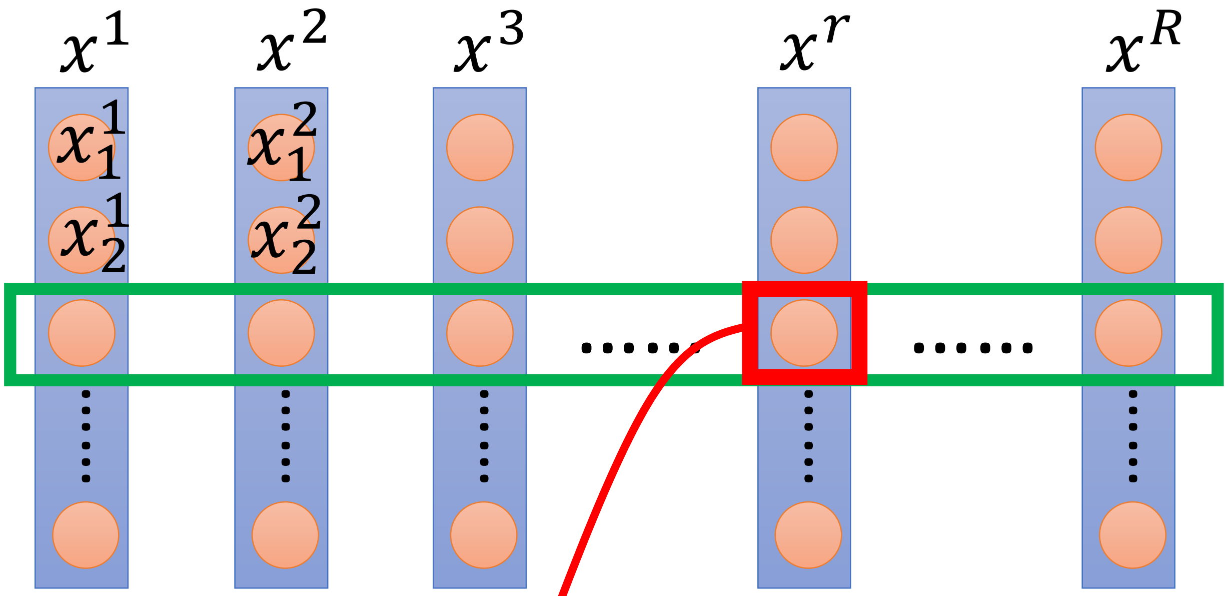
Batch Normalization

Feature Scaling



For each dimension i:

mean: m_i

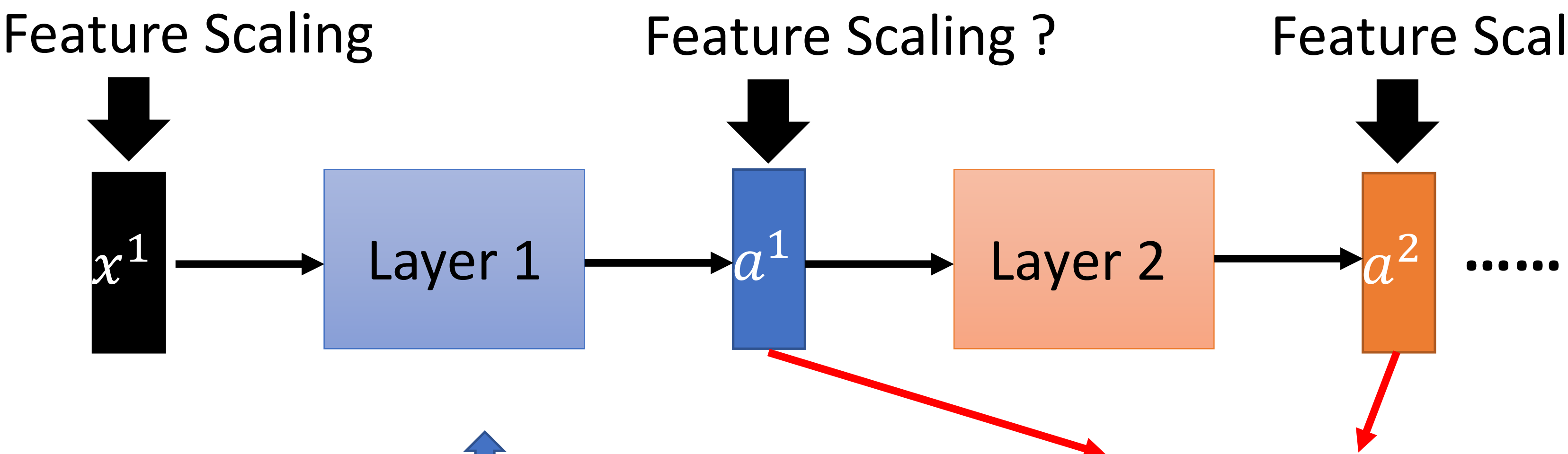
standard deviation: σ_i

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

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

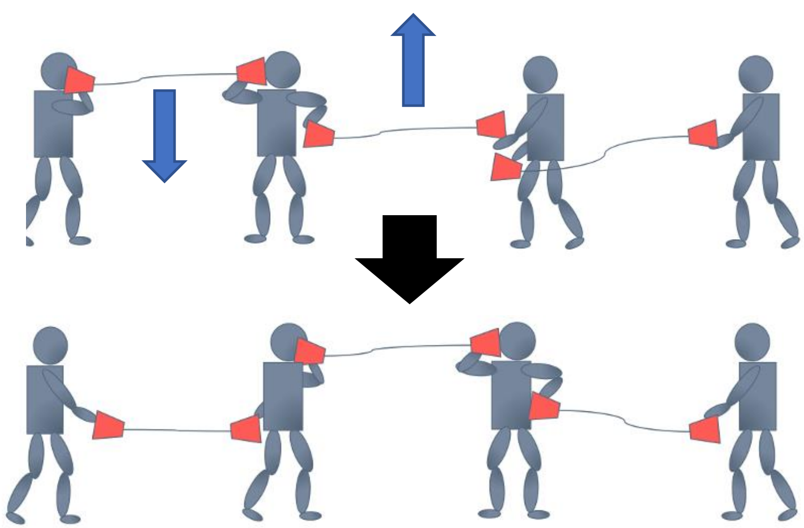
In general, gradient descent converges much faster with feature scaling than without it.

How about Hidden Layer?



