这堂是讲 training 的 tips

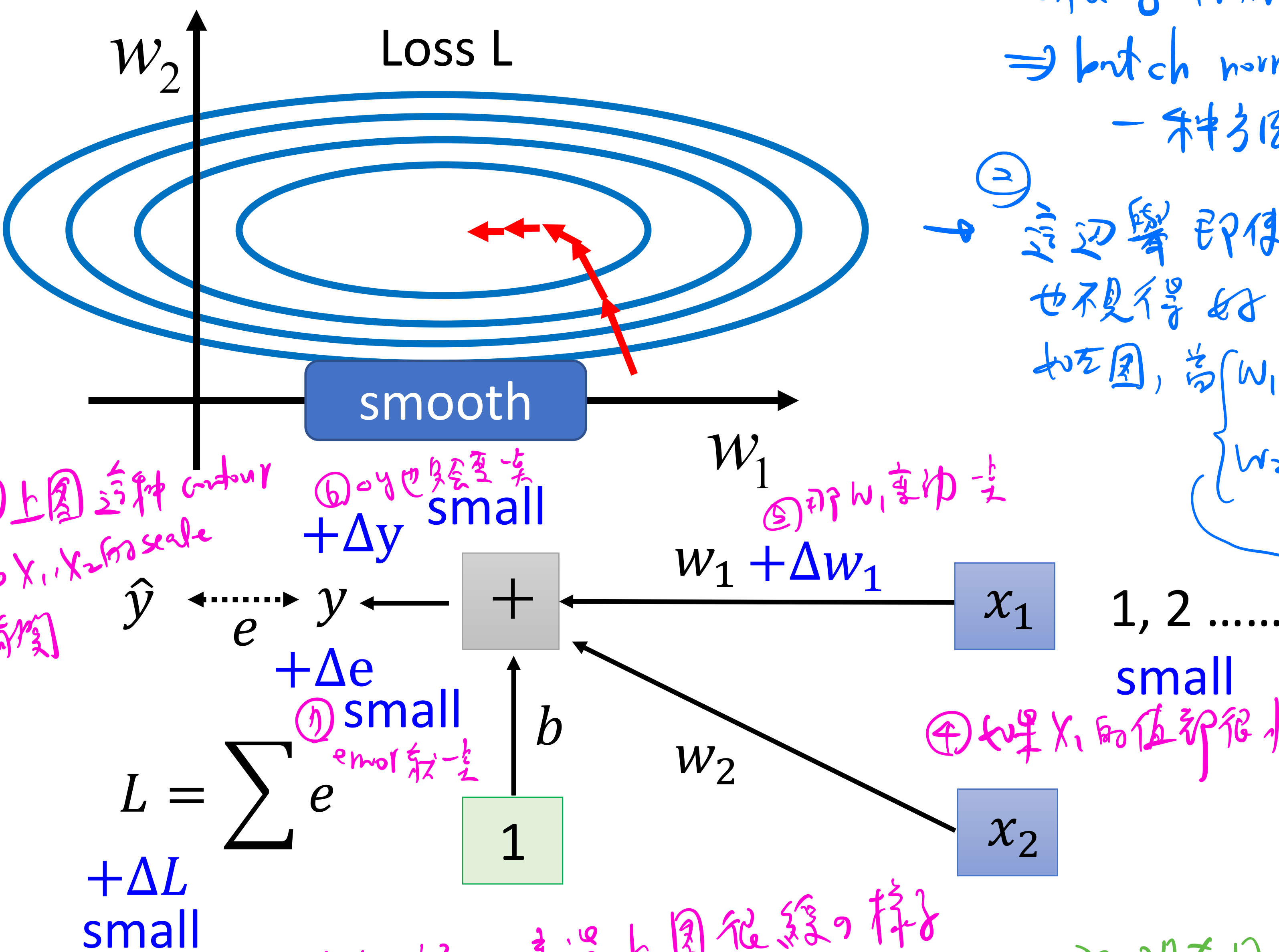
等下的 CNN, seq2seq 都会用到

这堂右手讲

Quick Introduction of Batch Normalization

Hung-yi Lee 李宏毅

Changing Landscape



①error surface 很崎嶇的時候比較難 train，那我們能不能直接把山剷平，讓他變的比較好 train？
⇒ batch normalization 就是一種方法

