

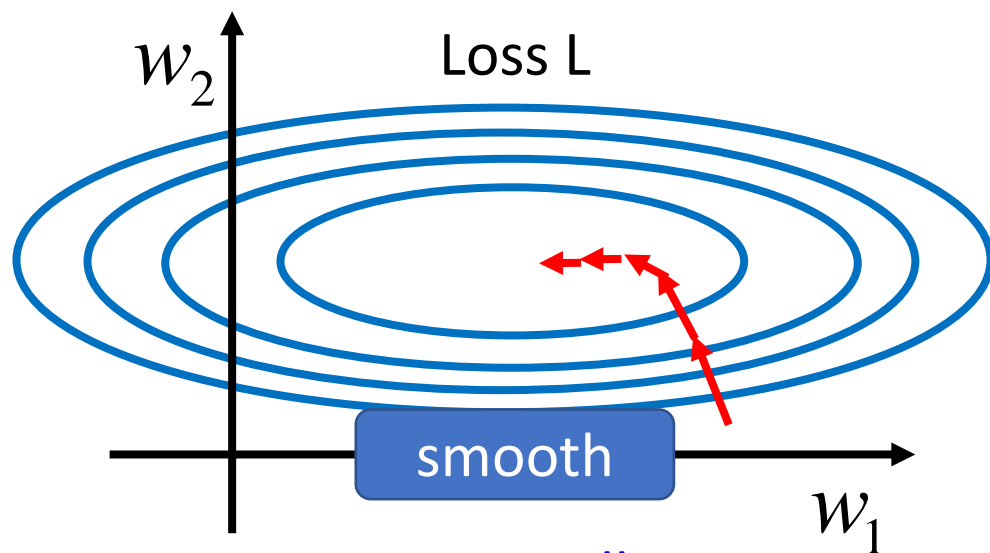
# Quick Introduction of Batch Normalization