Difficulty: their statistics change during the training …

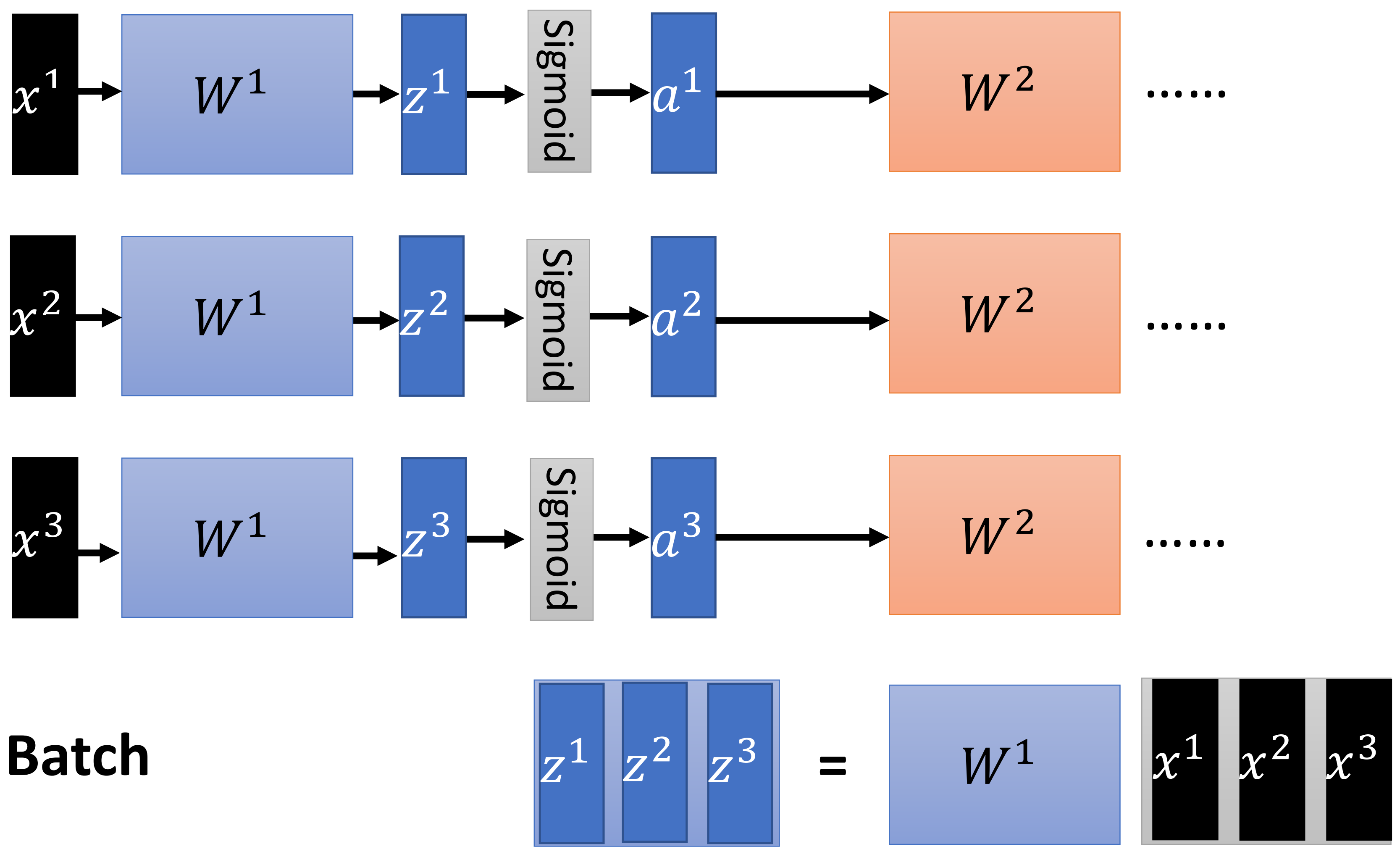
Batch normalization



Internal Covariate Shift

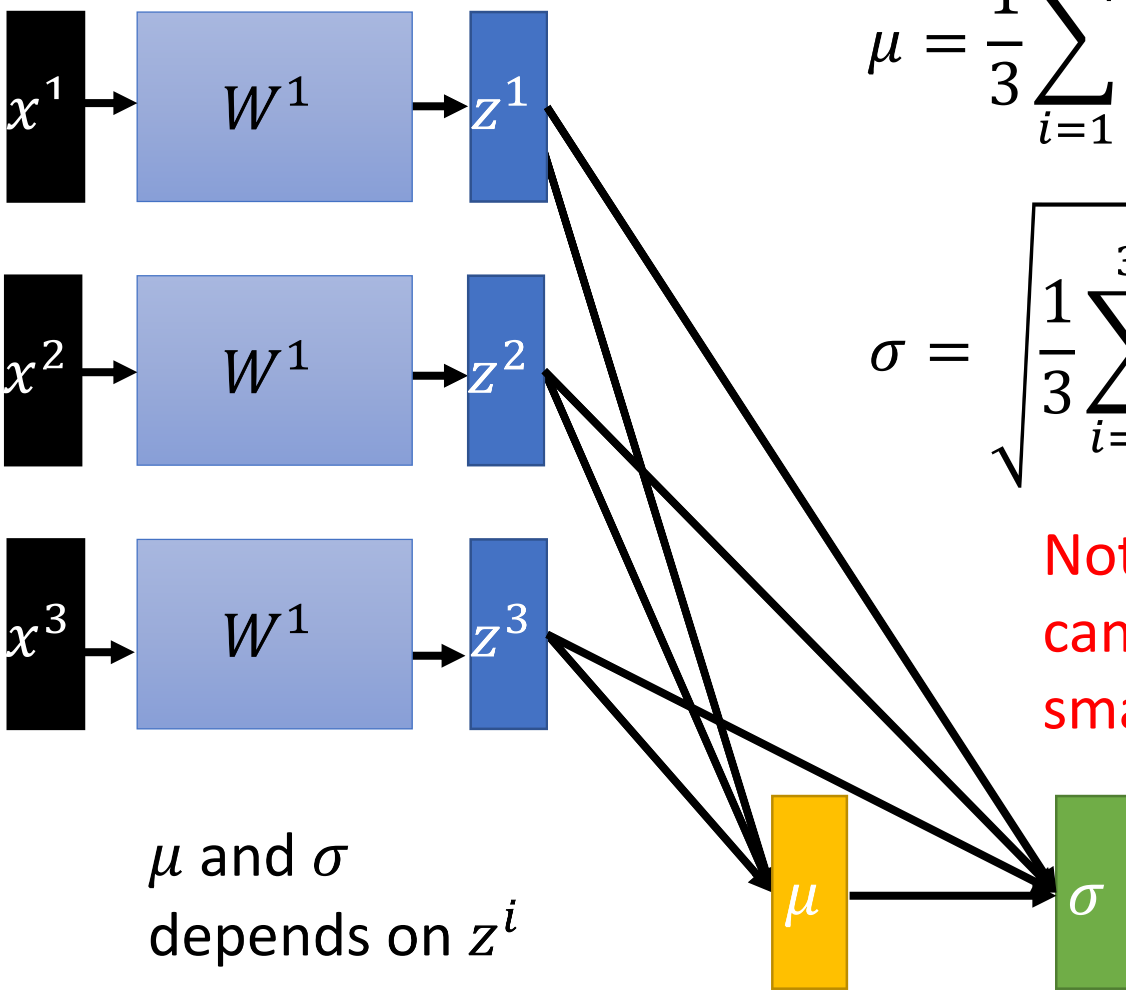
Smaller learning rate can be helpful, but the training would be slower.

Batch



x^1
W^1
z^1
Sigmoid
a^1
W^2
......
x^2
W^1
z^2
Sigmoid
a^2
W^2
......
x^3
W^1
z^3
Sigmoid
a^3
W^2
......
Batch
z^1 z^2 z^3
=
W^1
x^1 x^2 x^3