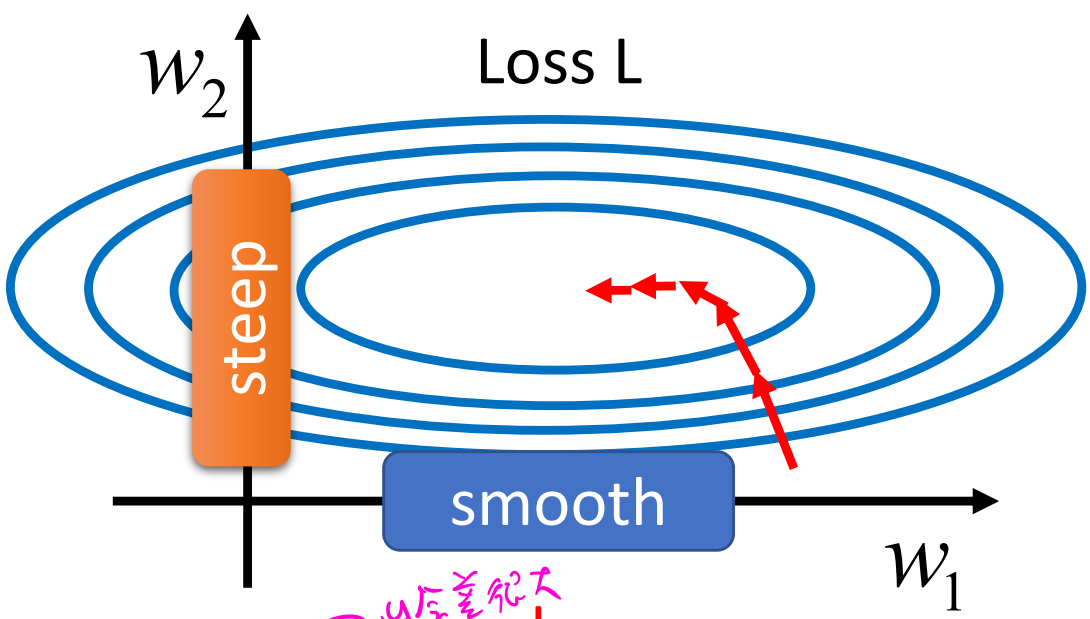
②這邊舉即使是 convex 的問題也很難 train，
如左圖，當 w_1 較緩（走很大步，loss 才降一點），w_2 較陡（走一小步，loss 就降很多）
用固定的 learning rate 也沒用，
∴才有各種調 lr 的方式
e.g. Adam, momentum

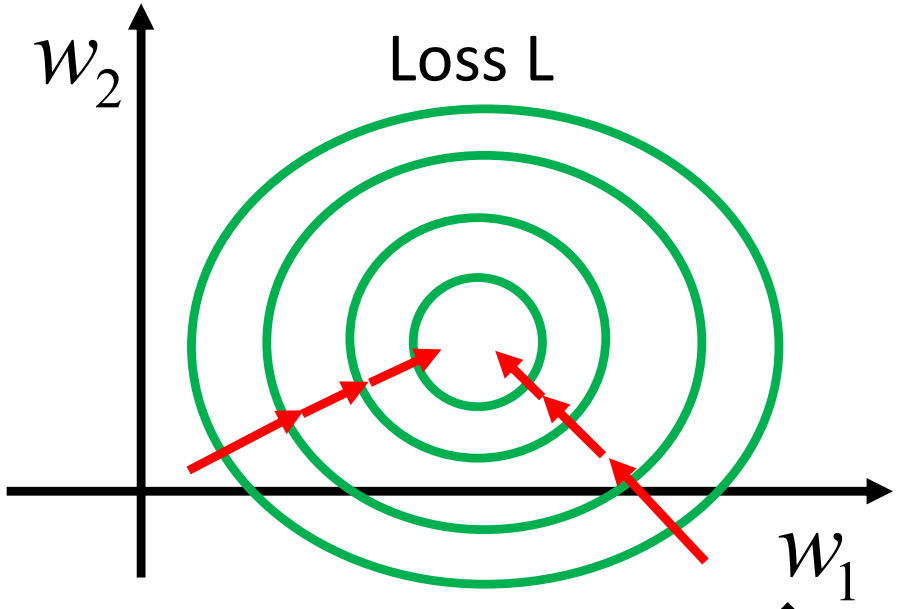
③上圖這種 contour 和 x_1, x_2 的 scale 有關
④如果 x_1 的值都很小
⑤那 w_1 變動一點
⑥ y 也只會變一點
⑦ error 就一點
⑧ Loss 也只變一點 ∴ 就是上圖很緩的樣子

