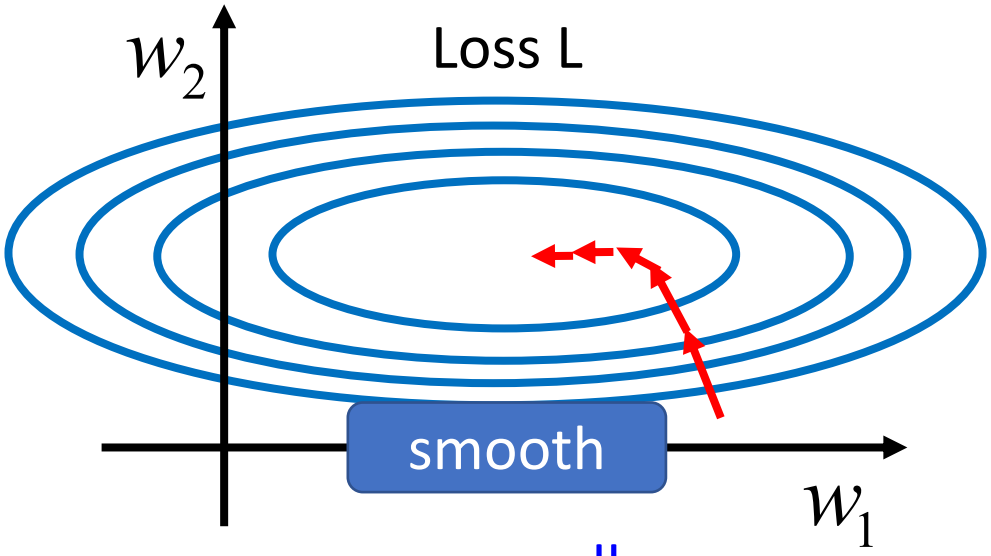
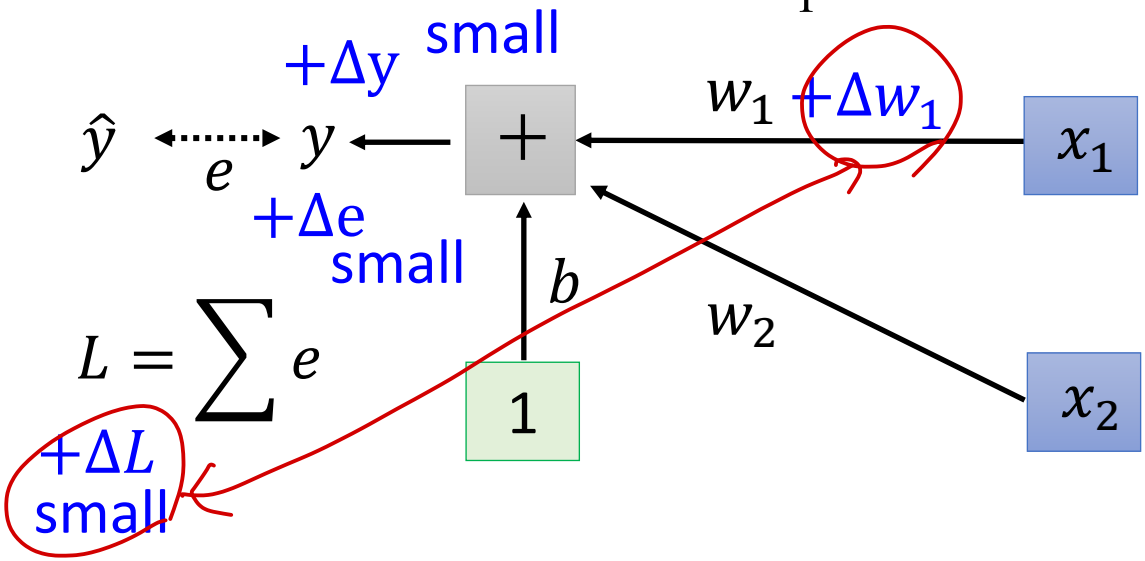
Quick Introduction of Batch Normalization

Hung-yi Lee 李宏毅

Changing Landscape



adaptive learning rate.



