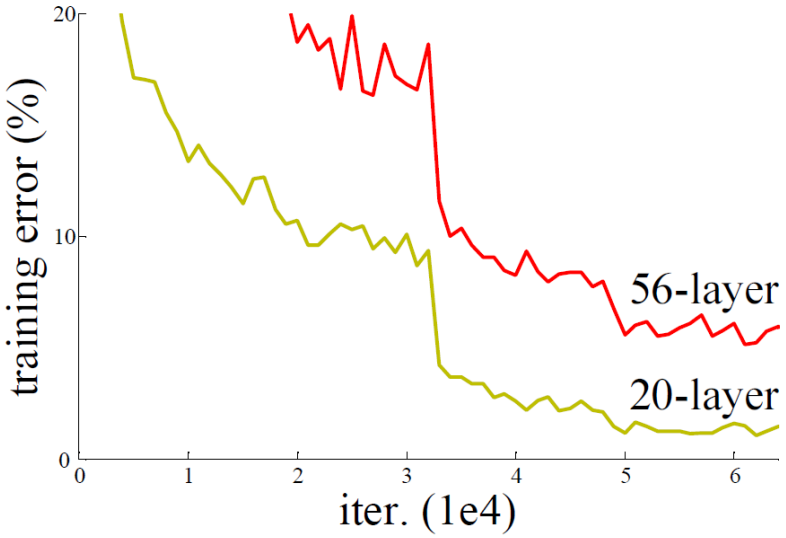
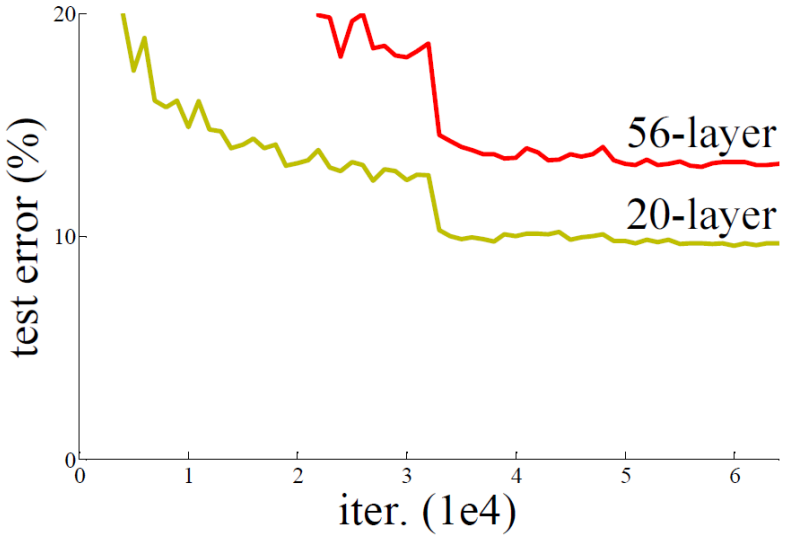
Sharp Minima Can Generalize For Deep Nets

MinKi Jo
KAIST

1. Intro

The residual learning make it possible to stack over 20 layers on the neural network. However, there was no clear explanation why the deep network without skip connection is extremely hard to train.

