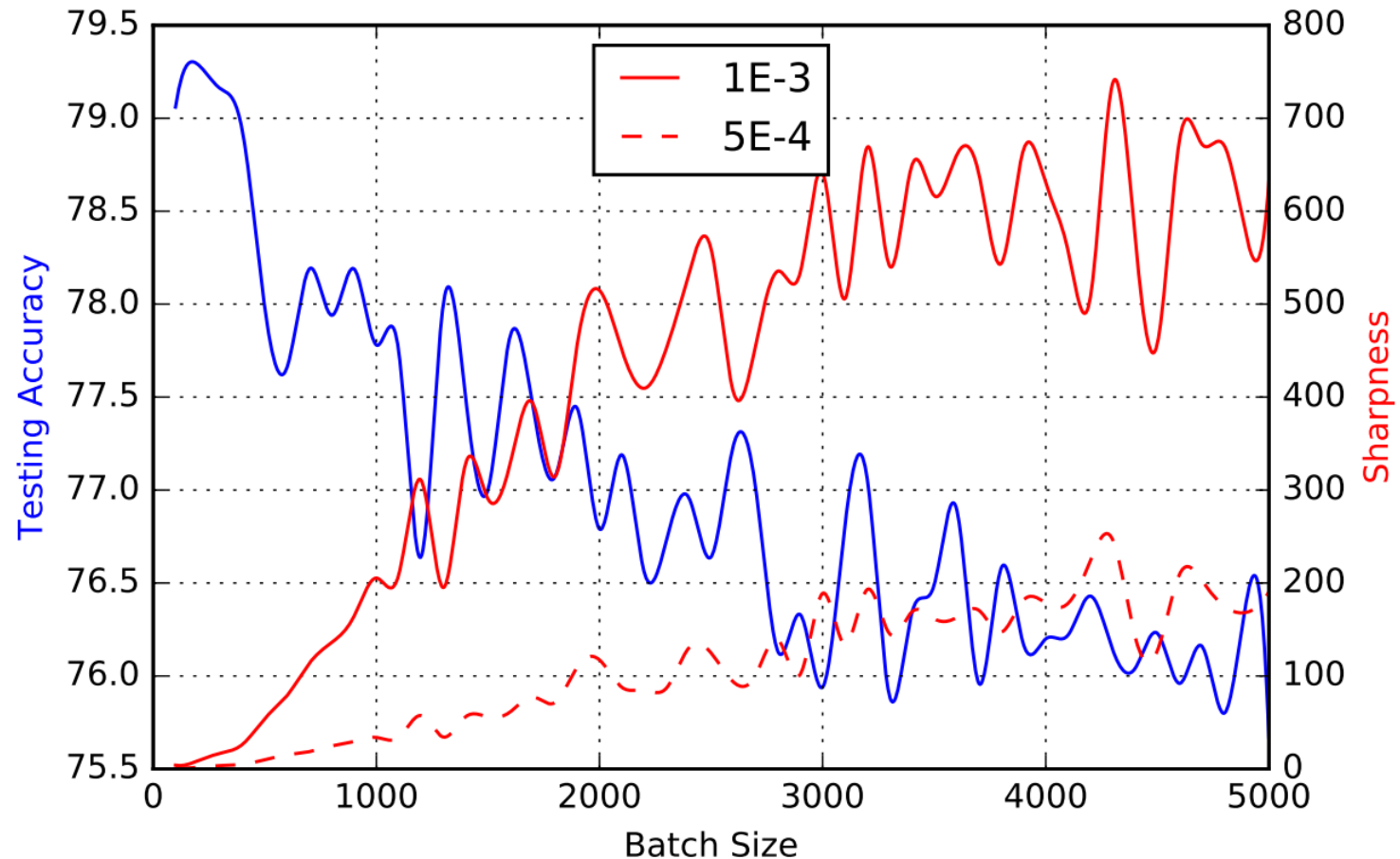
Name	Training Accuracy		Testing Accuracy	
	SB	LB	SB	LB
$F_1$	99.66% $\pm$ 0.05%	99.92% $\pm$ 0.01%	98.03% $\pm$ 0.07%	97.81% $\pm$ 0.07%
$F_2$	99.99% $\pm$ 0.03%	98.35% $\pm$ 2.08%	64.02% $\pm$ 0.2%	59.45% $\pm$ 1.05%
$C_1$	99.89% $\pm$ 0.02%	99.66% $\pm$ 0.2%	80.04% $\pm$ 0.12%	77.26% $\pm$ 0.42%
$C_2$	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%

		$\epsilon = 10^{-3}$		$\epsilon = 5 \cdot 10^{-4}$	
		SB	LB	SB	LB
SB = 256	$F_1$	$1.23 \pm 0.83$	$205.14 \pm 69.52$	$0.61 \pm 0.27$	$42.90 \pm 17.14$
	$F_2$	$1.39 \pm 0.02$	$310.64 \pm 38.46$	$0.90 \pm 0.05$	$93.15 \pm 6.81$
LB = 0.1 x data set	$C_1$	$28.58 \pm 3.13$	$707.23 \pm 43.04$	$7.08 \pm 0.88$	$227.31 \pm 23.23$
	$C_2$	$8.68 \pm 1.32$	$925.32 \pm 38.29$	$2.07 \pm 0.86$	$175.31 \pm 18.28$
	$C_3$	$29.85 \pm 5.98$	$258.75 \pm 8.96$	$8.56 \pm 0.99$	$105.11 \pm 13.22$
	$C_4$	$12.83 \pm 3.84$	$421.84 \pm 36.97$	$4.07 \pm 0.87$	$109.35 \pm 16.57$

# Batch Size v.s. Sharpness



# Batch Size v.s. Sharpness



# Concluding Remarks

# Summary

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

知道什麼意的network可以對generalization有幫助，就把他加入net

# 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 true label v.s. random label
- 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 想辦法把找flat的process加入training
- 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