影响小的情况

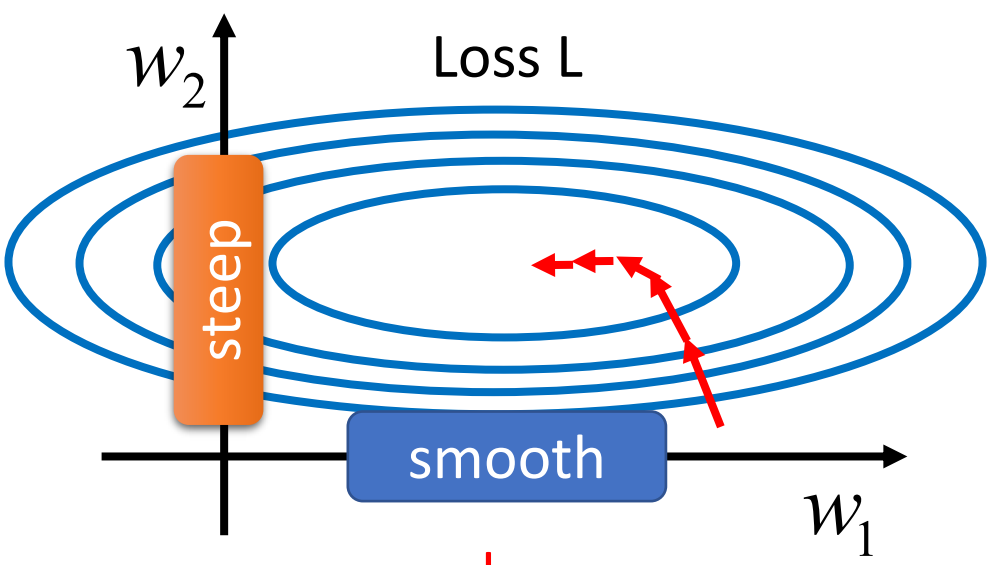
① x_1很小

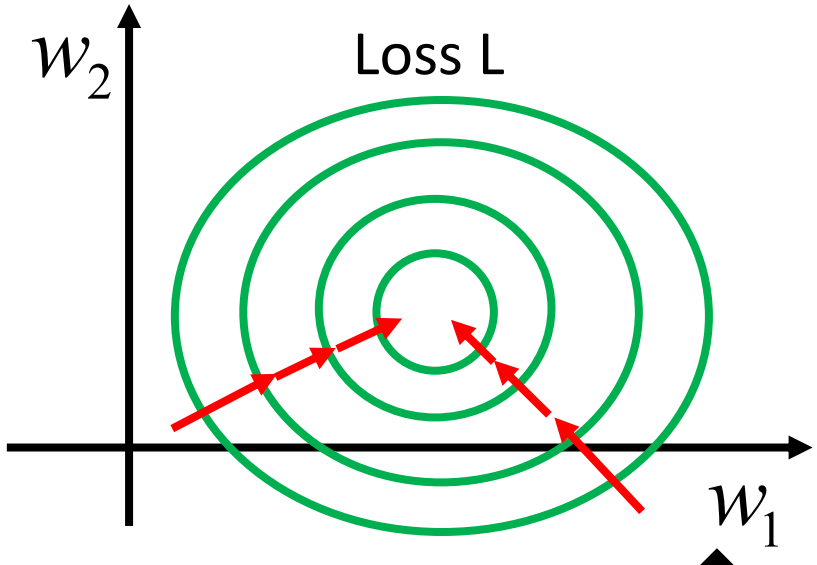
1, 2 .....