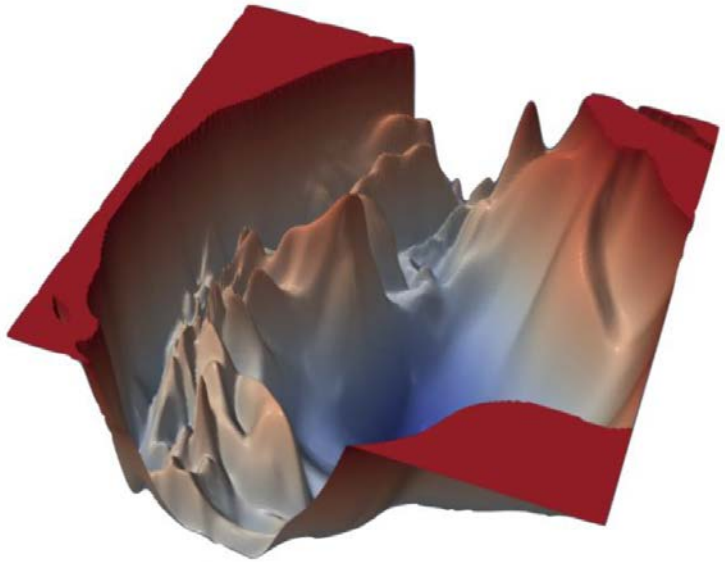




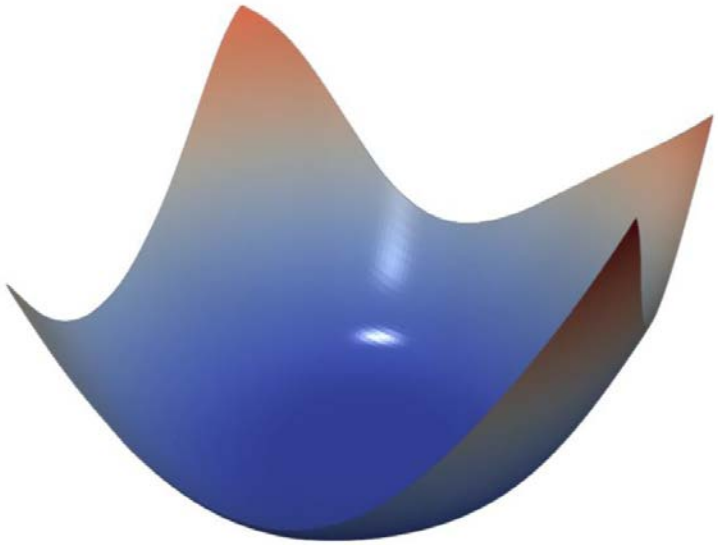
[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

1. Intro

One of the compelling reason of it is that there are many bad local optima, and it is hard to escape from there during the training.



(a) 110 layers, no skip connections



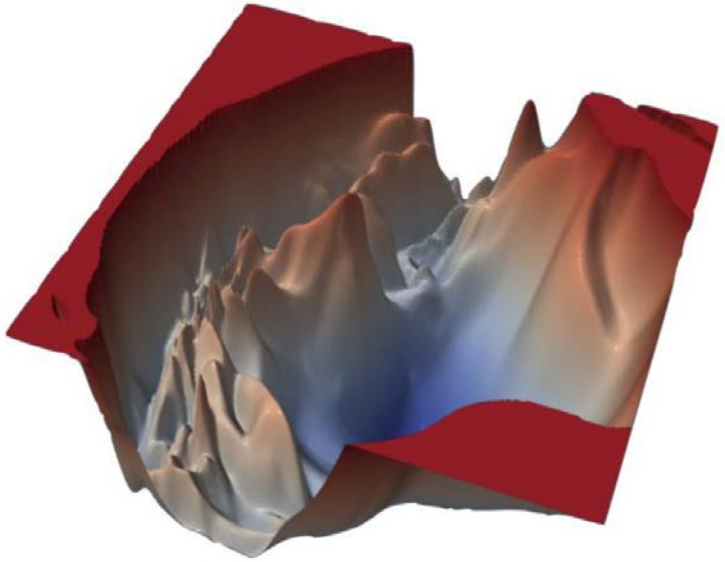
(b) DenseNet, 121 layers

The sharpness concept used to applied to analyze the the neural network

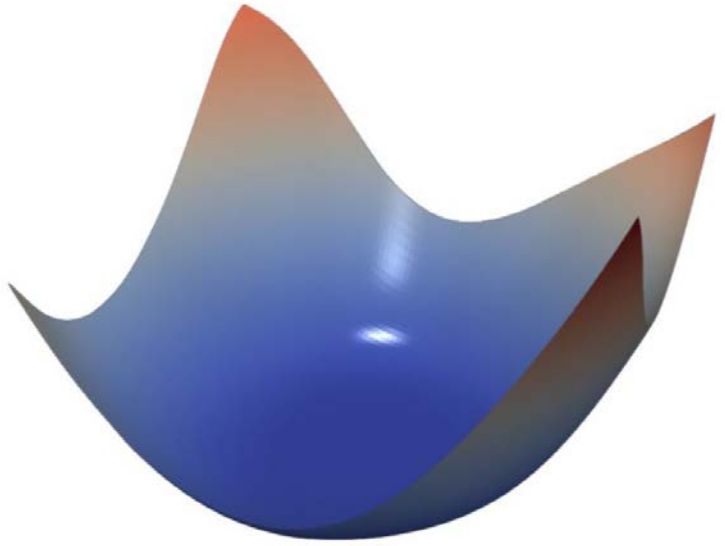
[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

1. Intro

However, this paper assert that the current concept of the flatness and sharpness has a loophole, which is **not perfect metric** to explain the generalization of the network.



(a) 110 layers, no skip connections



(b) DenseNet, 121 layers

[Laurent Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

2. Definition and Properties

There are 3 different definitions of the flatness and sharpness

that are suggested by previous studies.