批归一化

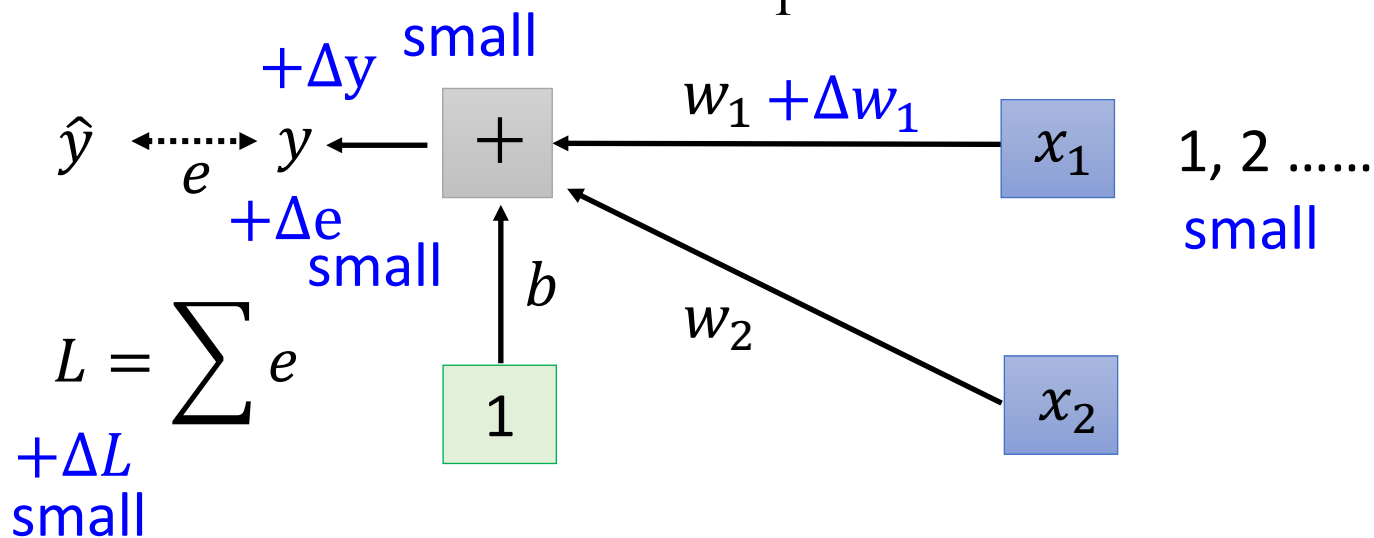
Hung-yi Lee 李宏毅

改变error surface的面貌

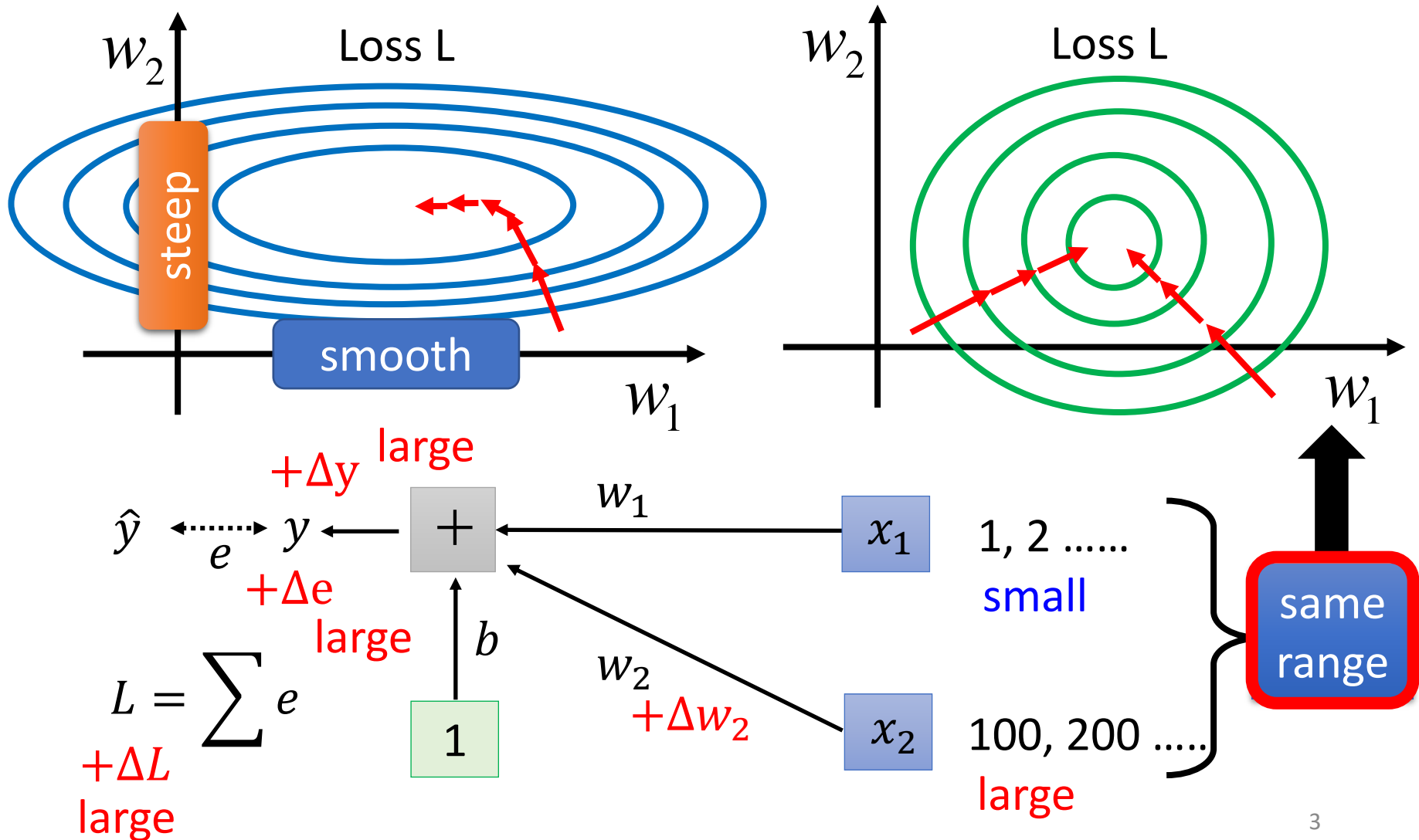
