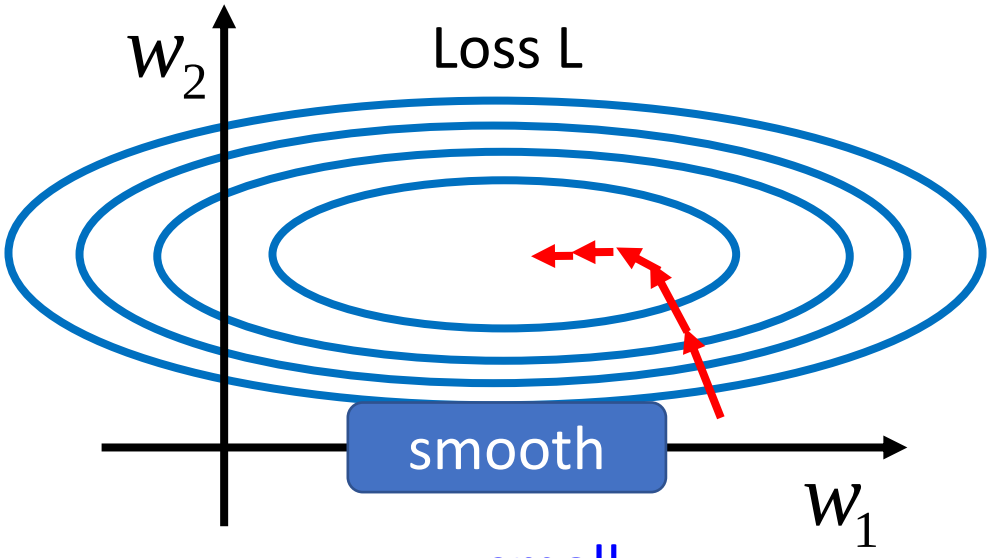
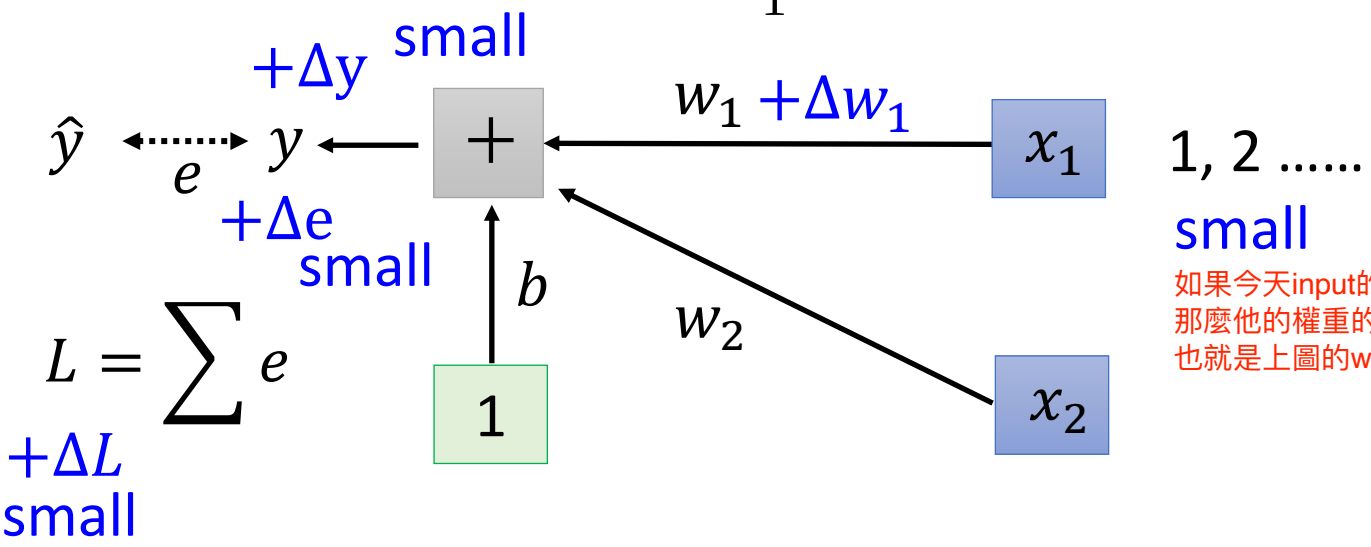
Quick Introduction of Batch Normalization