1. **volume ϵ-flatness**
   $\mathrm{C}(\mathrm{L}, \theta, \epsilon)$: the largest connected set containg θ such that
   $\forall \theta' \in C(L, \theta, \epsilon), L(\theta') < L(\theta) + \epsilon$.
   The flatness is defined as the volume of the $\mathrm{C}(\mathrm{L}, \theta, \epsilon)$

2. **Hessian-based measure (curvature)**

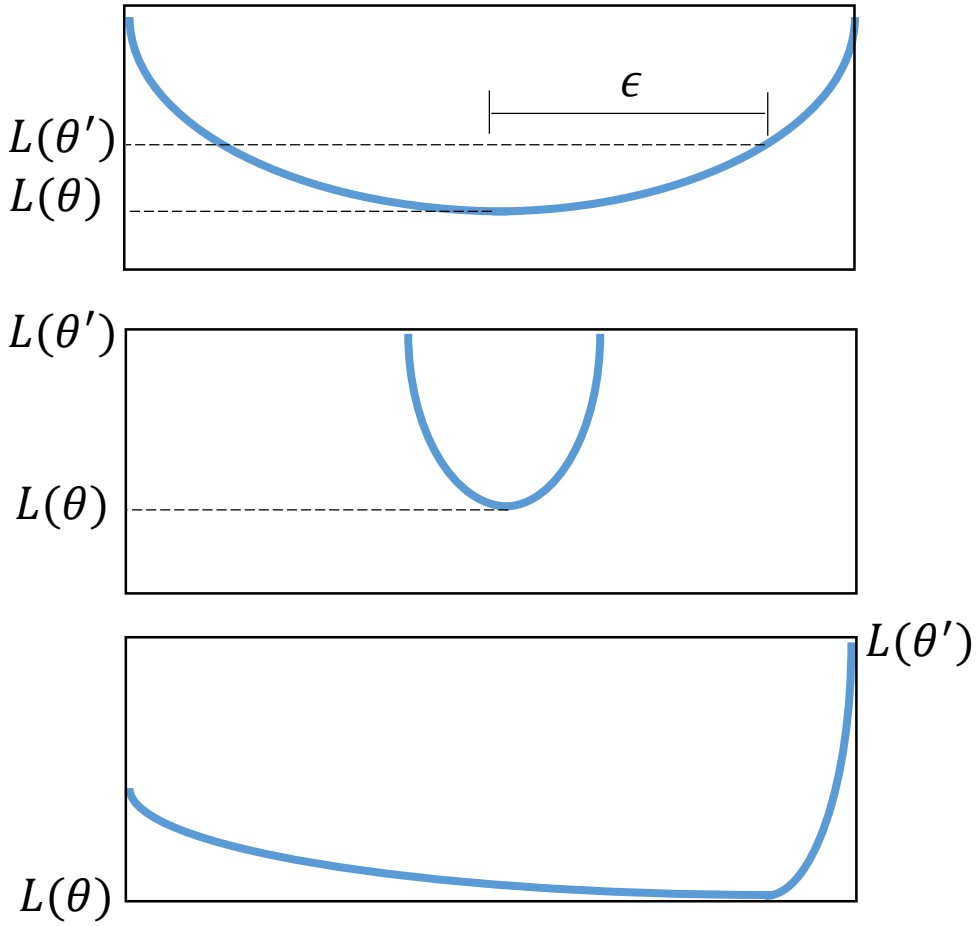
3. **ϵ-sharpness**
   When the $\mathrm{B}_2(\epsilon, \theta)$ is the ball that radius is ϵ *where the center is* θ,
   the sharpness defined as

$$\frac{\max_{\theta' \in B_2(\epsilon, \theta)} \left(L(\theta') - L(\theta) \right)}{1 + L(\theta)}$$

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

2. Definition and Properties

Examples



High Flatness

Low Sharpness

Low Flatness

High Sharpness

High Flatness

Low Sharpness

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

2. Definition and Properties

In order to give the explicit explanation, the author provides the specific setting of the network model and the problem

1. Supervised scalar output
2. Rectifier unit for the activation function
3. Linear output layer

Particularly, the activation function must be the **Relu function.**

Otherwise, following assertion does not hold.