那現在用另一個角度去想，

⑭ 直接把 error surface 改成 → 那即使是最基本的 gradient descent 也能好 train 起来

Changing Landscape





⑪ y 差很大 $+\Delta y$ large

$\hat{y}$ ⋯⋯ y ← $+$ ← w_1 ← x_1 1, 2 ..... small

e

$+\Delta e$

⑫ large

b

w_2

$+\Delta w_2$

⑩ 即 w_2 变一点点

x_2 100, 200 ..... large

same range

⑮ 依据很简单，

$L = \sum e$

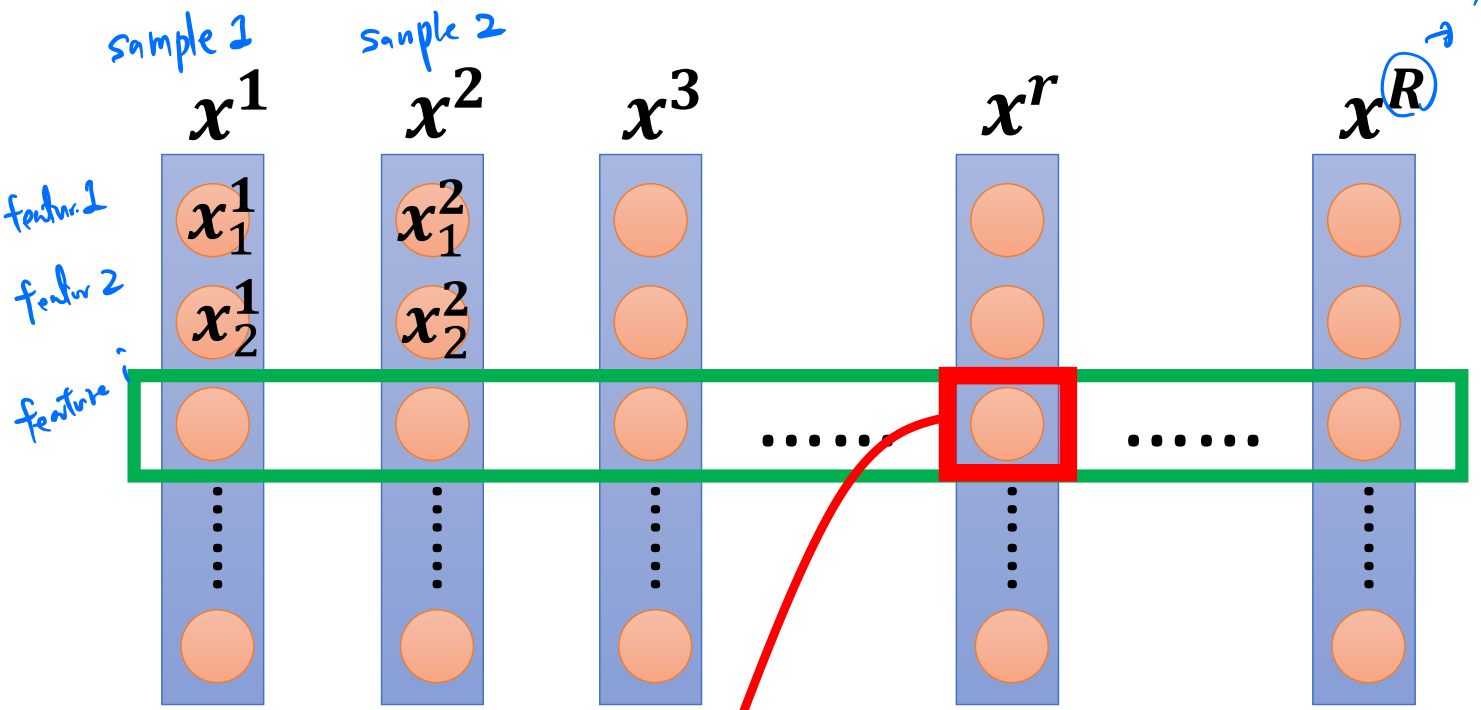
$+\Delta L$

large

⑬

⑨ 如果 x_2 值很大

Feature Normalization



For each dimension i:

mean: m_i

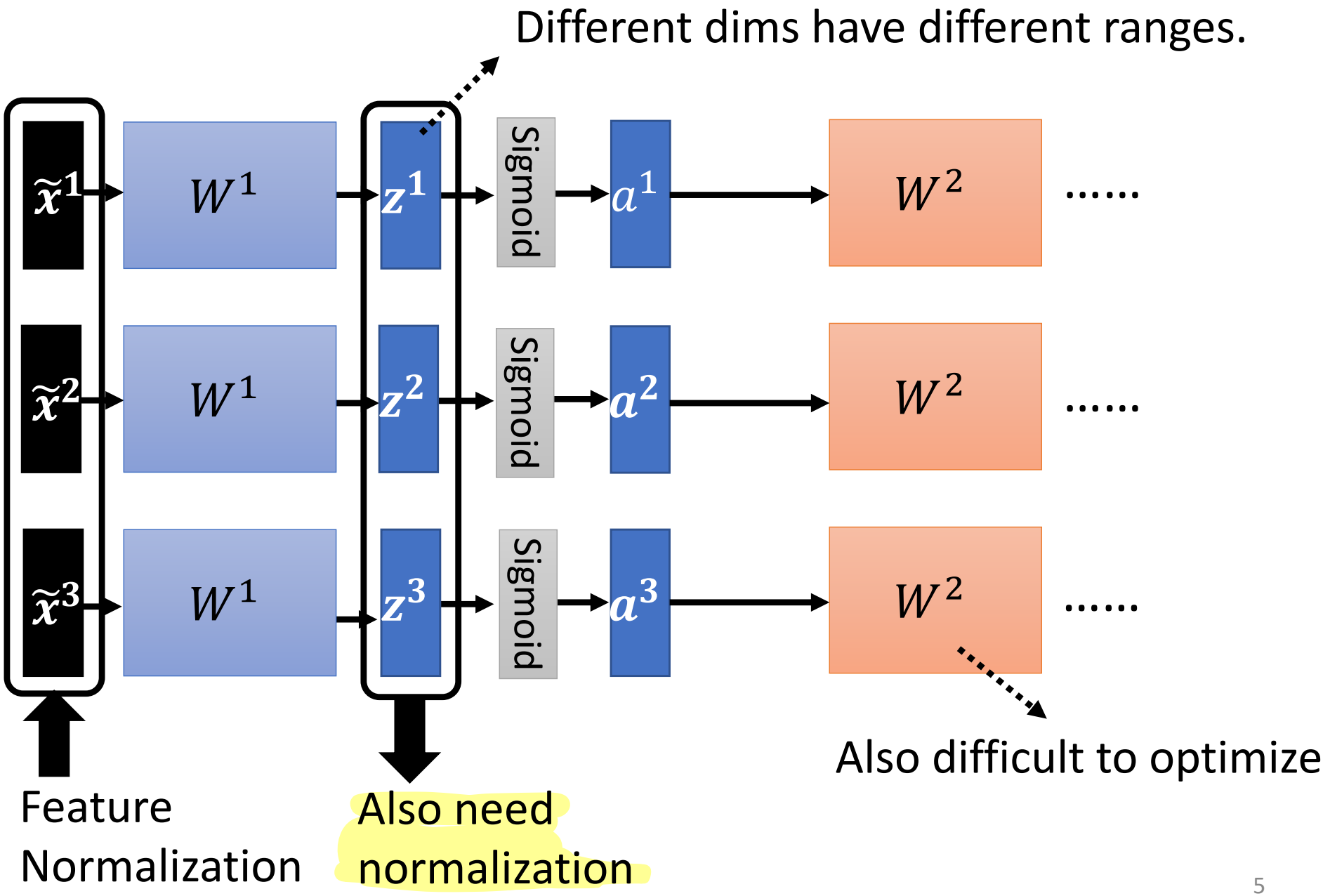
standard deviation: σ_i

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

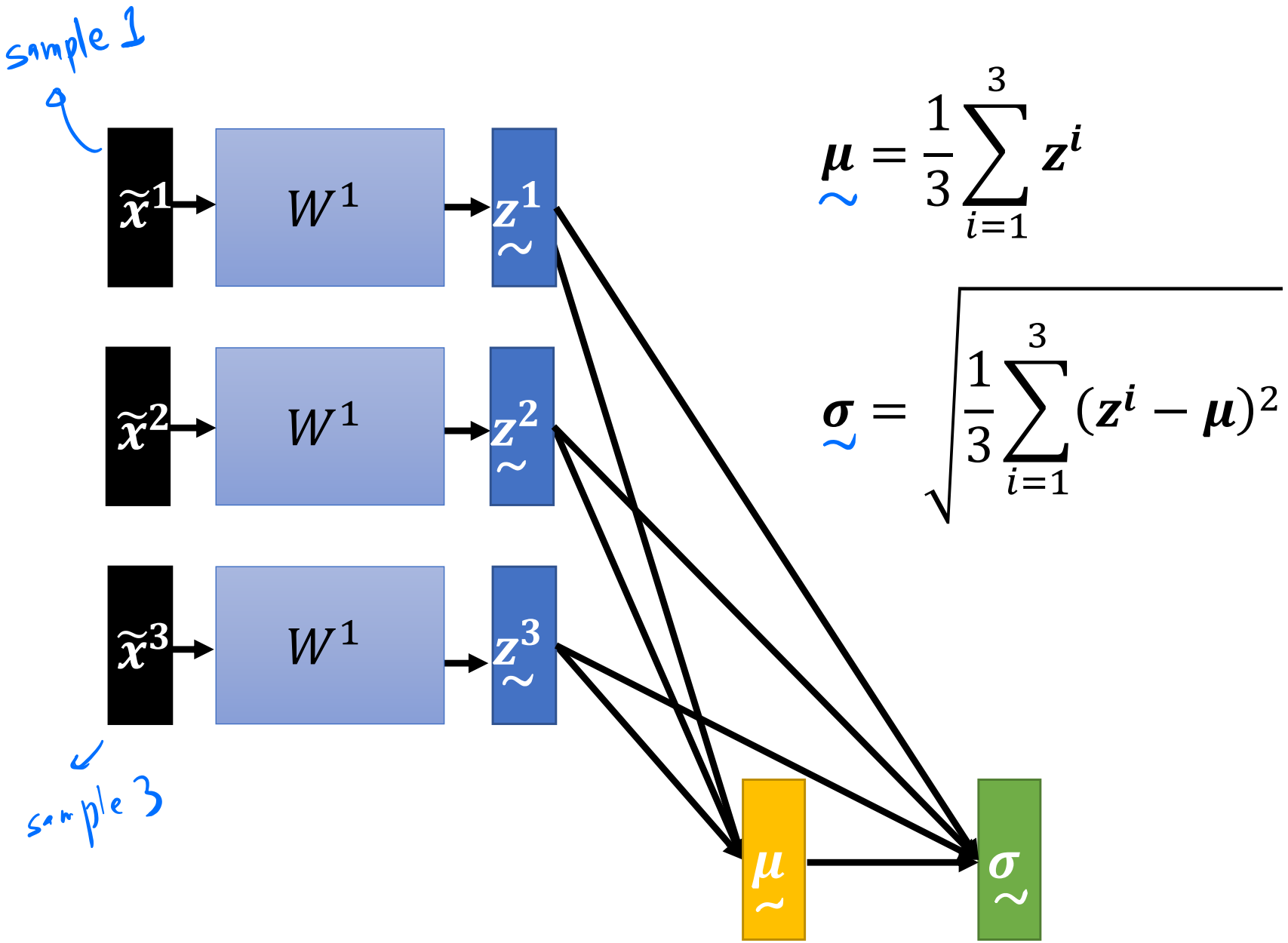
The means of all dims are 0, and the variances are all 1