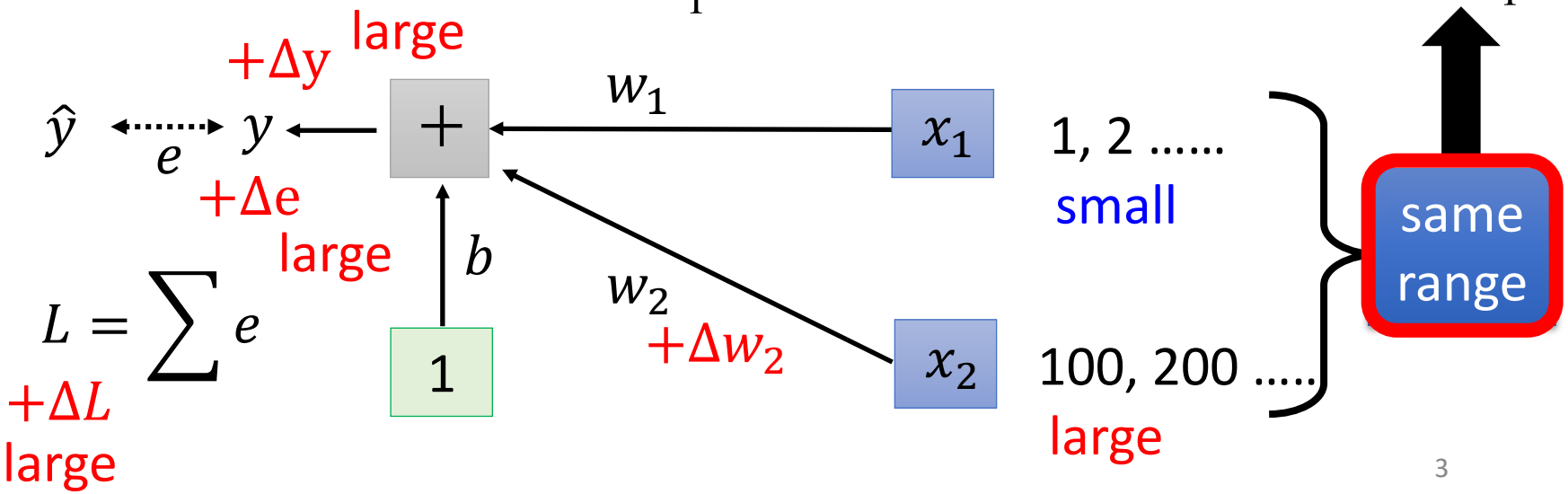
small

② x_2很大

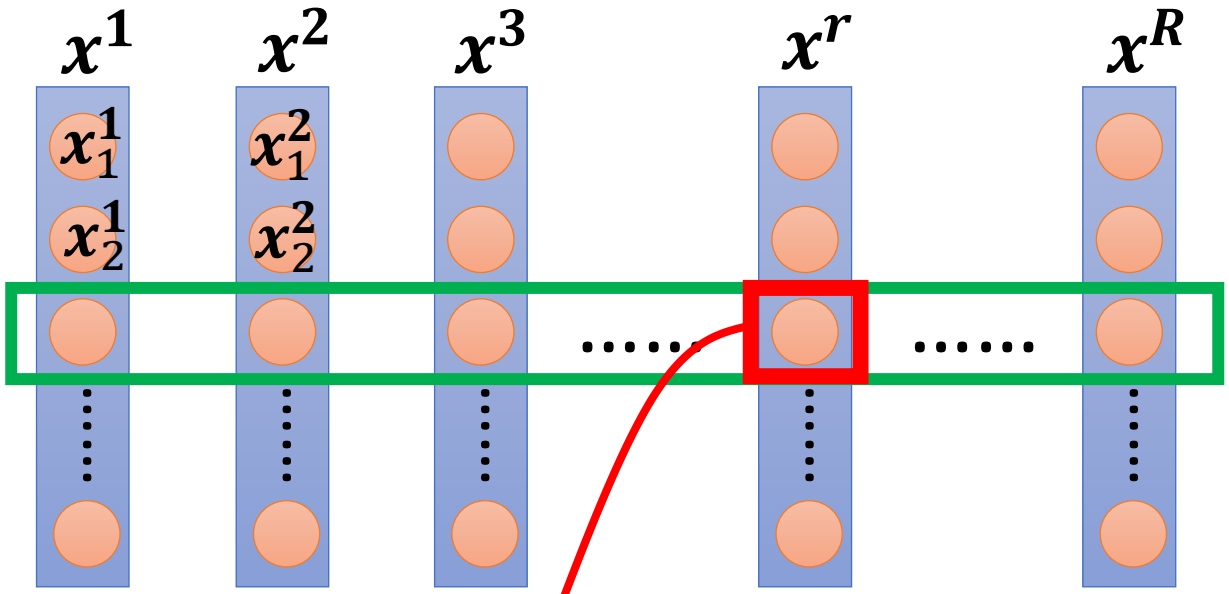
Changing Landscape







Feature Normalization



For each dimension i:

mean: m_i

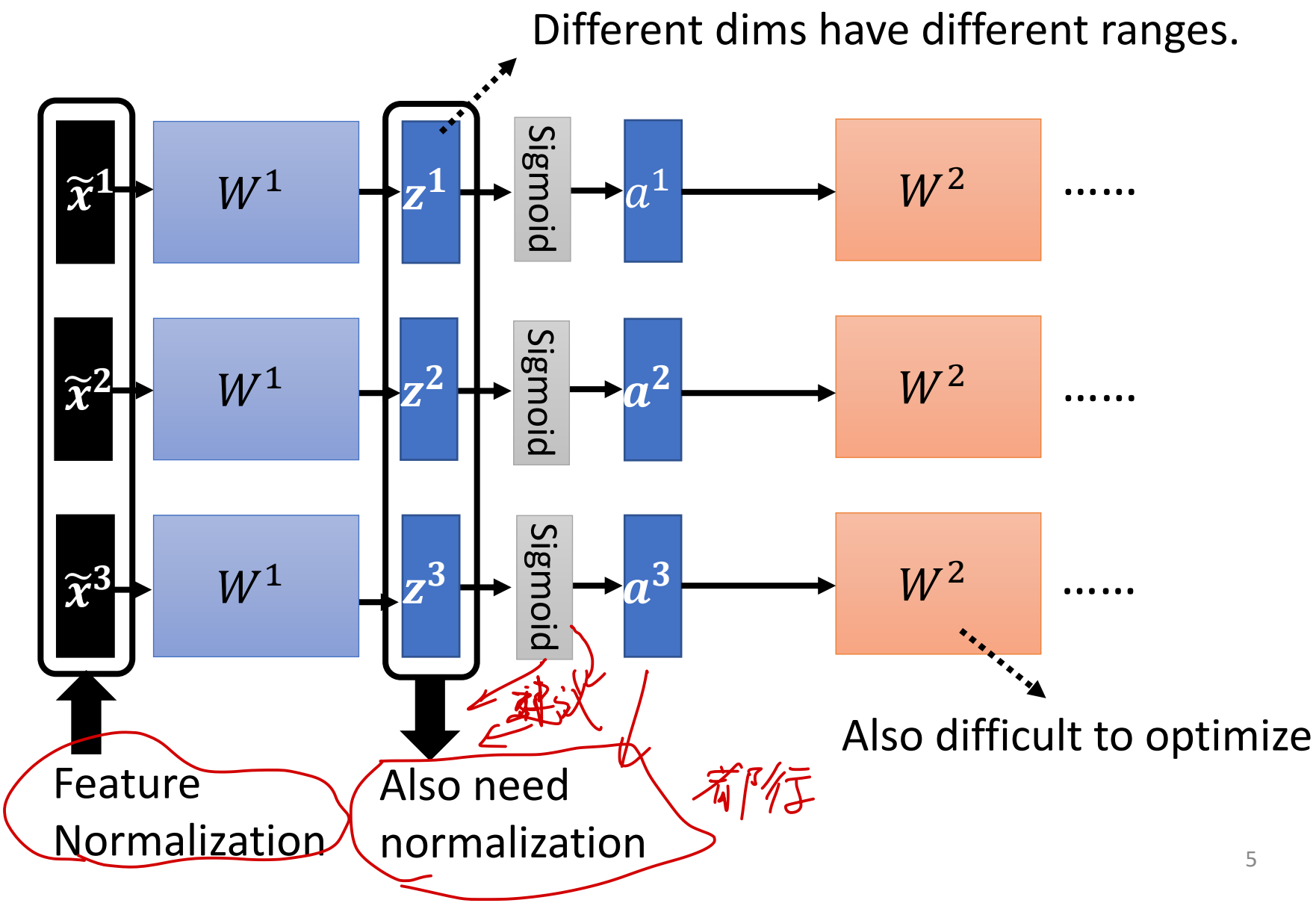
standard deviation: σ_i

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

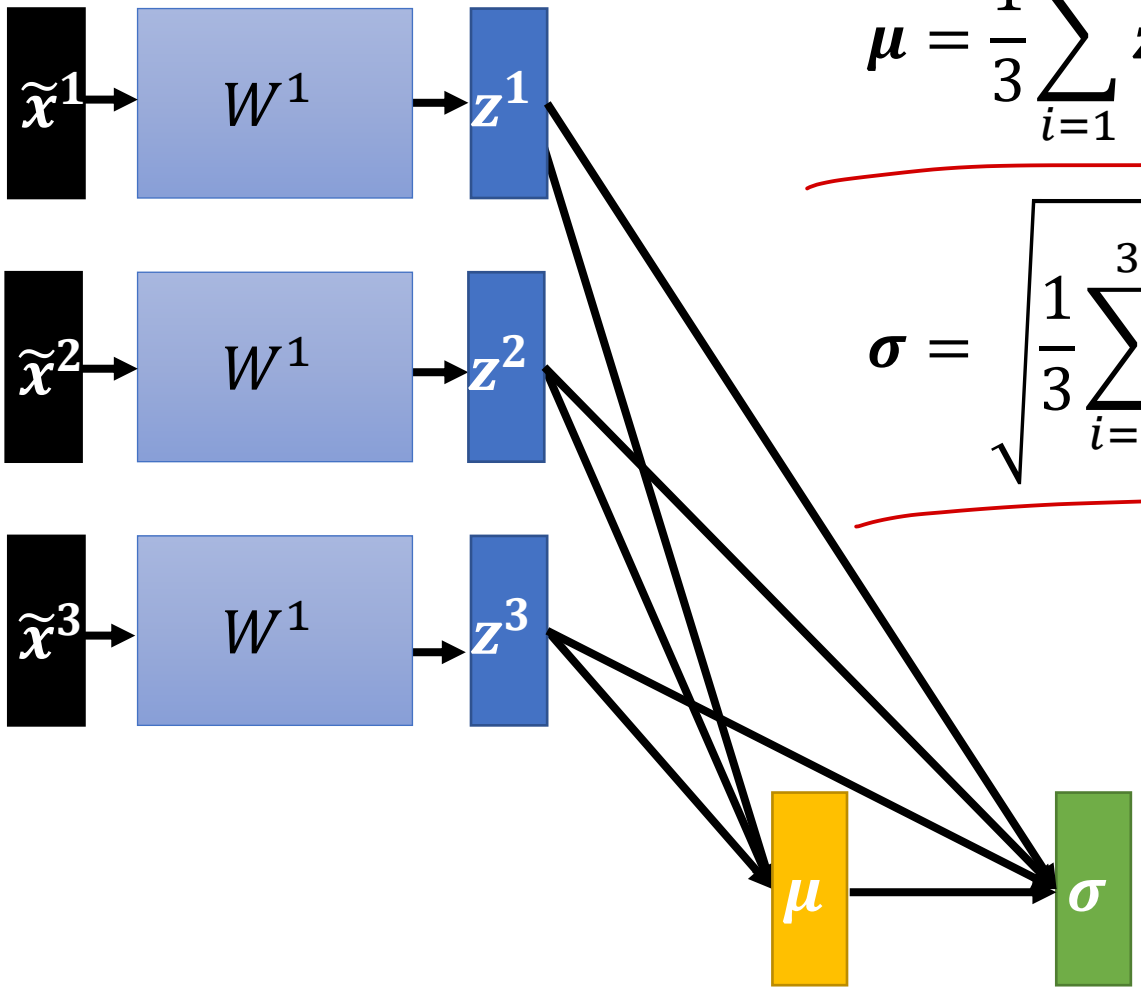
The means of all dims are 0, and the variances are all 1