# Changing Landscape



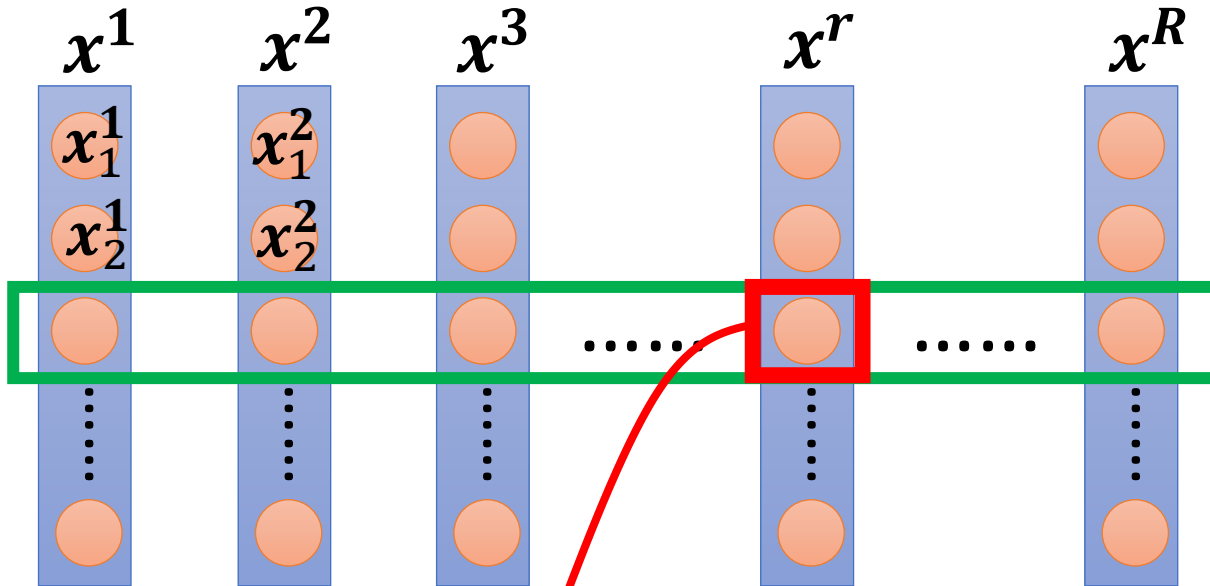
$w_1$ 的改变会导致 $L$ 的改变。  
什么情况下这个影响很小呢？  
答： $x_1$ 的值很小时