In general, feature normalization makes gradient descent converge faster.

Considering Deep Learning



Considering Deep Learning



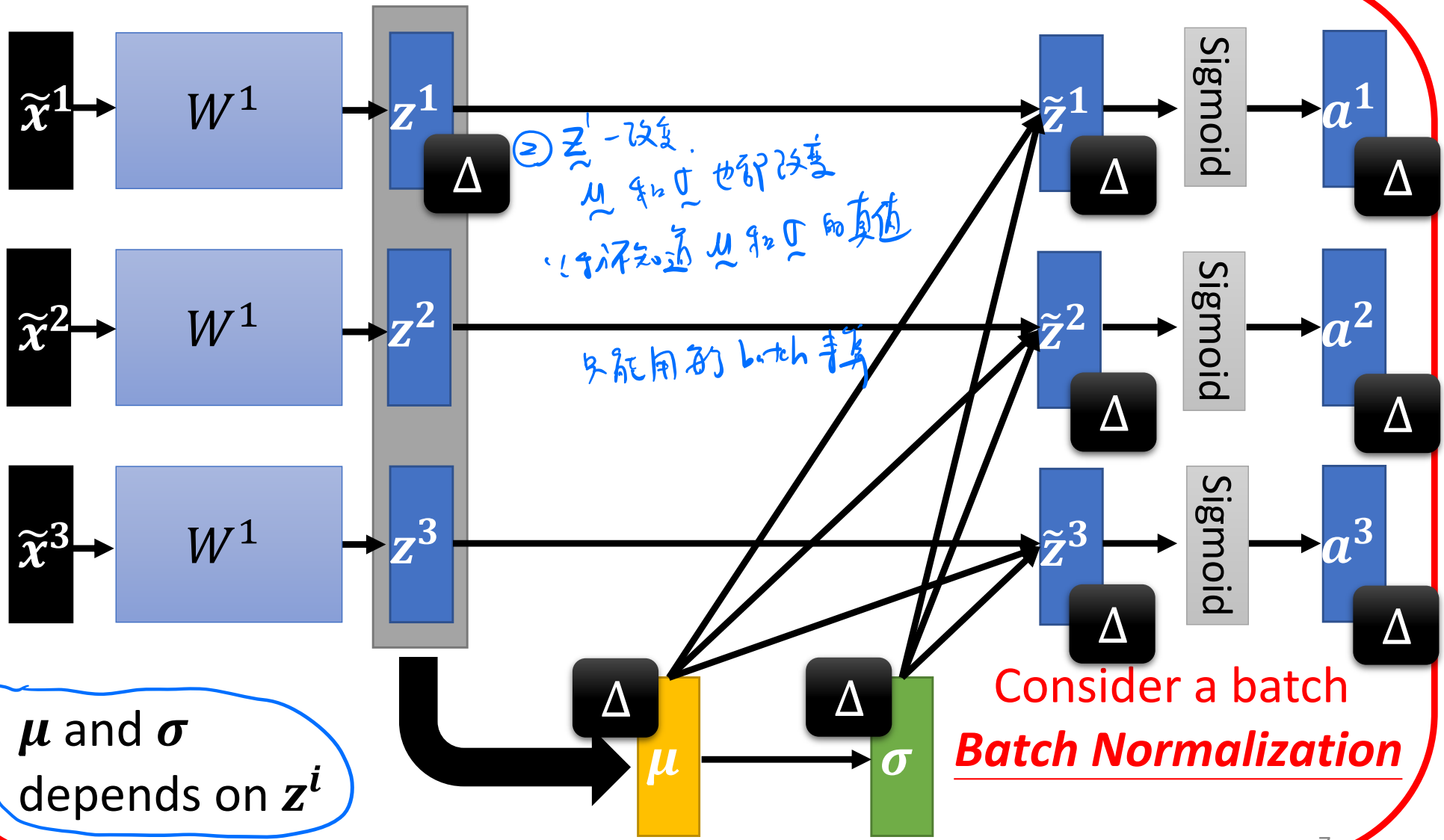
$$\tilde{\mu} = \frac{1}{3}\sum_{i=1}^{3} z^i$$