Hung-yi Lee 李宏毅

Changing Landscape

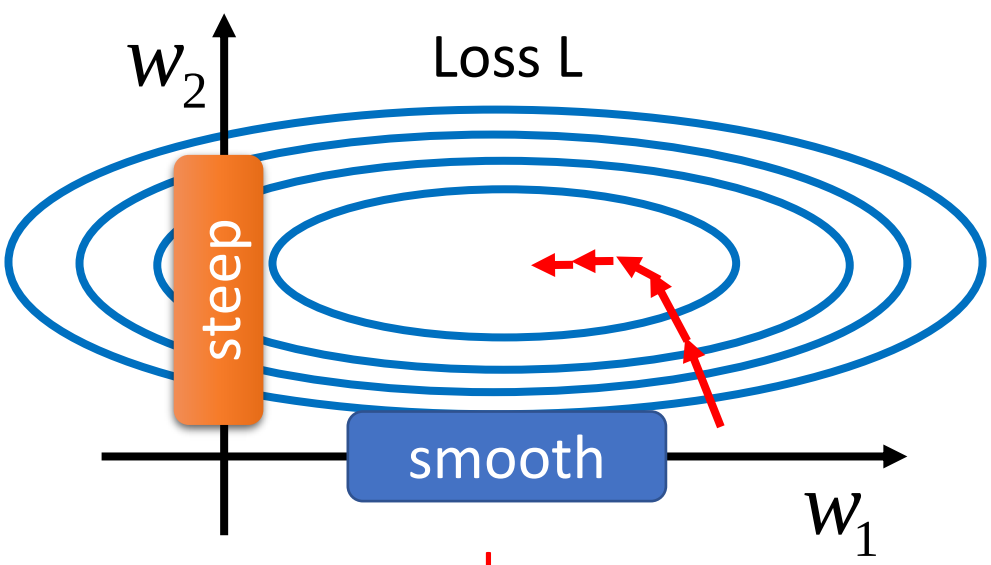


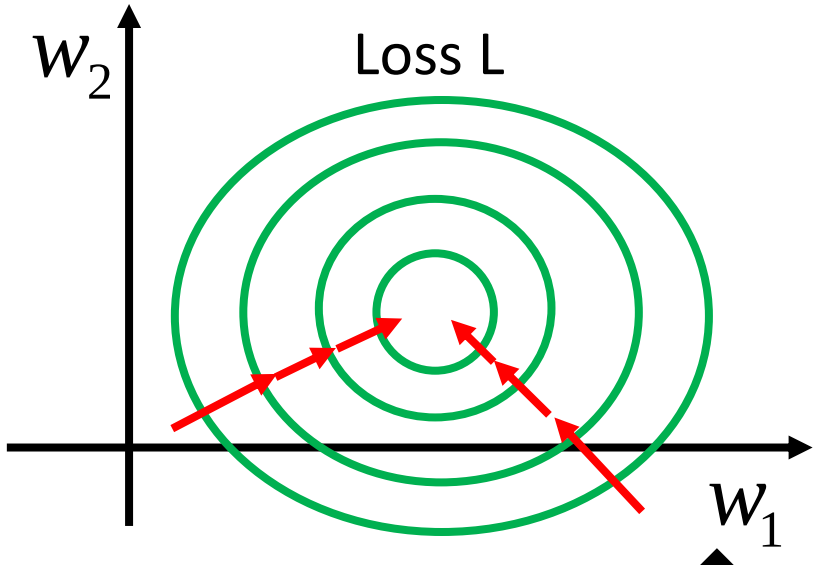
如果兩個weight的斜率差很多
（如左圖所示）
那麼就會不太好train

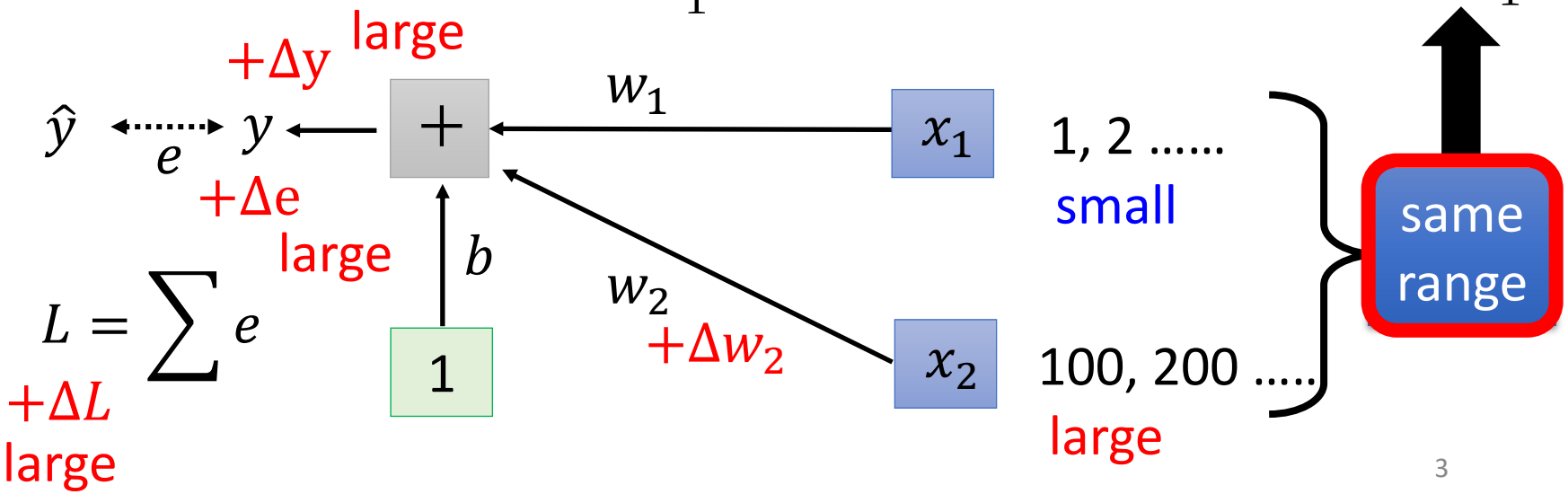


如果今天input的某一維度都很小的話
那麼他的權重的改變也會對loss很小
也就是上圖的w1

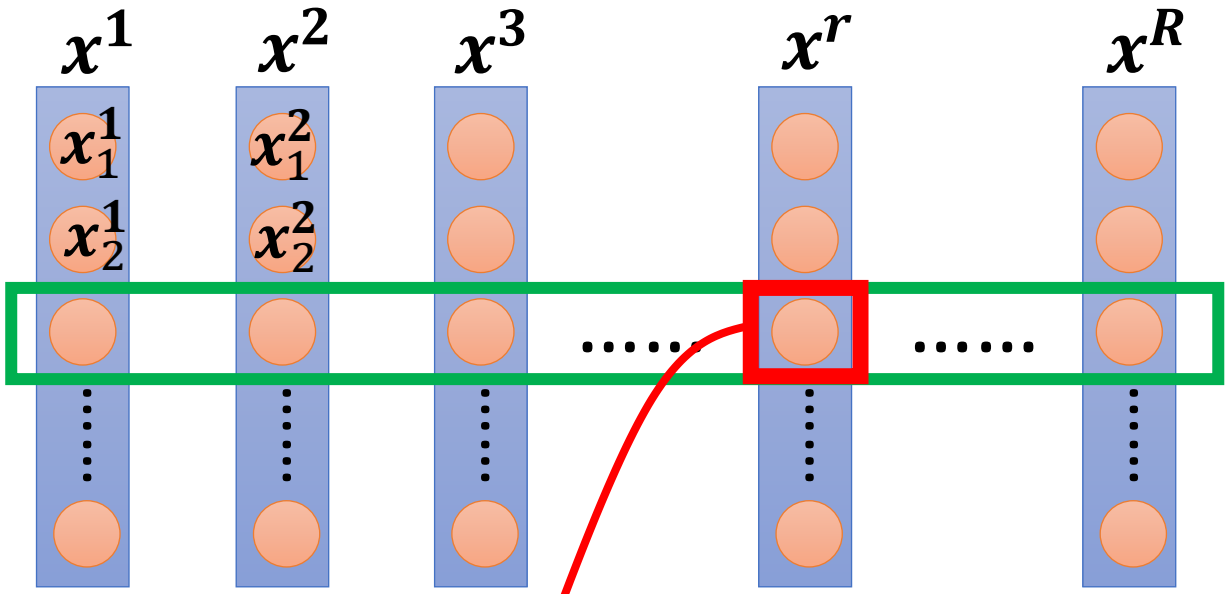