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

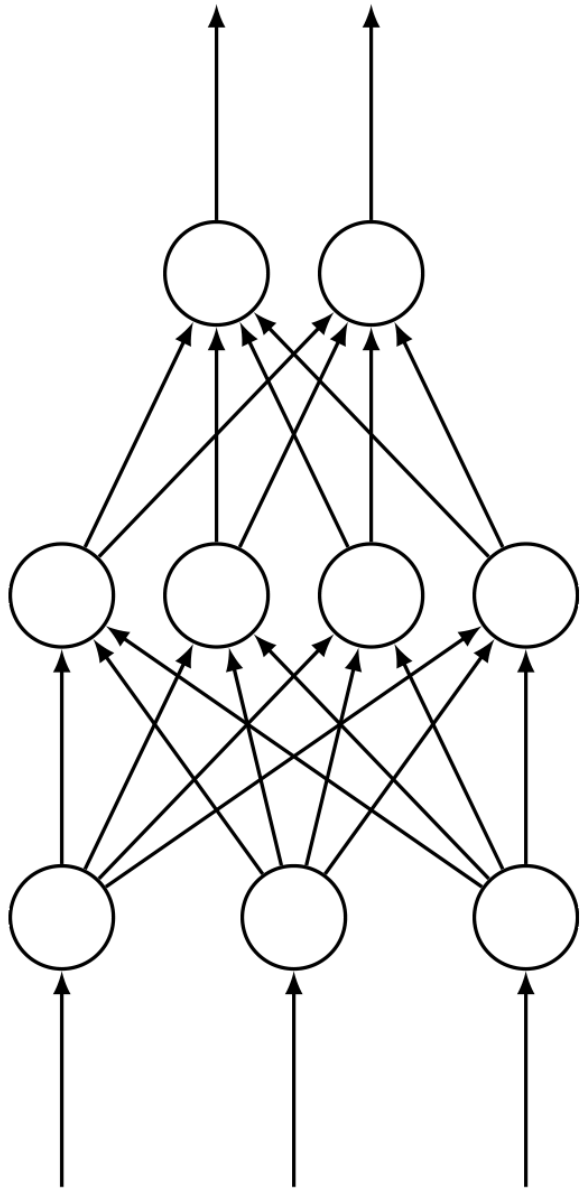
2. Definition and Properties

The rectified feedforward networks with a linear output layer can be represented by:

$$y = \phi_{rect}\Big(\phi_{rect}\big(\cdots\phi_{rect}(x\cdot\theta_1)\cdots\big)\cdot\theta_{K-1}\Big)\cdot\theta_K$$

where the θ_k is the weights of the kth layer, and ϕ_{rect} is the rectified activation function.

When the input domain is X and output domain is Y, the mapping function of the network is $F_\Theta: X \to Y$



[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

2. Definition and Properties

According to the non-negative homogeneous of the rectified function, following equation holds.

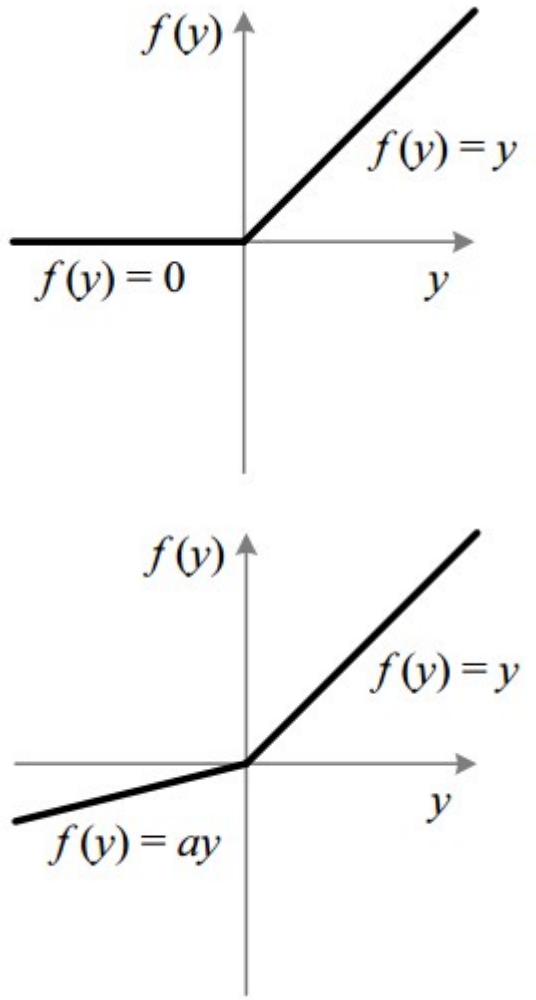
$$\phi_{rect}\big(x \cdot (\alpha\theta_1)\big) \cdot \theta_2 = \phi_{rect}(x \cdot \theta_1) \cdot (\alpha\theta_2)$$

Which means that after the weights of the two layer go through the ***α*-scale transform** such as,

$$T_\alpha : (\theta_1, \theta_2) \mapsto (\alpha\theta_1, \alpha^{-1}\theta_2)$$

the mapping function of the network is still same.

$$F_{(\theta_1,\theta_2)} = F_{(\alpha\theta_1,\alpha^{-1}\theta_2)}$$



[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

2. Definition and Properties

According to the non-negative homogeneous of the rectified function, following equation holds.