In general, feature normalization makes gradient descent converge faster.

Considering Deep Learning



Considering Deep Learning



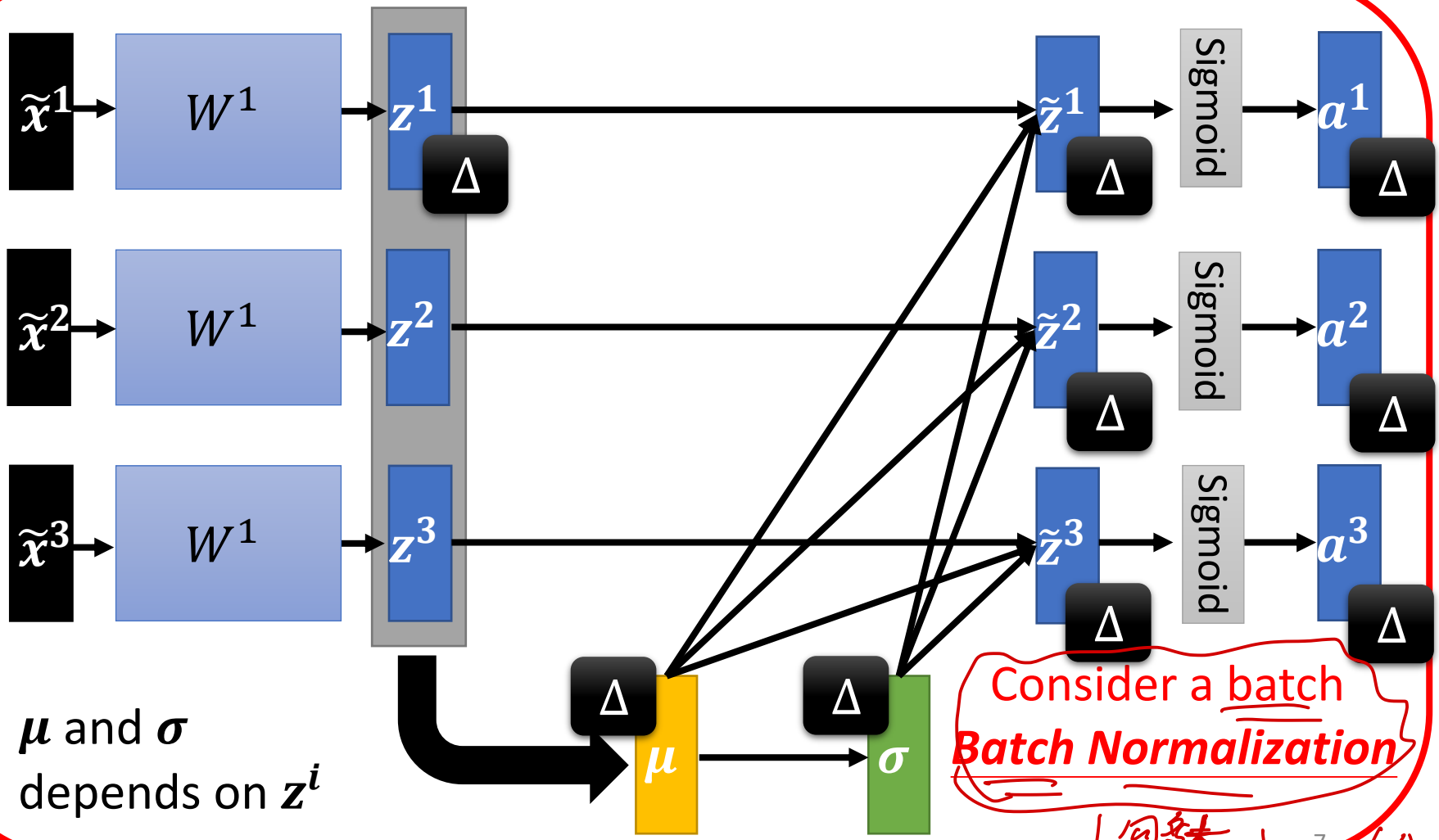
$$\mu = \frac{1}{3}\sum_{i=1}^{3} z^i$$