Changing Landscape







Feature Normalization



For each dimension i:

mean: m_i

standard deviation: σ_i

$$\tilde{x}_i^r \leftarrow \frac{x_i^r - m_i}{\sigma_i}$$

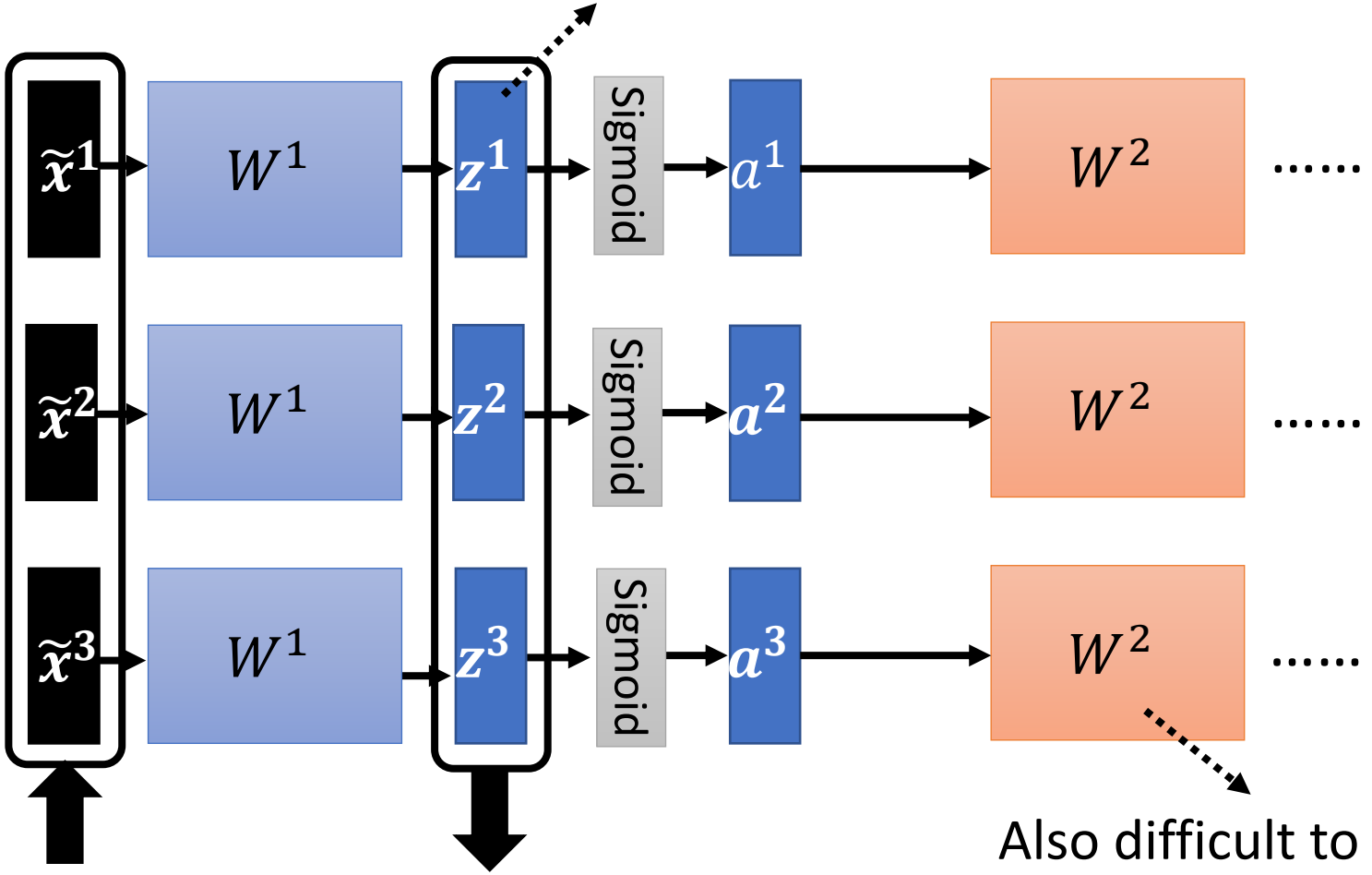
The means of all dims are 0, and the variances are all 1