Batch normalization



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

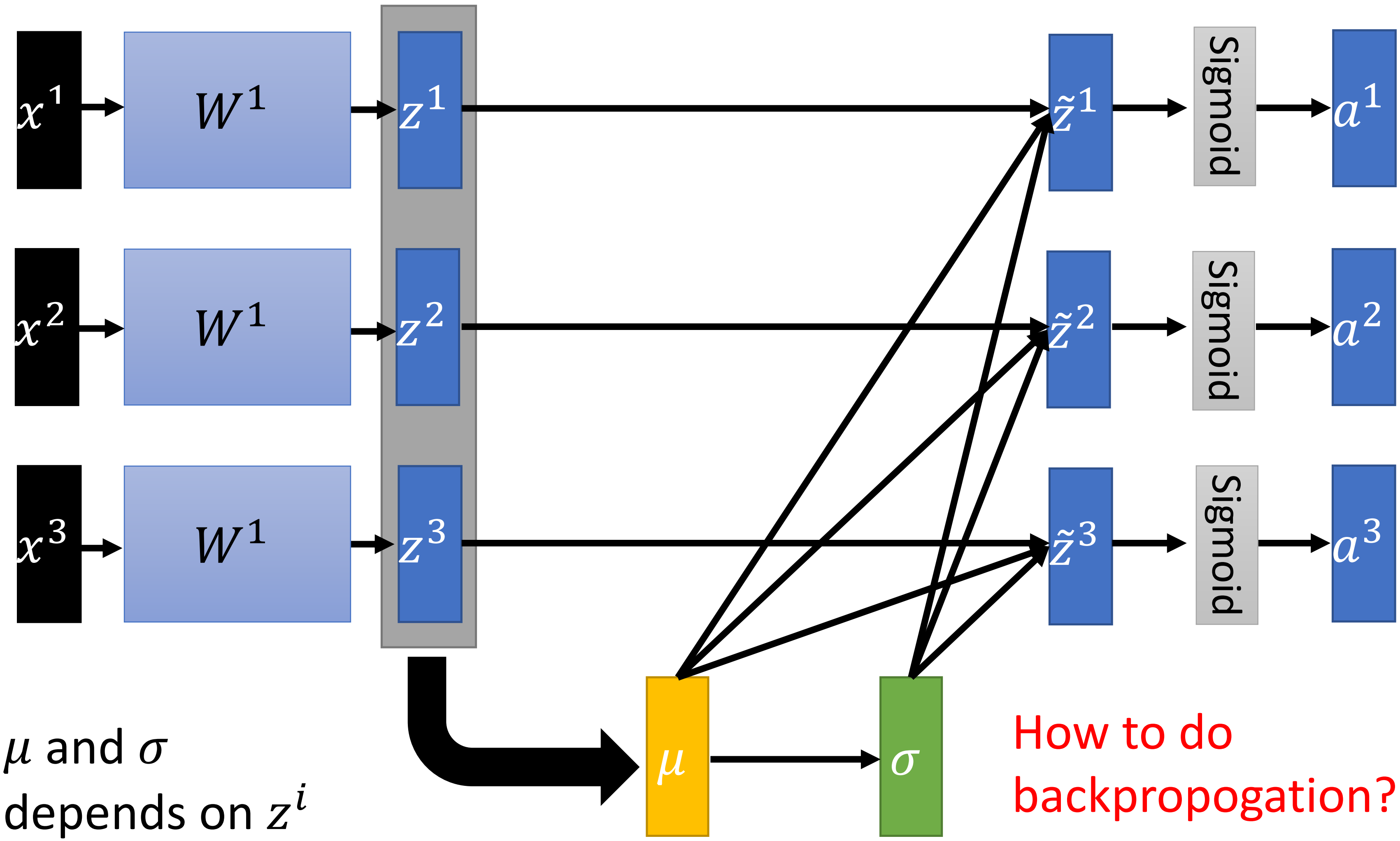
$$\sigma = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(z^i-\mu)^2}$$

Note: Batch normalization cannot be applied on small batch.