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

Considering Deep Learning

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

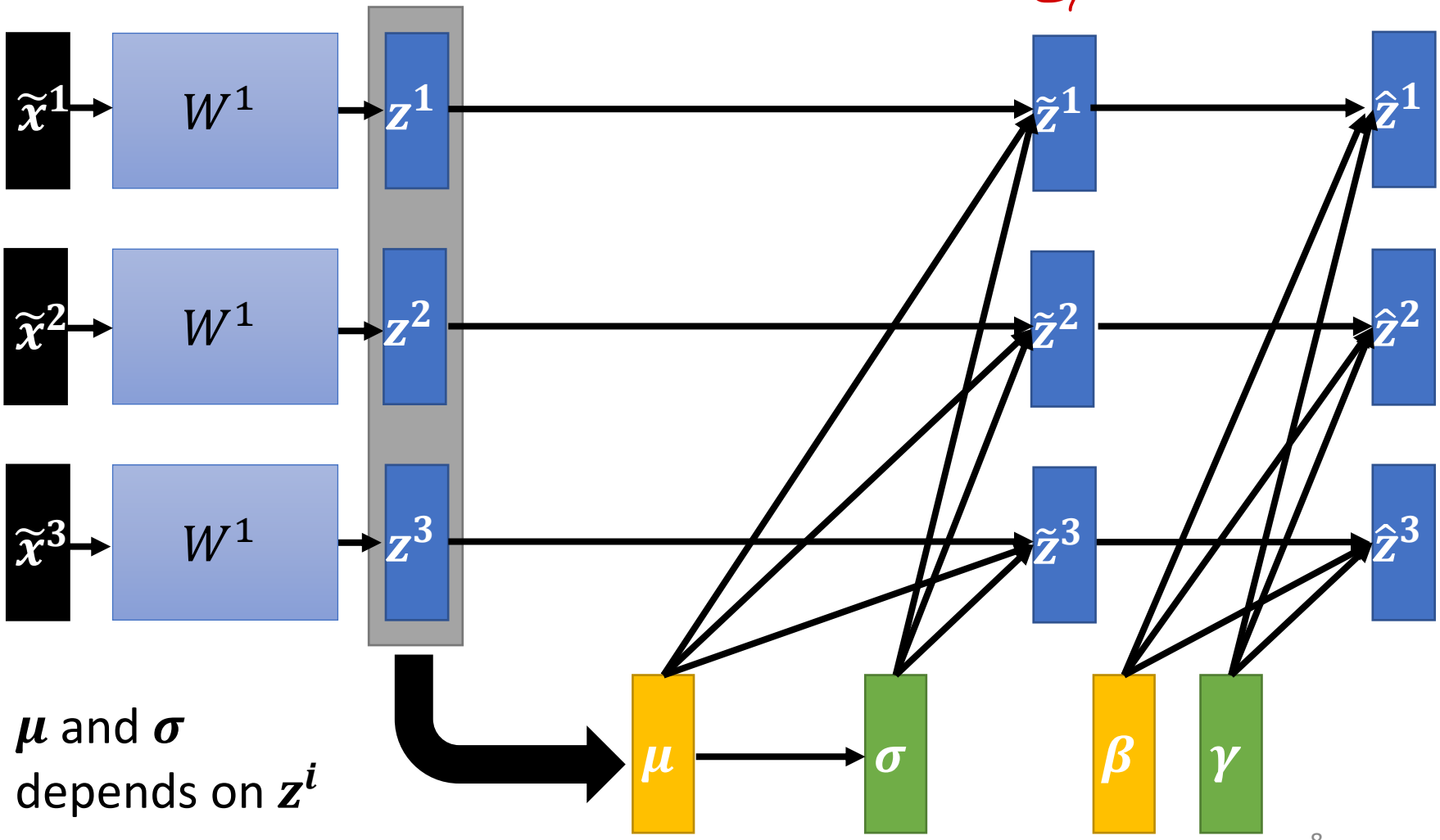
This is a large network!