In general, feature normalization makes gradient descent converge faster.

Considering Deep Learning

在實作上要再activation function前後做normalization其實都可以
如果是sigmoid則建議在之前做

Different dims have different ranges.

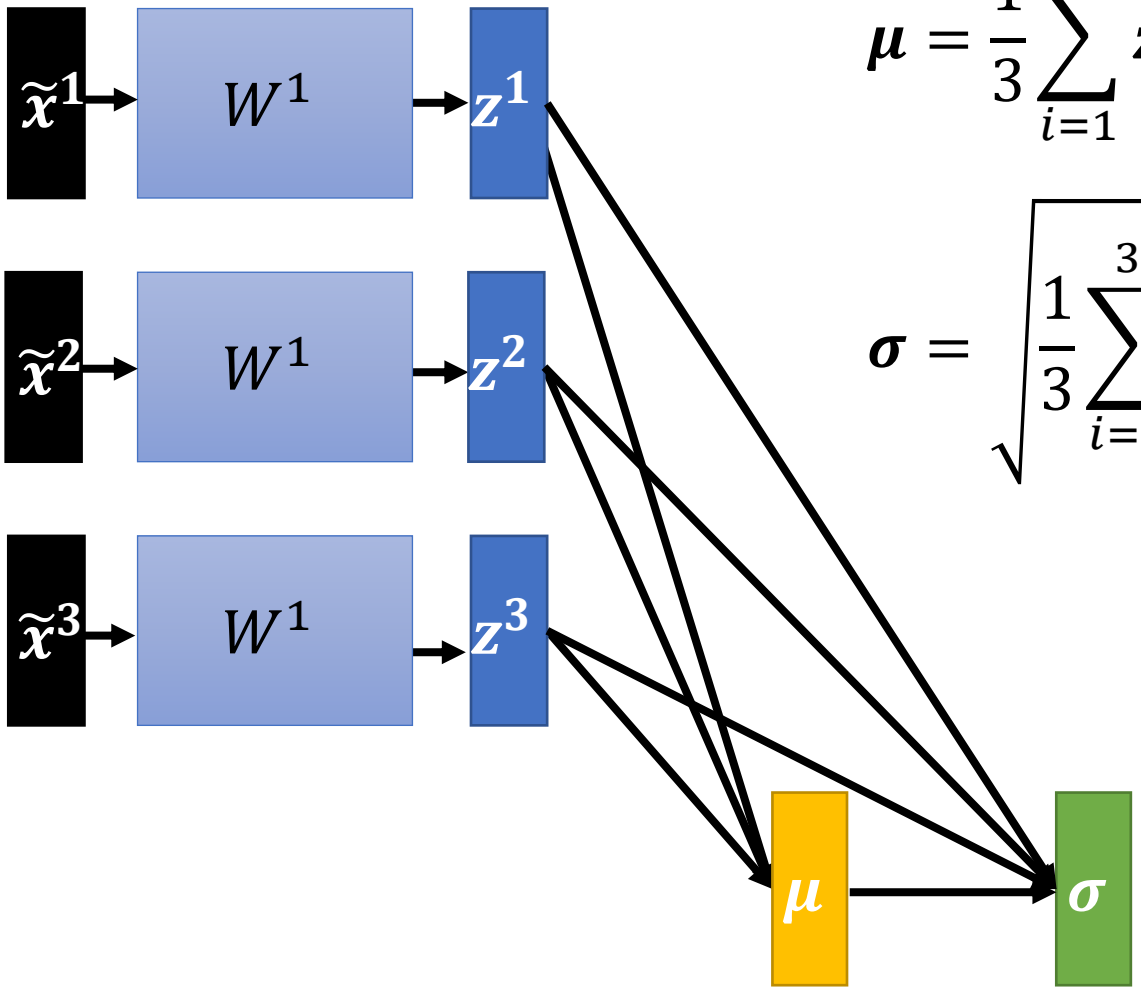


Also difficult to optimize

Feature Normalization

Also need normalization

Considering Deep Learning