# Changing Landscape



# Feature Normalization = feature scaling



For each dimension  $i$ :  
mean:  $m_i$   
standard deviation:  $\sigma_i$

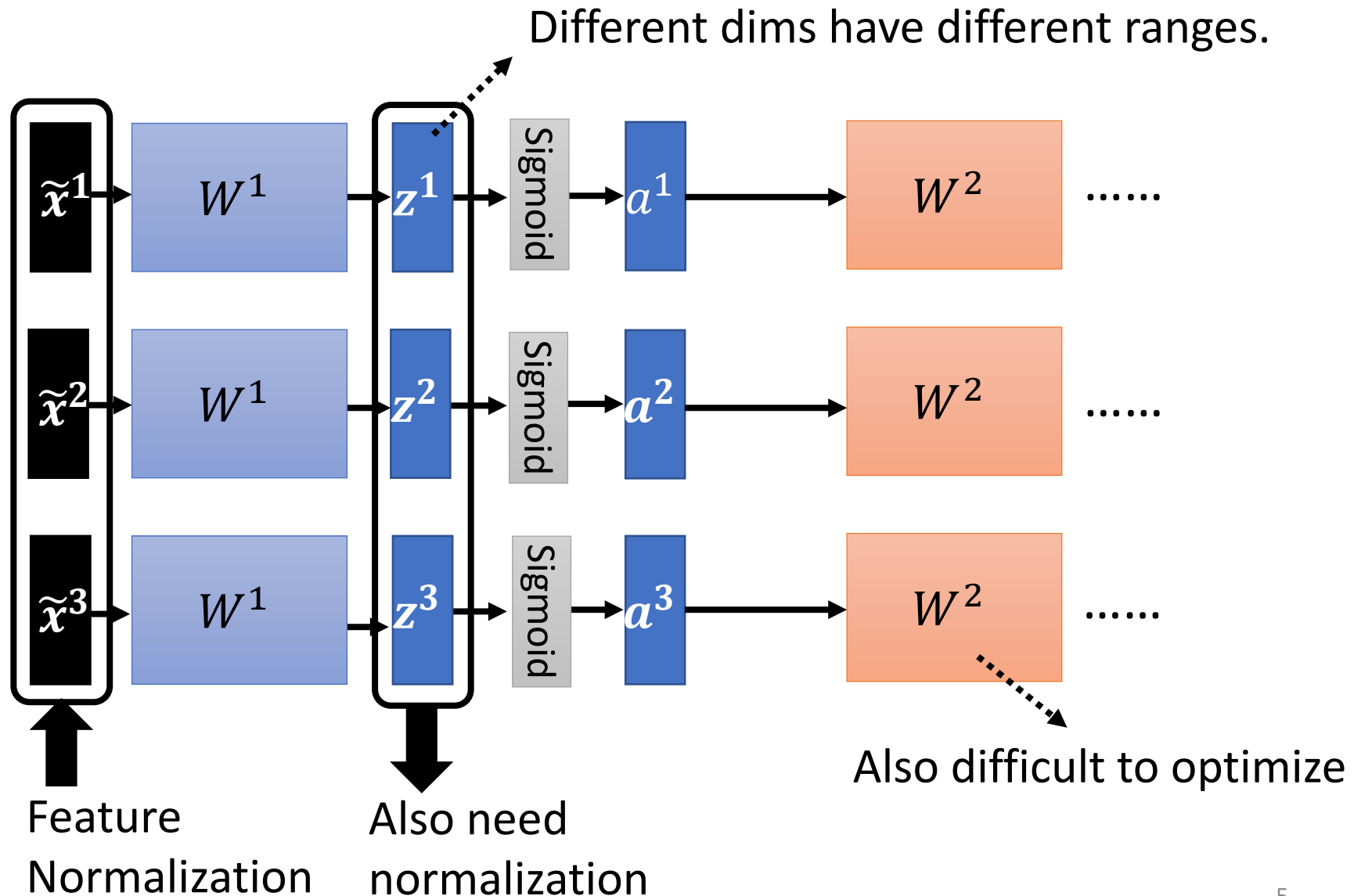
standardization 标准化

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

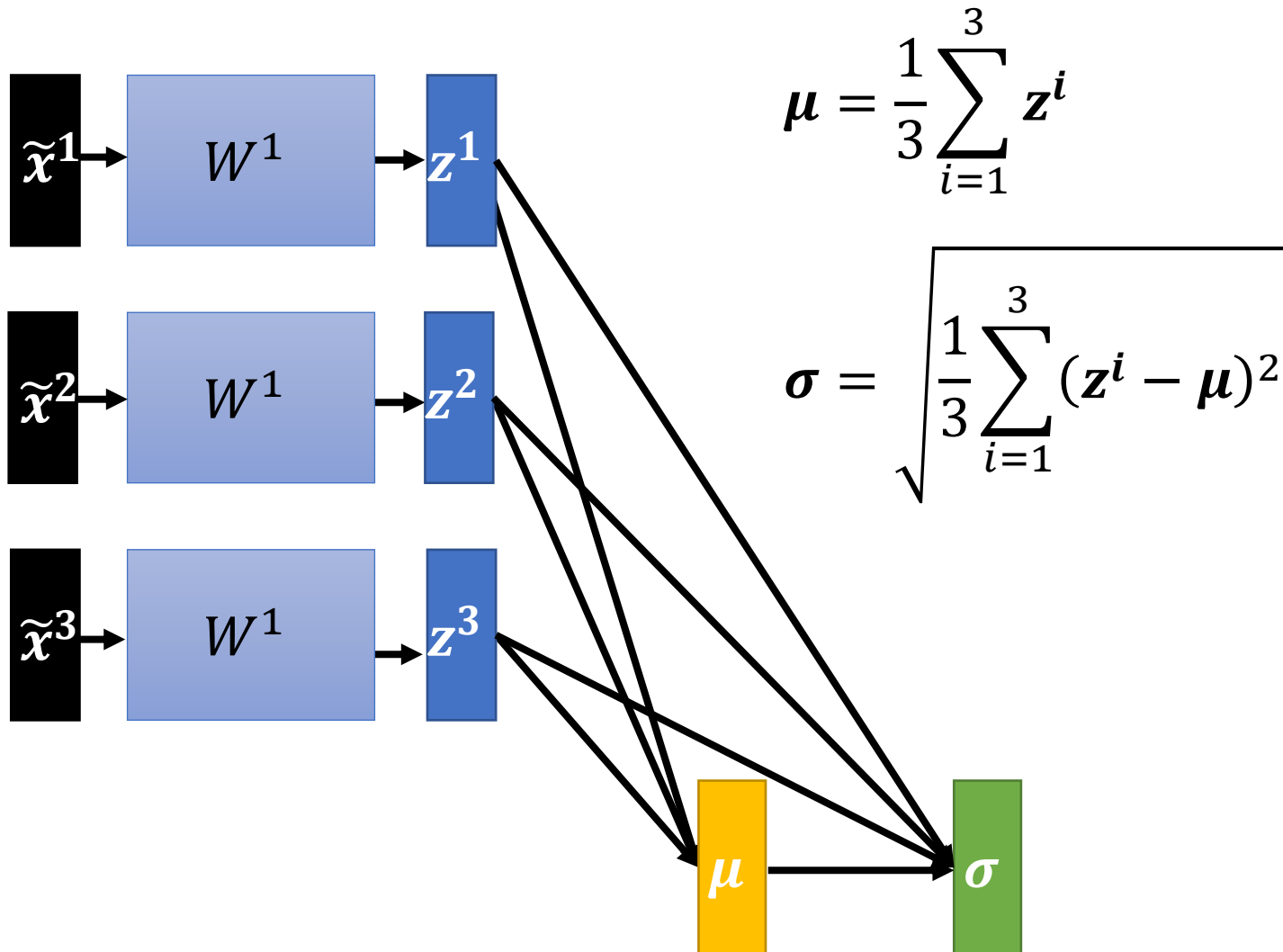
The means of all dims are 0,  
and the variances are all 1