Batch normalization

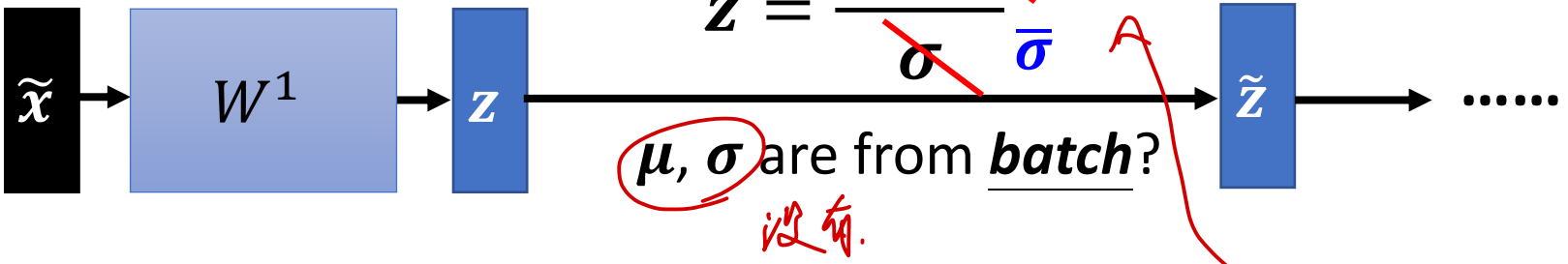
$$\tilde{z}^i = \frac{z^i - \mu}{\sigma}$$

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



μ and σ depends on z^i

Batch normalization – Testing



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

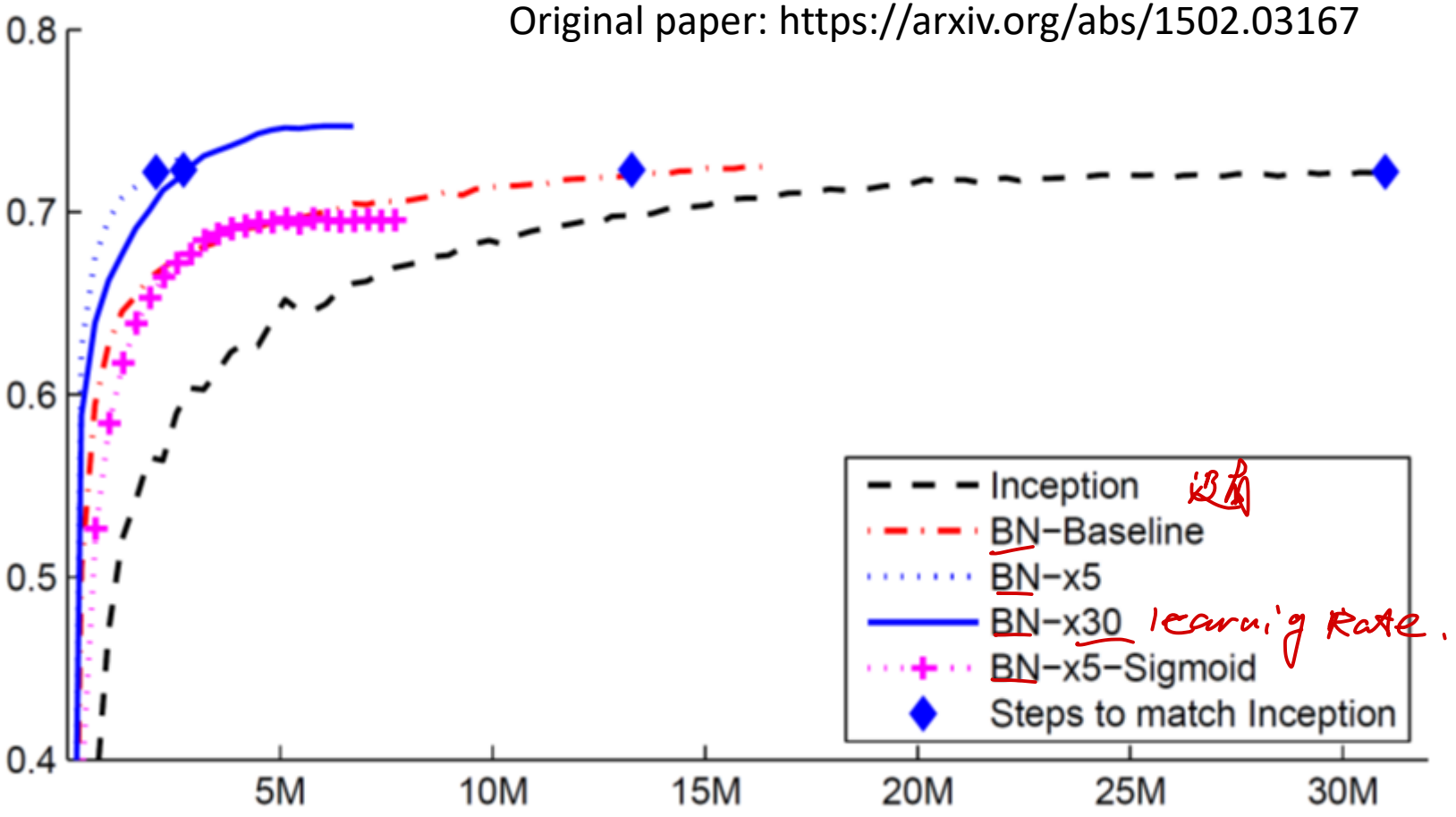
Computing the moving average of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ of the batches during training.