$$\mu = \frac{1}{3}\sum_{i=1}^{3} z^i$$

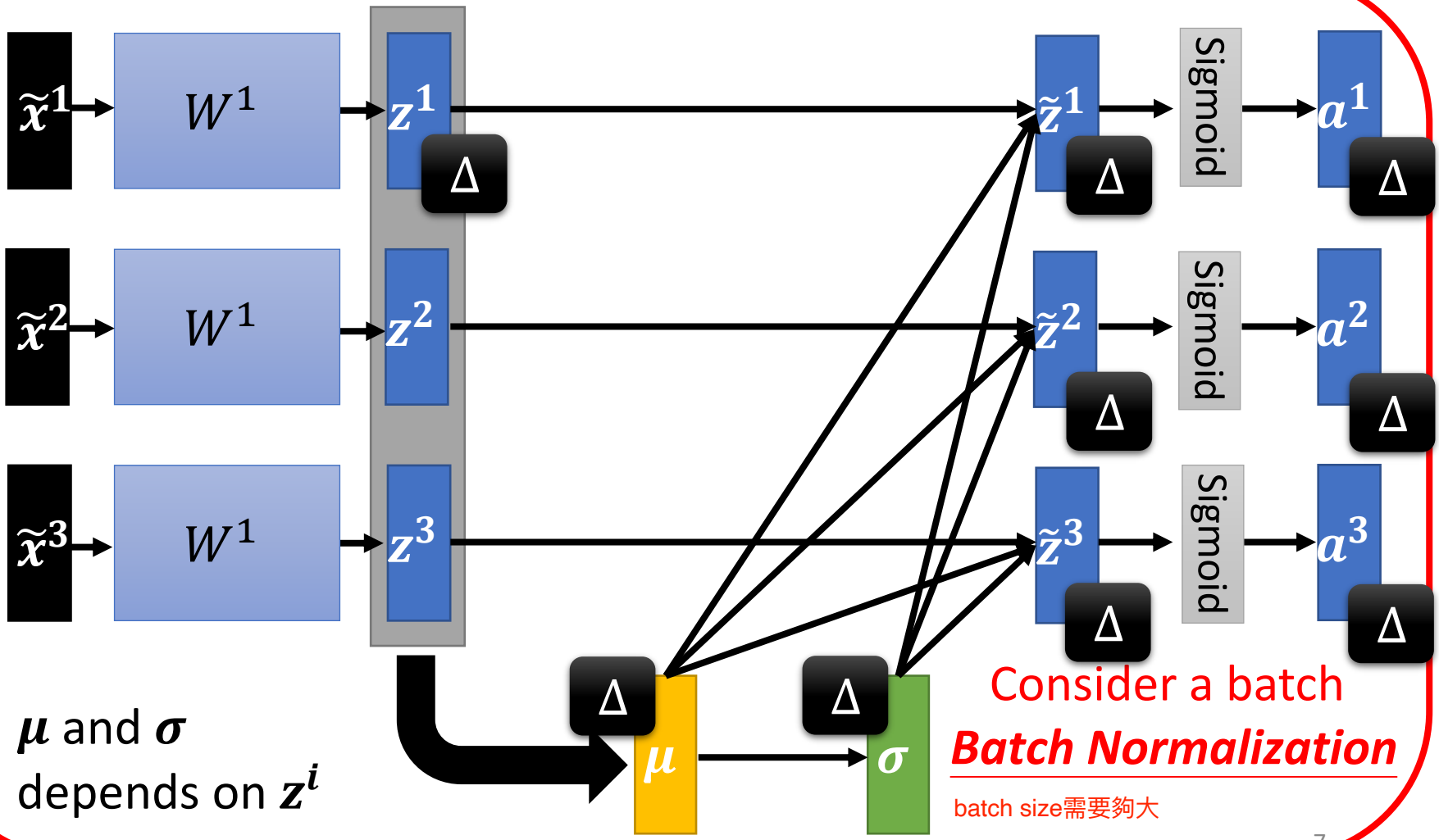
$$\sigma = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(z^i - \mu)^2}$$

Considering Deep Learning

若要對每一個layer都去做normalize，則需要把所有的data都算到那一層layer才能去做運算
而且這樣對testing會有問題（你要拿誰去算mean and variance?）

$$\tilde{z}^i = \frac{z^i - \mu}{\sigma}$$

This is a large network!