In general, feature normalization makes gradient descent converge faster.

# Considering Deep Learning



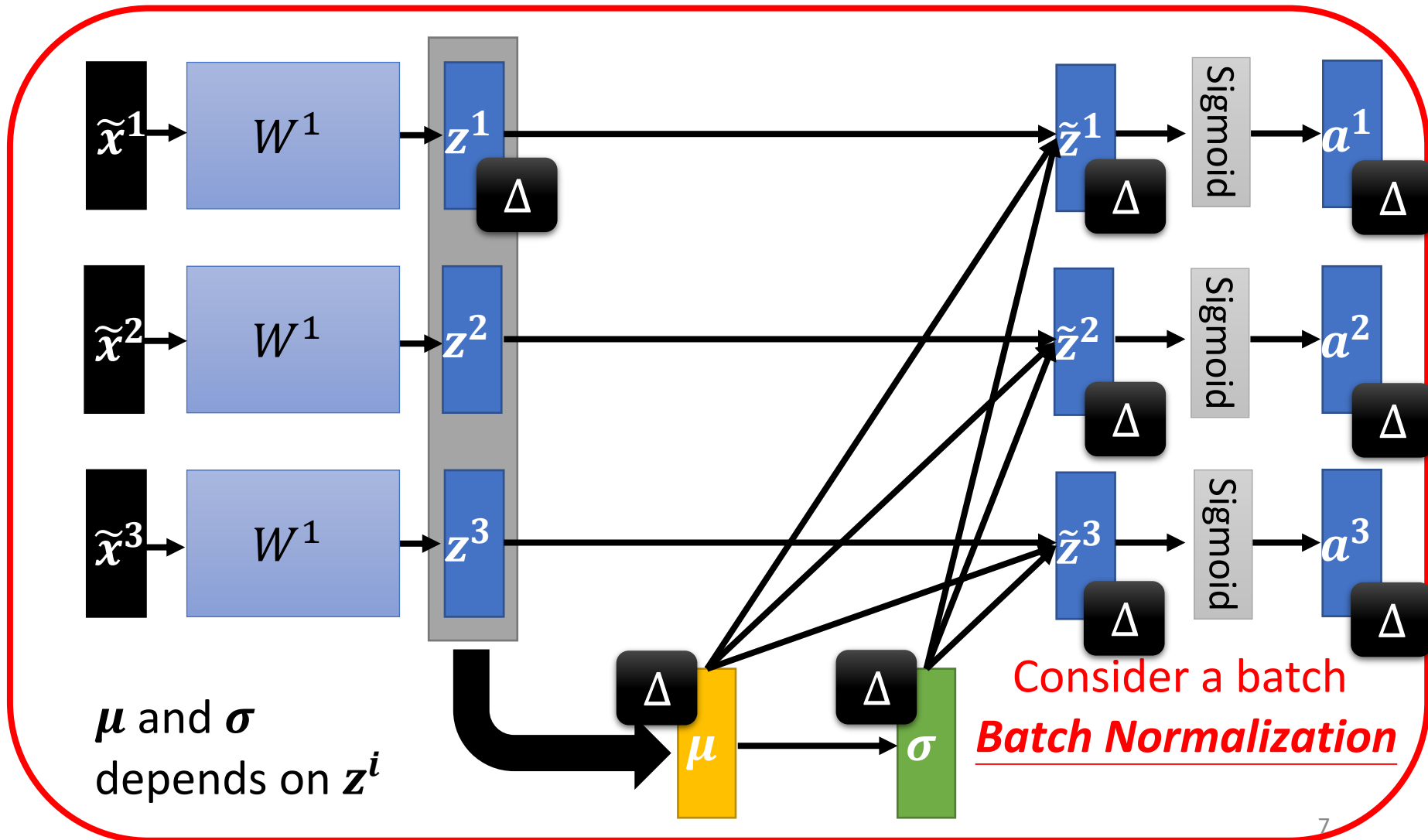
# Considering Deep Learning



# Considering Deep Learning

This is a large network!