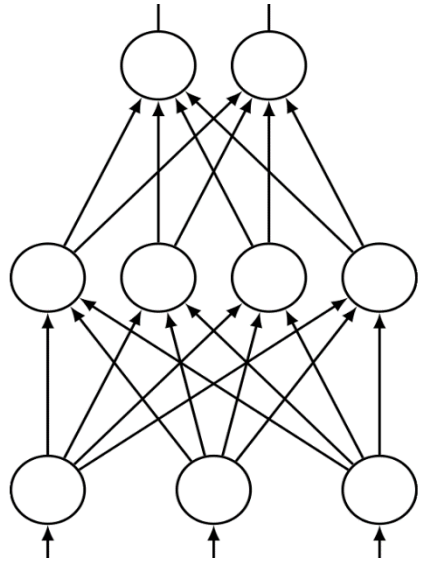
$$\phi_{rect}\big(x \cdot (\alpha\theta_1)\big) \cdot \theta_2 = \phi_{rect}(x \cdot \theta_1) \cdot (\alpha\theta_2)$$

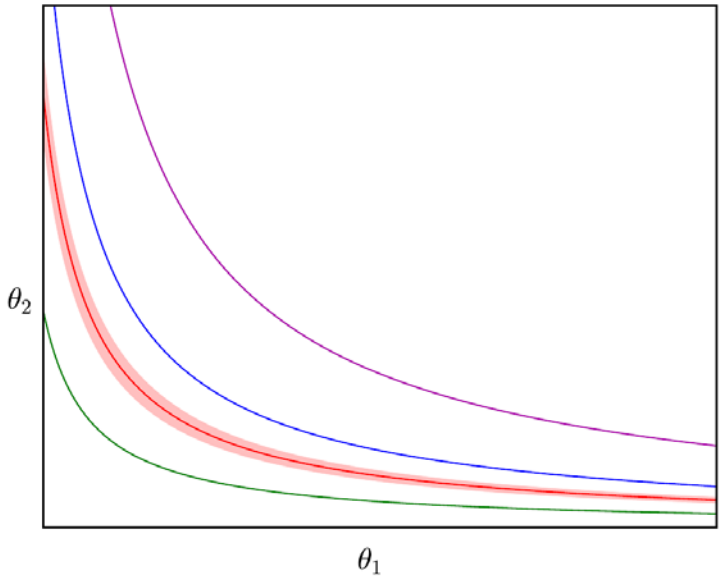
Which means that after the weights of the two layer go through the ***α*-scale transform** such as,

$$T_\alpha : (\theta_1, \theta_2) \mapsto (\alpha\theta_1, \alpha^{-1}\theta_2)$$

the mapping function of the network is still same.

$$F_{(\theta_1,\theta_2)} = F_{(\alpha\theta_1,\alpha^{-1}\theta_2)}$$





All points on the same line are the same network which has same mapping function but has different weight value.

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

3. Shortcoming of the flatness

The author provide the various transformation trick to show that

1. volume ϵ-flatness.
2. Hessian-based meausre
3. ϵ-sharpness

Are **not sufficient metric** to represent the generalization of the network.