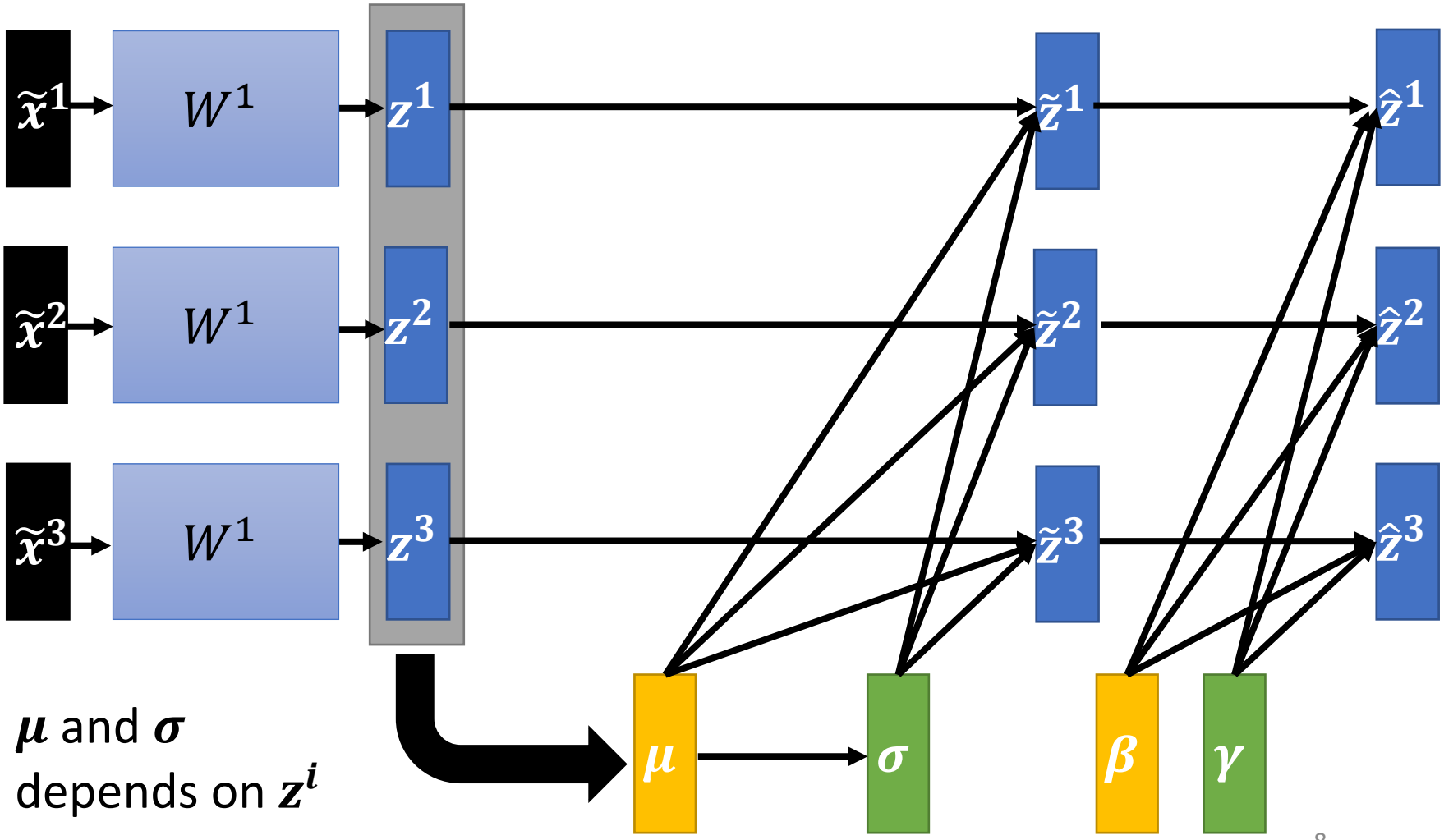


因為做normalization，等於是對nn做限制
例如平均會是零，那這可能會帶來一些不好的影響
因此再將算出來的值乘上gamma再加上beta

Batch normalization

$$\tilde{z}^i = \frac{z^i - \mu}{\sigma}$$

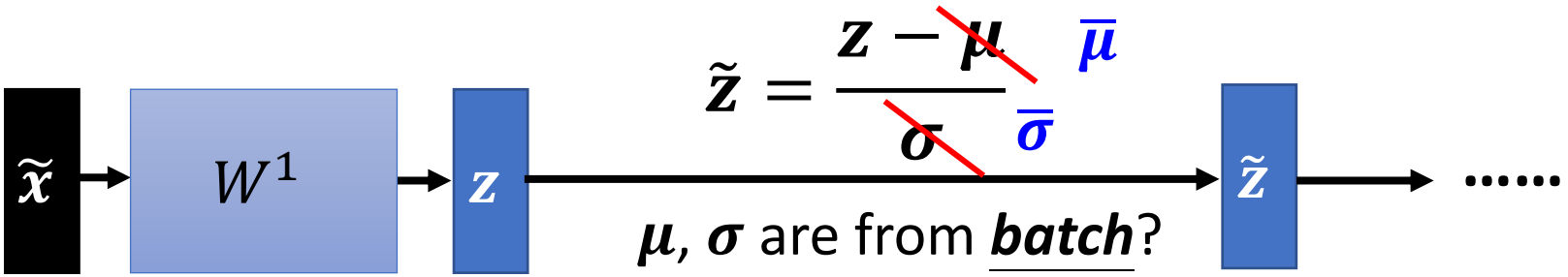
$$\hat{z}^i = \gamma \odot \tilde{z}^i + \beta$$