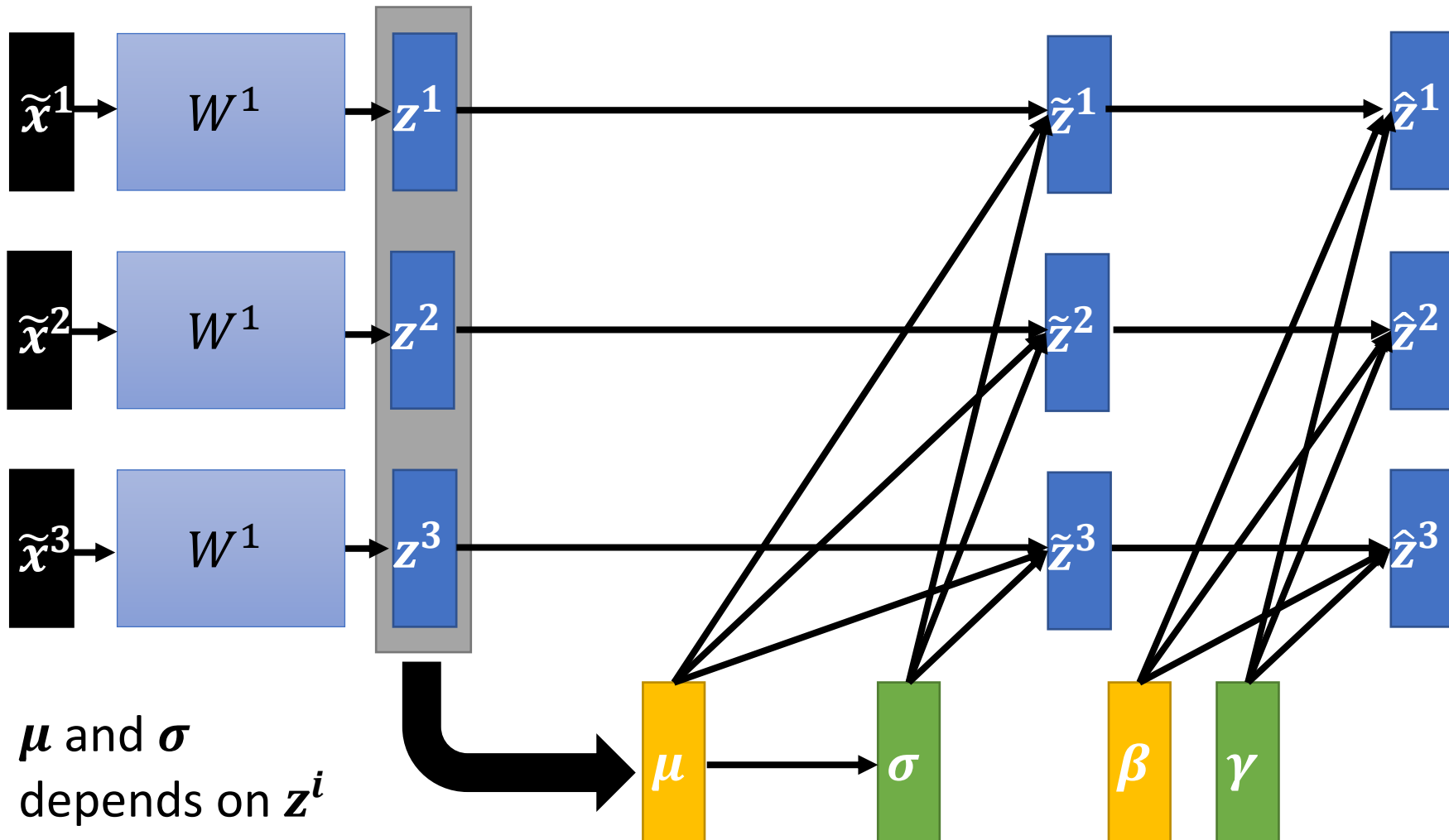
$$\tilde{z}^i = \frac{z^i - \mu}{\sigma} \text{ element-wise}$$