μ and σ depends on z^i

Batch normalization

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



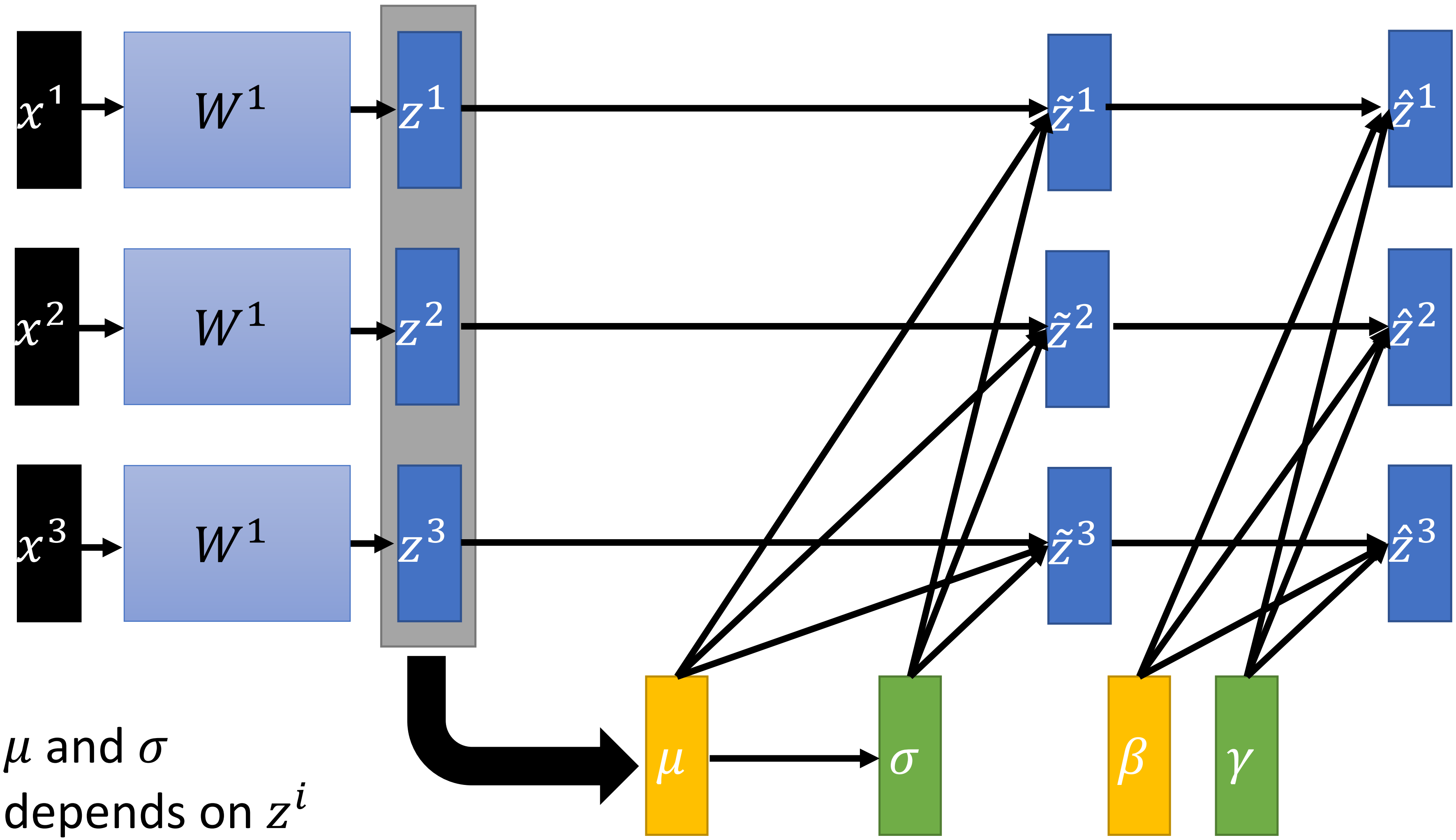
μ and σ depends on z^i

How to do backpropogation?