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

Considering Deep Learning

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

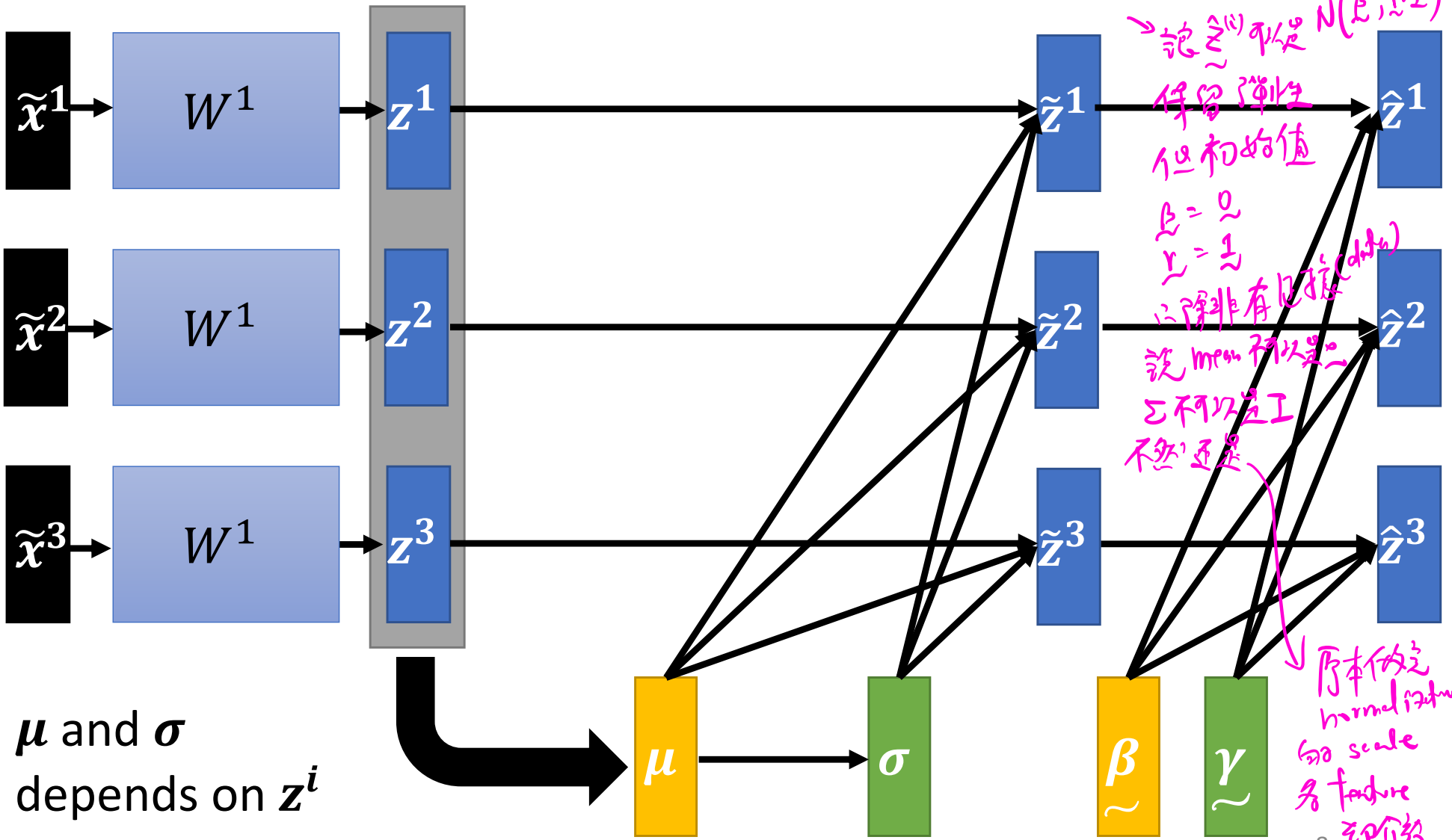
This is a large network!



Batch normalization

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

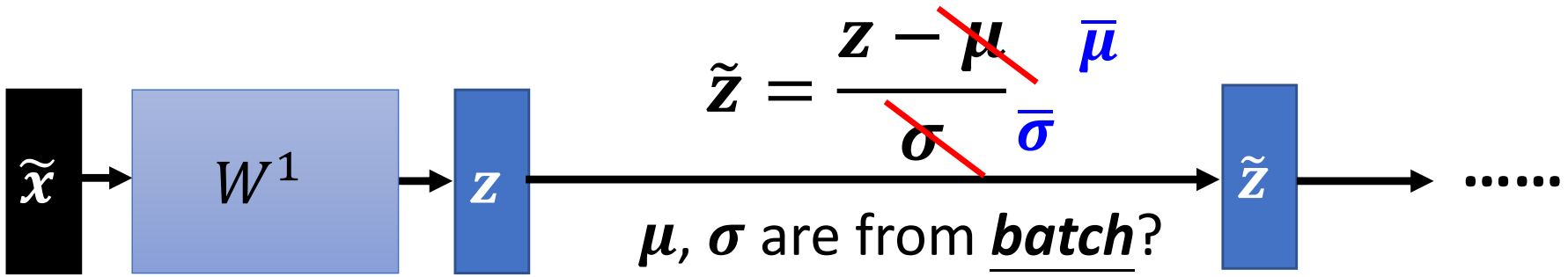
$$\hat{z}^i = \gamma \odot \tilde{z}^i + \beta$$