# Batch normalization

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

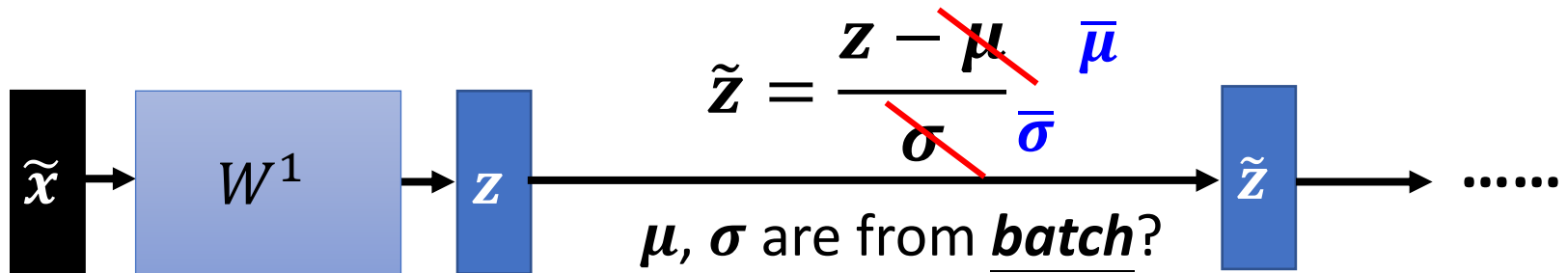
$$\hat{\mathbf{z}}^i = \gamma \odot \tilde{\mathbf{z}}^i + \beta$$





inference

# Batch normalization – Testing



We do not always have batch at testing stage.

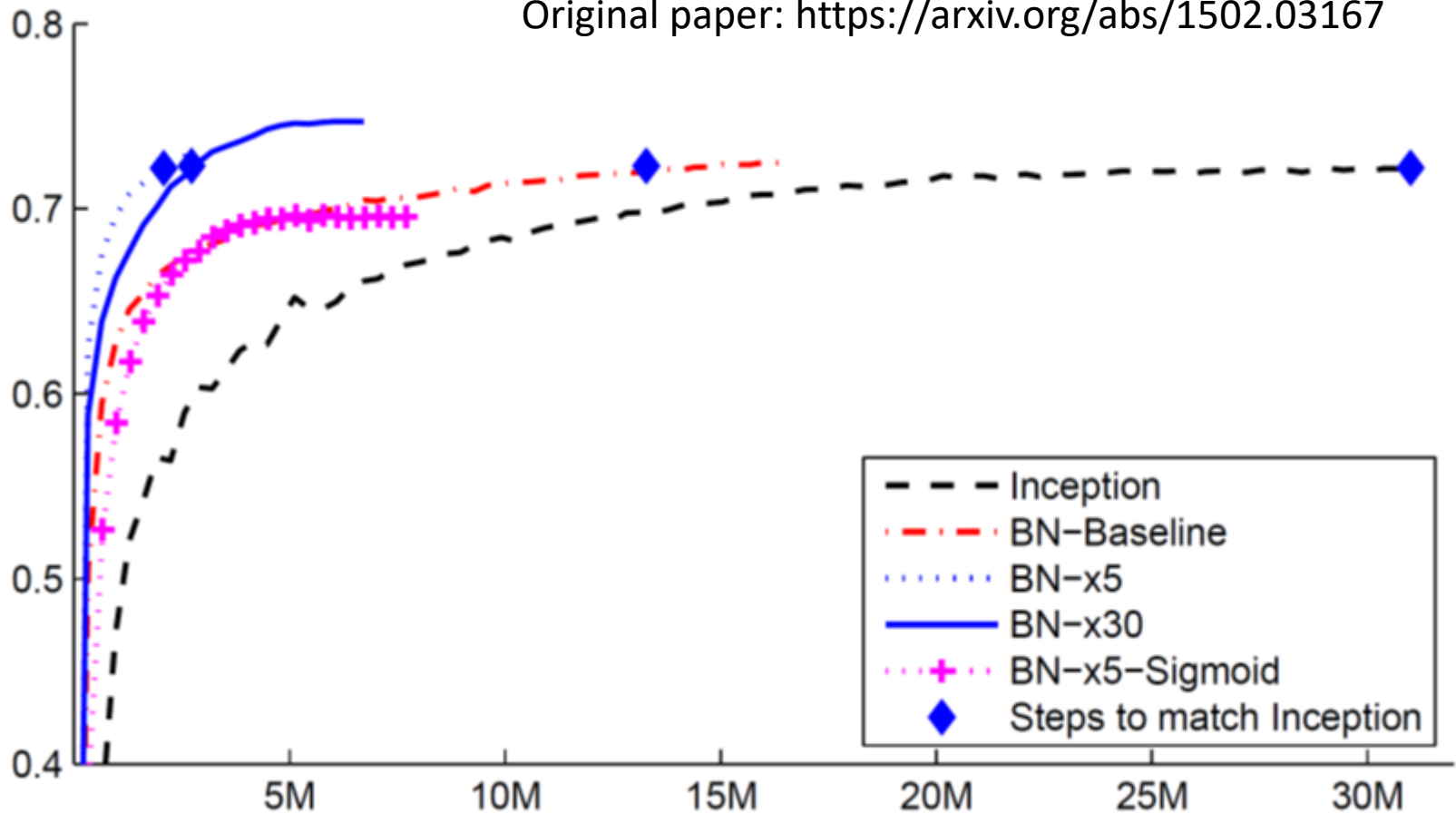
Computing the moving average of  $\mu$  and  $\sigma$  of the batches during training.