μ and σ depends on z^i

Batch normalization – Testing

testing的時候不一定能讓你在湊滿一個batch之後（有辦法算mean和variance）才算答案



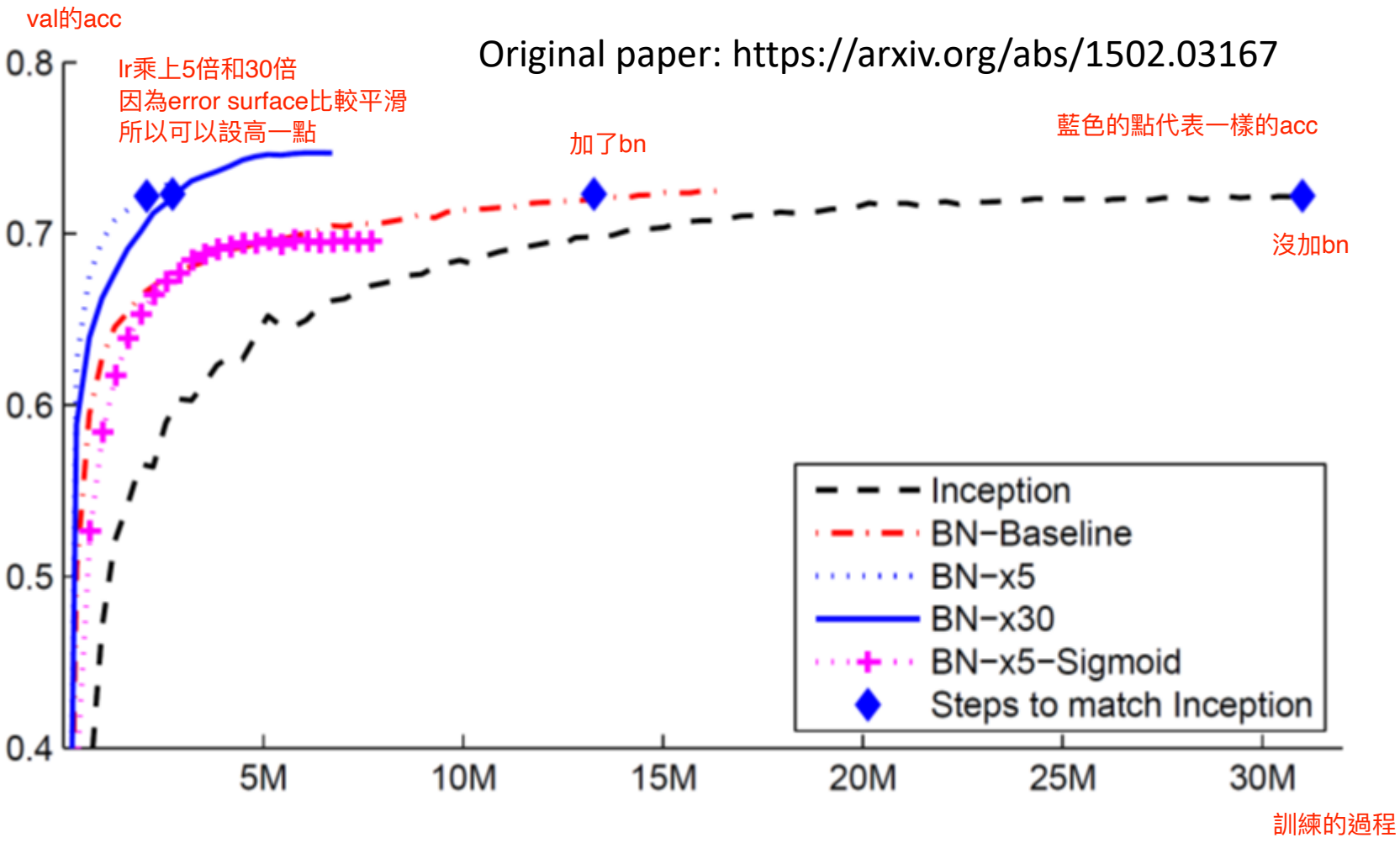
We do not always have ***batch*** at testing stage.

Computing the moving average of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of the batches during training.