μ and σ depends on z^i

Batch normalization – Testing

上线时，$\bar{\mu}$ 和 $\bar{\sigma}$ 是用 training 时得到的 $\bar{\mu}$ 和 $\bar{\sigma}$



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

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

$$\boldsymbol{\mu}^1 \quad \boldsymbol{\mu}^2 \quad \boldsymbol{\mu}^3 \quad \ldots\ldots \quad \boldsymbol{\mu}^t$$

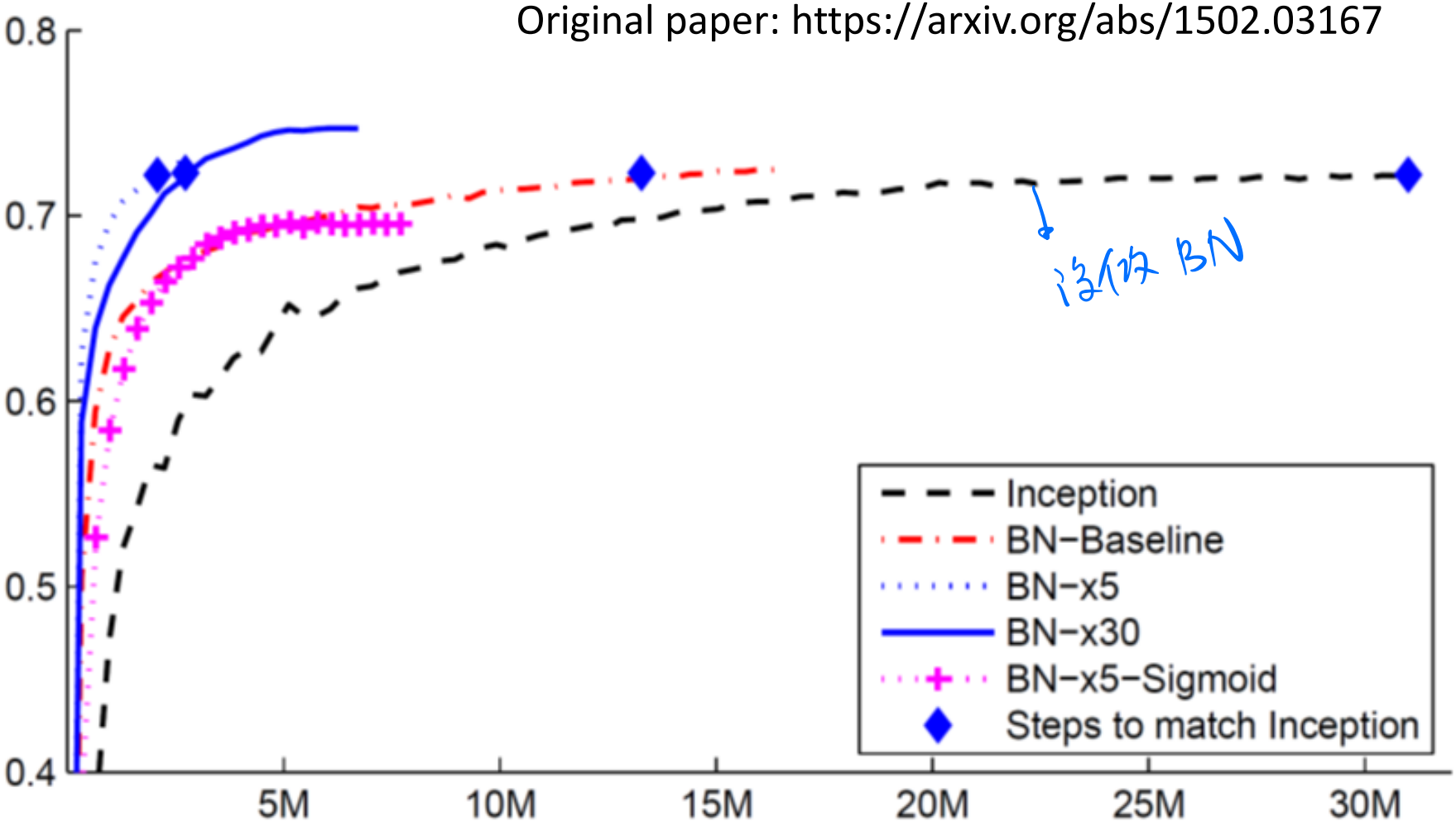
$$\bar{\boldsymbol{\mu}} \leftarrow p\bar{\boldsymbol{\mu}} + (1-p)\boldsymbol{\mu}^t$$