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

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

# Batch normalization CNN上的应用

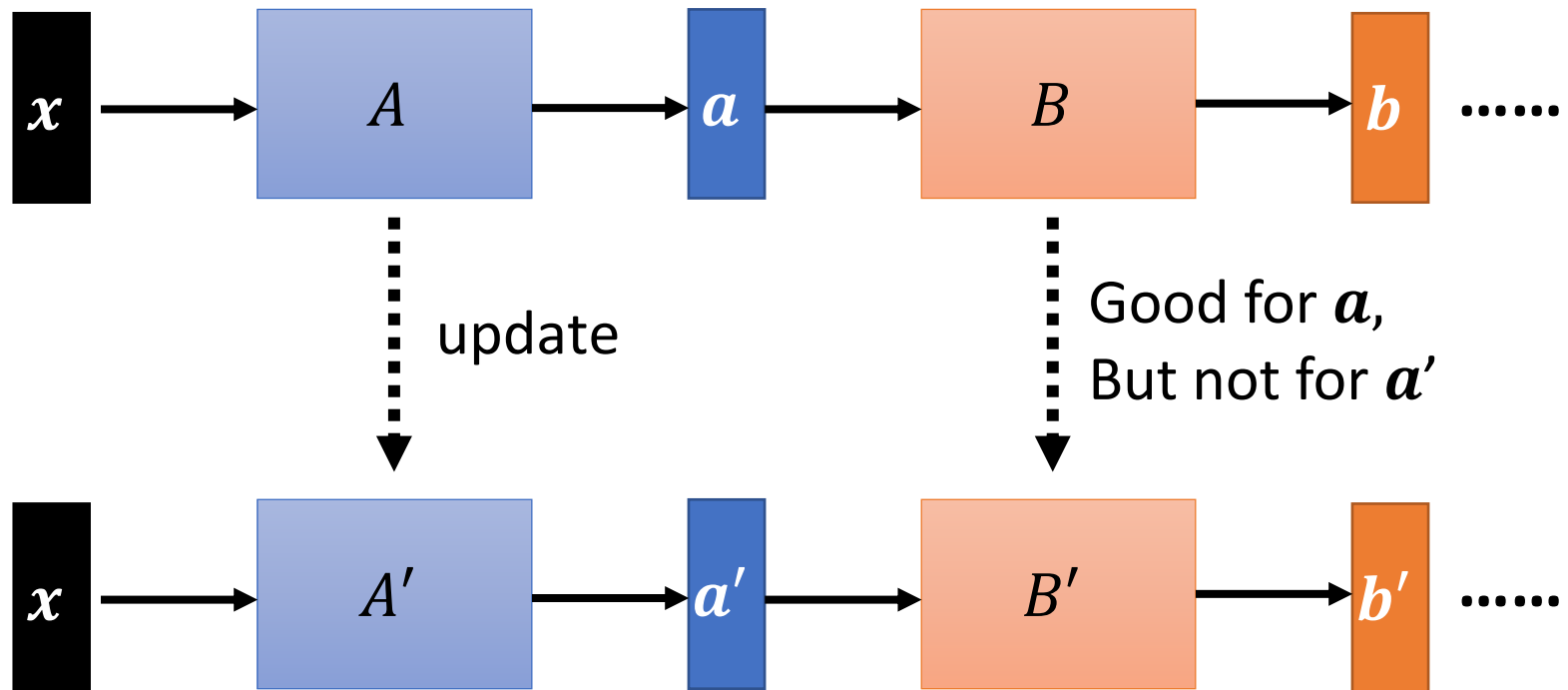
Original paper: <https://arxiv.org/abs/1502.03167>



# Internal Covariate Shift?

How Does Batch Normalization Help Optimization?

<https://arxiv.org/abs/1805.11604>



Batch normalization make  $a$  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>

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 ..... 有很多种normalization的方法

- 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>