$$\mu^1 \quad \mu^2 \quad \mu^3 \quad \ldots\ldots \quad \mu^t$$

$$\bar{\mu} \leftarrow p\bar{\mu} + (1-p)\mu^t$$

Batch normalization

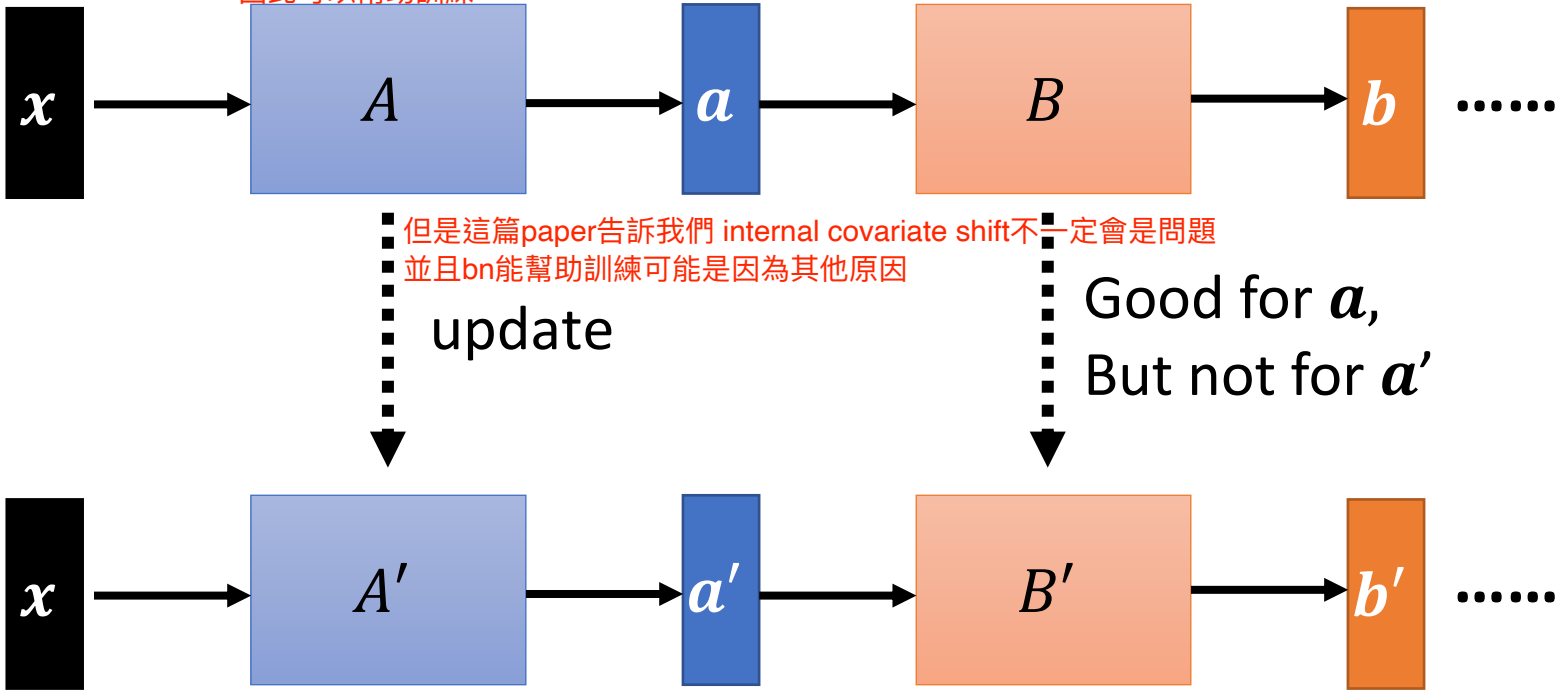


Internal Covariate Shift?

How Does Batch Normalization Help Optimization?

有可能這一次的update只對這次的data有幫助(internal covariate shift)
但是對其他data沒幫助
但若用bn後可以使data的數值分佈相近
因此可以幫助訓練

https://arxiv.org/abs/1805.11604



Batch normalization make a and a' have similar statistics.

Experimental results do not support the above idea.

Internal Covariate Shift?

How Does Batch Normalization Help Optimization?

https://arxiv.org/abs/1805.11604

bn能讓error surface變得比較好train也有實驗的支持了

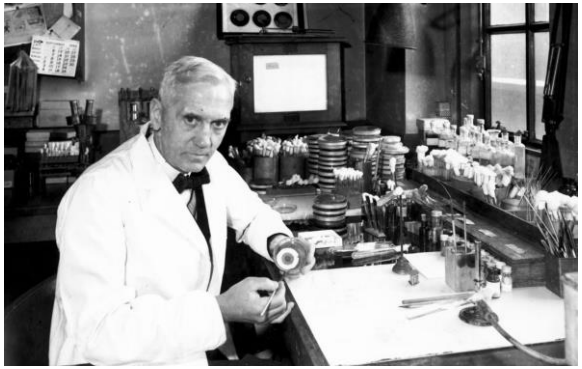
Experimental results (and theoretically analysis) support batch normalization change the landscape of error surface.

and 12 of Appendix B.) This suggests that the positive impact of BatchNorm on training might be somewhat serendipitous. Therefore, it might be valuable to perform a principled exploration of the design space of normalization schemes as it can lead to better performance.

serendipitous（偶然的)

這篇文章的作者認為bn也許是像penicillin一樣
偶然被發現的產物
（可能和原本想的理論方向不同）
但總之對訓練過程有幫助

penicillin



To learn more ……

- Batch Renormalization
  - https://arxiv.org/abs/1702.03275
- Layer Normalization
  - https://arxiv.org/abs/1607.06450
- Instance Normalization
  - https://arxiv.org/abs/1607.08022
- Group Normalization
  - https://arxiv.org/abs/1803.08494
- Weight Normalization
  - https://arxiv.org/abs/1602.07868
- Spectrum Normalization
  - https://arxiv.org/abs/1705.10941


Q&A