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

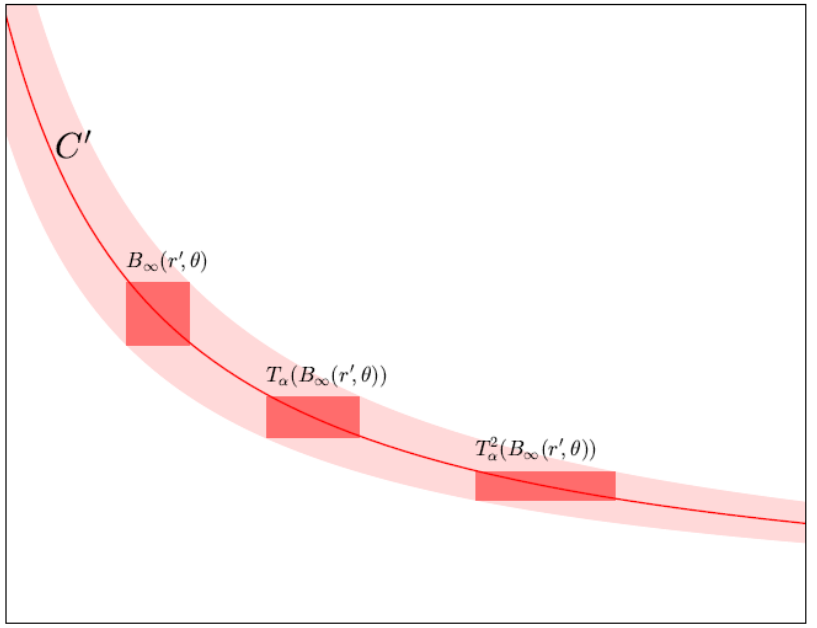
3.1 ϵ-flatness

Intuitively, you can find infinitely many different weight setting that has similar loss value.

Explicitly, for one layer network with Relu,

$(\forall \epsilon > 0)\ C(L, \theta, \epsilon)$ has an **infinite volume**.

In other word, according to this ϵ- flatness definition, **all minima are equally flat**.



The red boxes are in the C', has same volume and disjoint from each other.

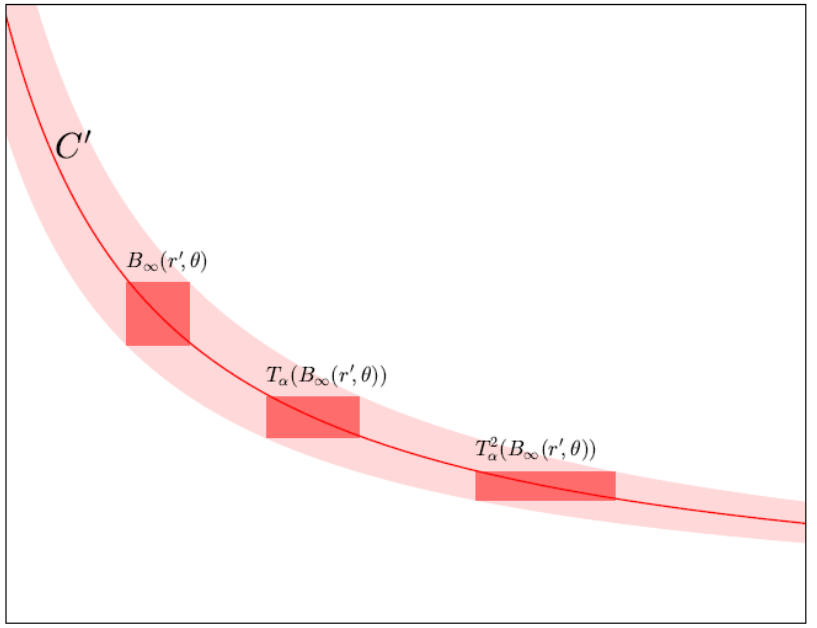
[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

3.1 ϵ-flatness

Let's say the ball $B_\infty(r, \theta)$ is in the $\mathrm{C}(L, \theta, \epsilon)$.

If the two dimension of the weights are same and the $\alpha = 2\frac{||\theta_1||_\infty + r}{||\theta_1||_\infty - r}$, the **volume of the $\boldsymbol{T_\alpha(B_\infty(r, \theta))}$ is same** with the original ball's one.

In addition, new ball is also in the C and **disjoint from the original ball**. Which means that by the scale transformation, you can find the **infinitely many disjoint balls with same volume** from the C. Therefore, the volume of the C can be represented by $\mathrm{V}(C) = v + v + v + \ \ldots = \infty$



The red boxes are in the C', has same volume and disjoint from each other.

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

3.2 Hessian based measure

The curvature of the space is heavily related to the **Hessian and Spectral norm**.

The author showed that the **Hessian of the loss can be modified** by the scale transformation.

Which means the curvature can be changed **without any difference of the model**.

$$(\nabla L)(\theta_1, \theta_2) = (\nabla L)(\alpha\theta_1, \alpha^{-1}\theta_2) \begin{bmatrix} \alpha \mathbb{I}_{n_1} & 0 \\ 0 & \alpha^{-1}\mathbb{I}_{n_2} \end{bmatrix}$$

$$\Leftrightarrow (\nabla L)(\alpha\theta_1, \alpha^{-1}\theta_2) = (\nabla L)(\theta_1, \theta_2) \begin{bmatrix} \alpha^{-1} \mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{bmatrix}$$