$\boldsymbol{\mu^1}$ $\boldsymbol{\mu^2}$ $\boldsymbol{\mu^3}$ ..... $\boldsymbol{\mu^t}$

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

Batch normalization

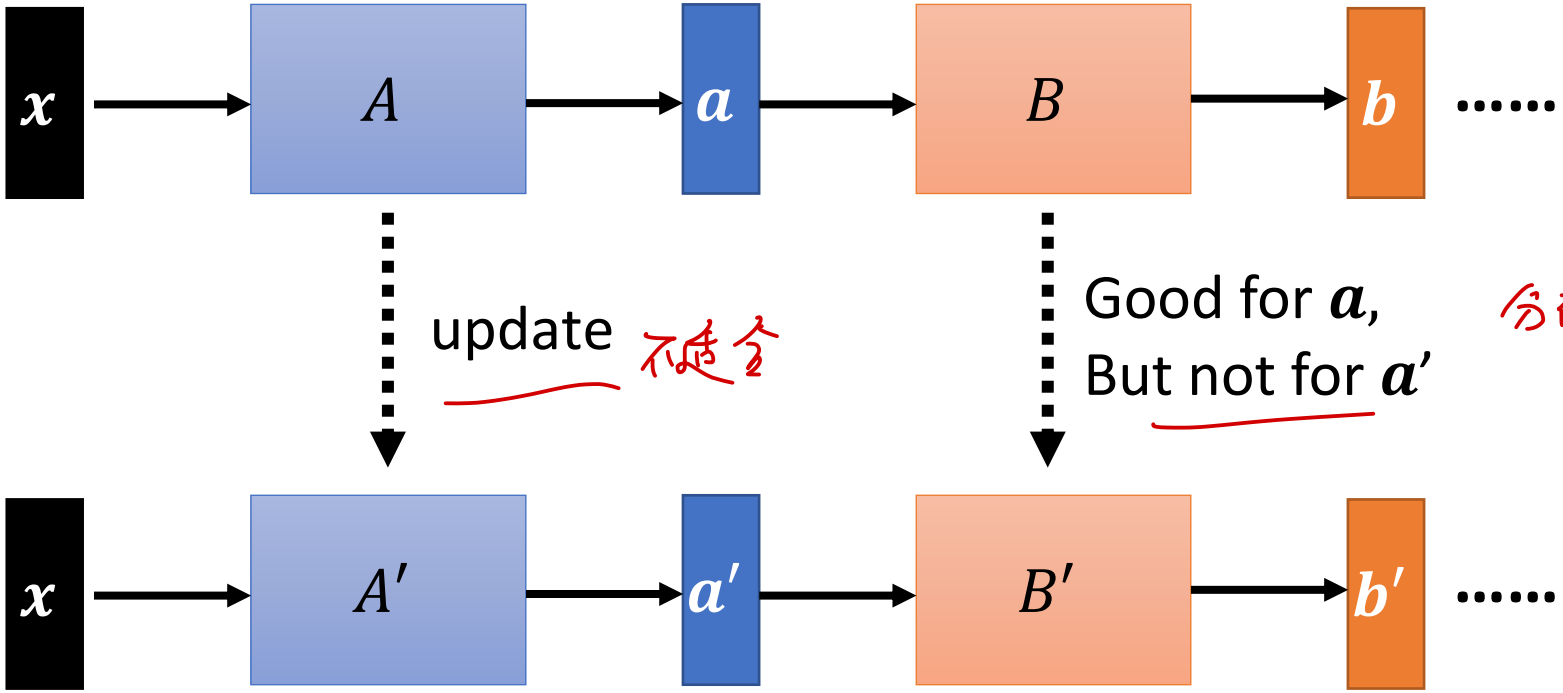
Original paper: https://arxiv.org/abs/1502.03167



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

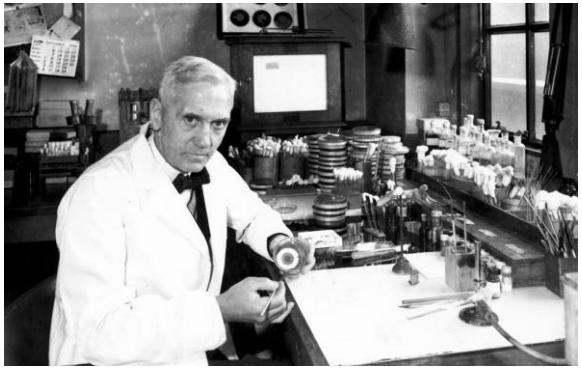
Experimental results (and theoretically analysis) support batch normalization change the landscape of error surface.

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.

意料之外的发现

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


Q&A