p: a hyper-parameter

$\boldsymbol{\mu}^t$: mean of the t-th batch

Batch normalization

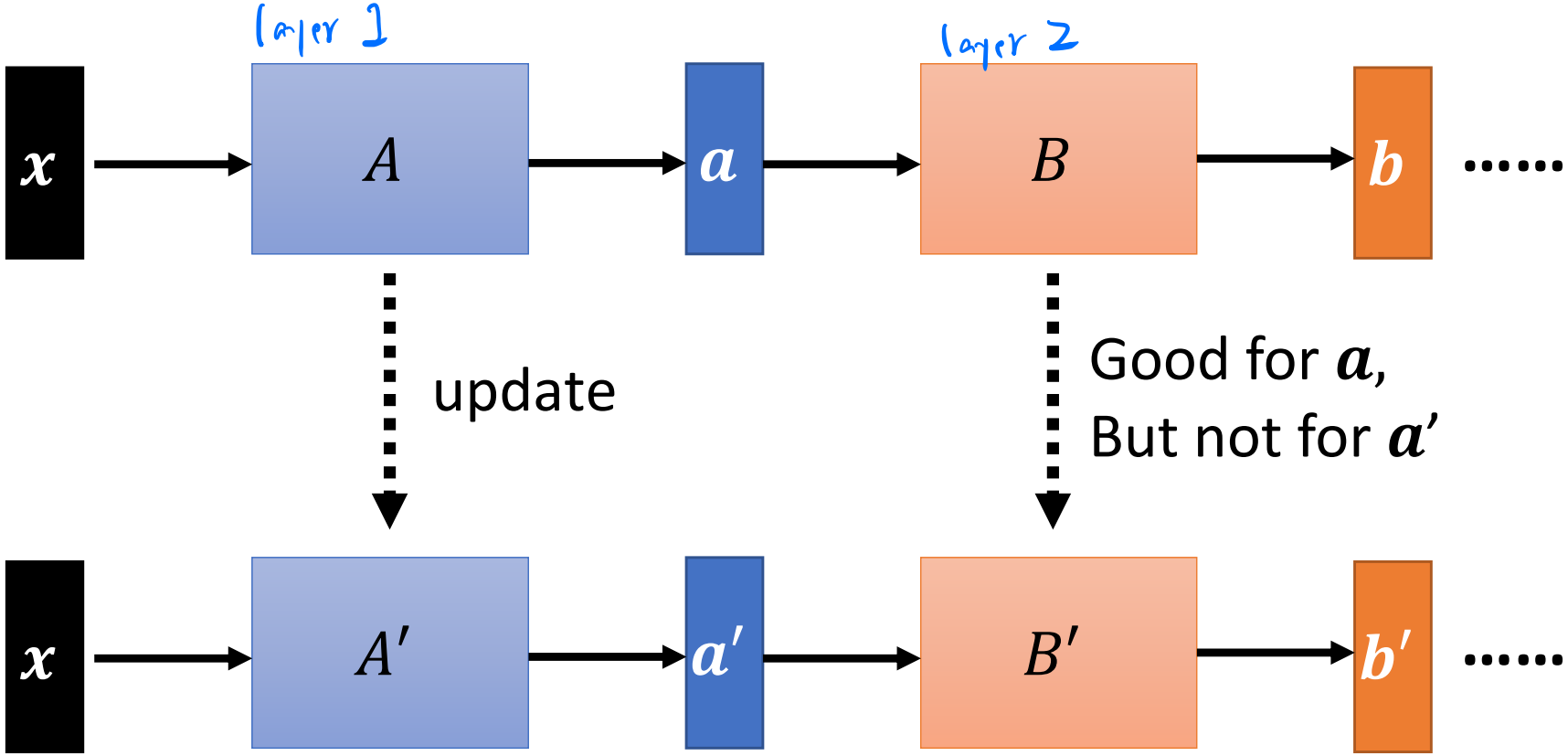
Original paper: https://arxiv.org/abs/1502.03167



Internal Covariate Shift?

How Does Batch Normalization Help Optimization?

https://arxiv.org/abs/1805.11604



Batch normalization make $\boldsymbol{a}$ and $\boldsymbol{a}'$ have similar statistics.

Experimental results do not support the above idea.

Internal Covariate Shift?

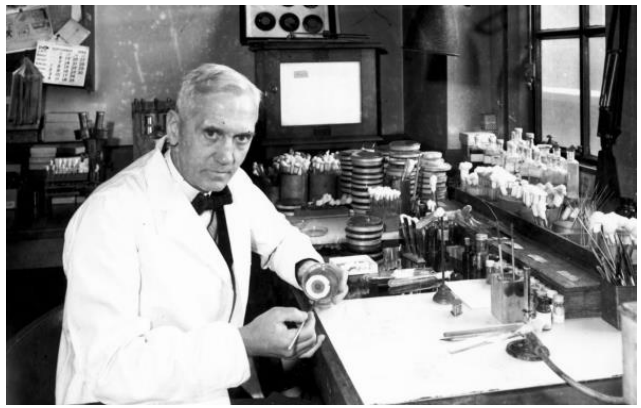
How Does Batch Normalization Help Optimization?

https://arxiv.org/abs/1805.11604

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)



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

还有一张 normalization 的方法
可以参考
(反正都是 training tips
有人试过这样改
normalization，有效
他就发出来)


Q&A