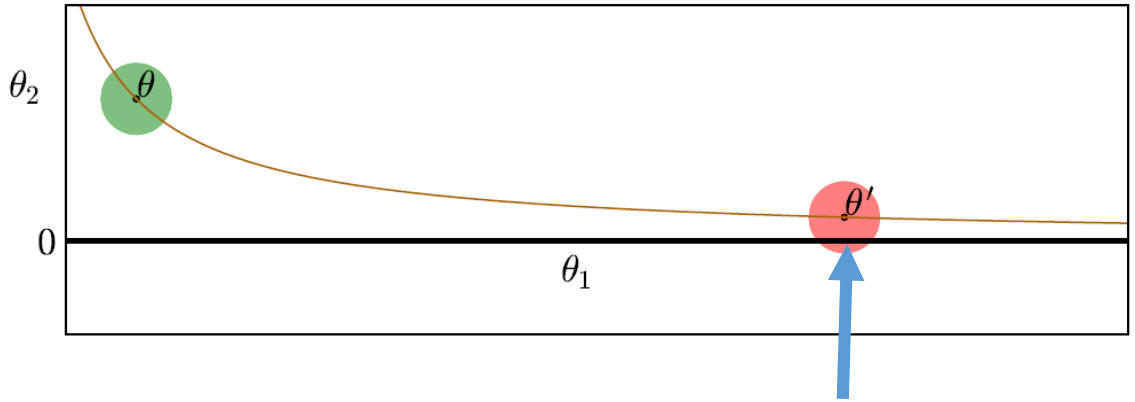
and

$$(\nabla^2 L)(\alpha\theta_1, \alpha^{-1}\theta_2)$$
$$= \begin{bmatrix} \alpha^{-1} \mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{bmatrix} (\nabla^2 L)(\theta_1, \theta_2) \begin{bmatrix} \alpha^{-1} \mathbb{I}_{n_1} & 0 \\ 0 & \alpha\mathbb{I}_{n_2} \end{bmatrix}.$$

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

3.3 ϵ-sharpness

In similar sense, when the $\left|\left|\theta_1\right|\right|_2 \leq \left|\left|\theta_2\right|\right|_2$ and $\alpha = \frac{\epsilon}{\left|\left|\theta_1\right|\right|_2}$, let's consider the parameter $\mathrm{T}_\alpha(\theta_1, \theta_2) = \left(\frac{\epsilon}{\left|\left|\theta_1\right|\right|_2}, \alpha^{-1}\theta_2\right)$



From that parameter, the $(0, \alpha^{-1}\theta_2)$ is in the $\mathrm{B}_2\left(\epsilon, T_\alpha(\theta)\right)$ which has the **maximum loss value**.

In other word, you can get extremely high sharpness without changing the network function.

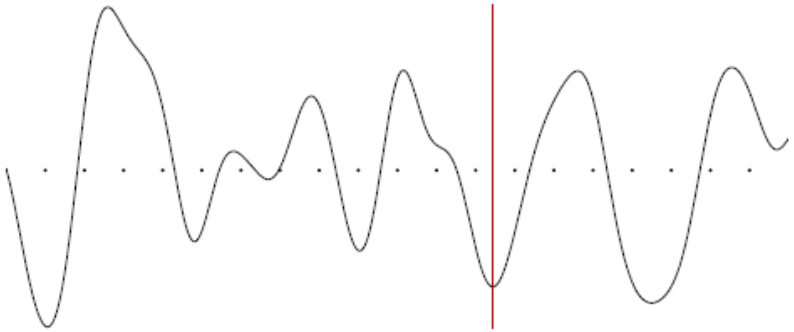
[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

3.4 Reparametrization

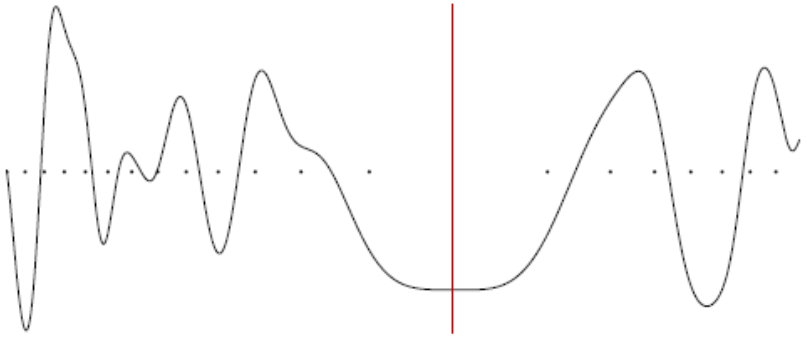
Previous sections showed the **fixed parametrization**. On the other hand, there is another way to obtain the **different sharpness property without changing the mapping function** of the network.

- Reparametrization

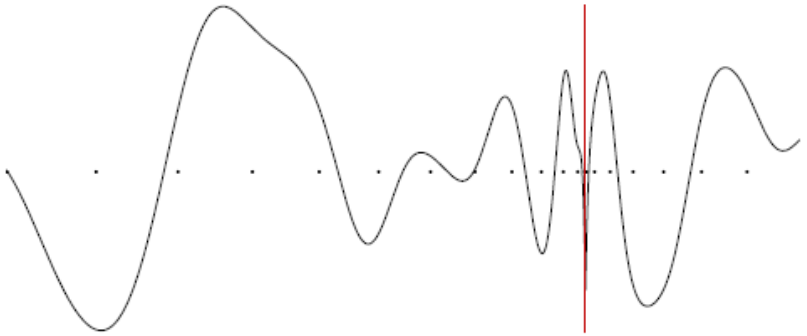
$$\hat{F}(x_1) = F(\phi(x_1)), \phi \; must \; be \; one\text{–}by\text{–}one \; mapping$$



(a) Loss function with default parametrization



(b) Loss function with reparametrization



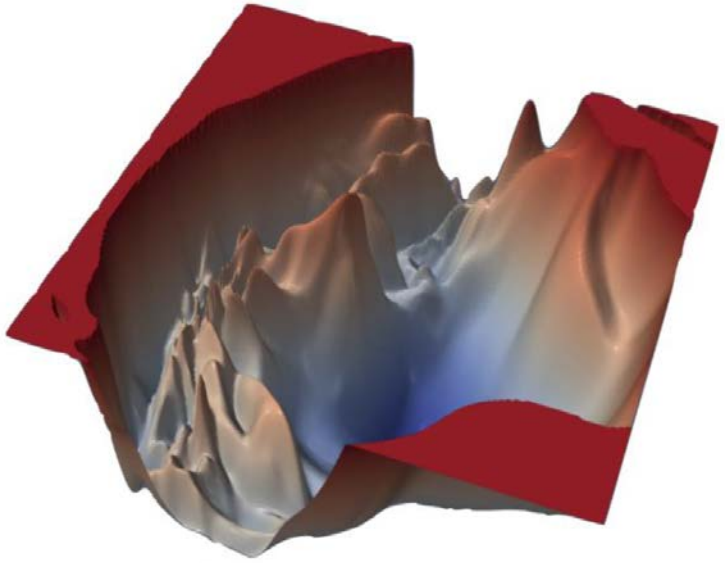
(c) Loss function with another reparametrization

[Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017

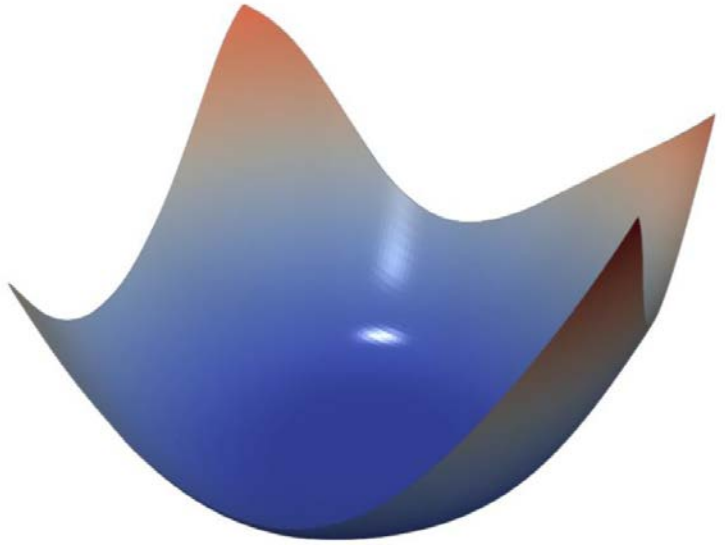
4. Conclusion

The author assert that the current metric for the sharpness and flatness is **insufficient** to represent the generalization of the network. However, this work **does not imply that those metrics are irrelevant or useless**. Actually those metrics have been showed good performance for the generalization by many experiments.

For example, the curvature concept of the flatness has improved the optimization of the network by **moving from the high curvature minima**. (Desjardins et al., 2015; Salimans & Kingma,2016)



(a) 110 layers, no skip connections



(b) DenseNet, 121 layers

[Laurent Dinh et al. 17] Sharp Minima Can Generalize For Deep Nets, ICML 2017