Batch normalization

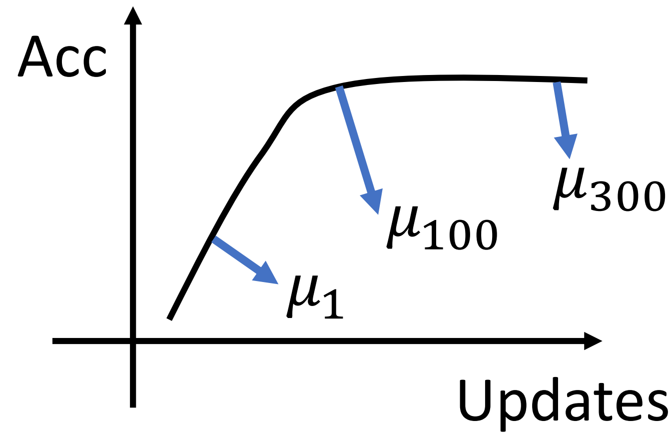
$$\tilde{z}^i = \frac{z^i - \mu}{\sigma}$$

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

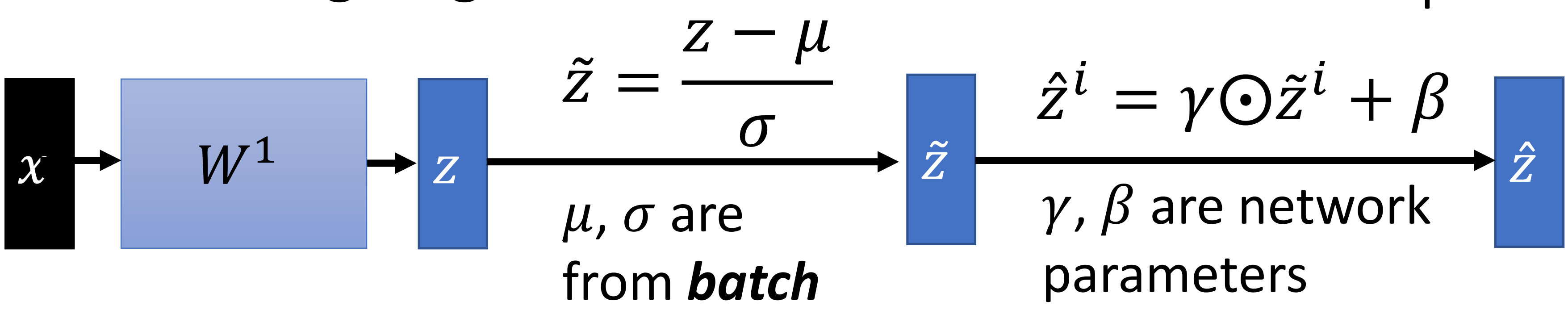


μ and σ depends on z^i

Batch normalization



- At testing stage:



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

Ideal solution:

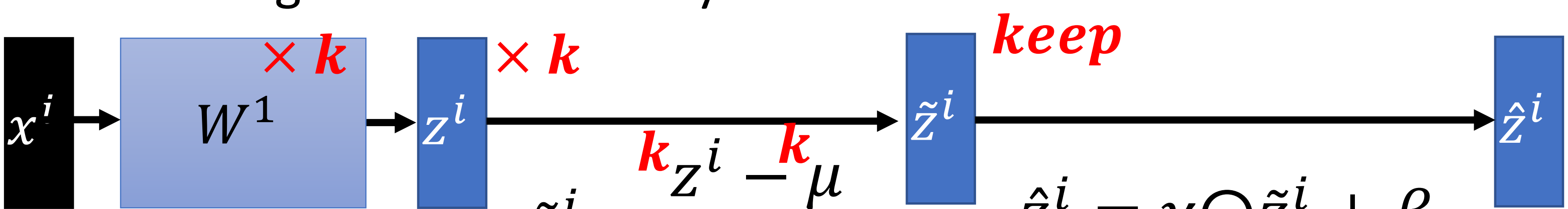
Computing μ and σ using the whole training dataset.

Practical solution:

Computing the moving average of μ and σ of the batches during training.

Batch normalization - Benefit

- BN reduces training times, and make very deep net trainable.
  - Because of less Covariate Shift, we can use larger learning rates.
  - Less exploding/vanishing gradients
    - Especially effective for sigmoid, tanh, etc.
- Learning is less affected by initialization.



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

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

- BN reduces the demand for regularization.

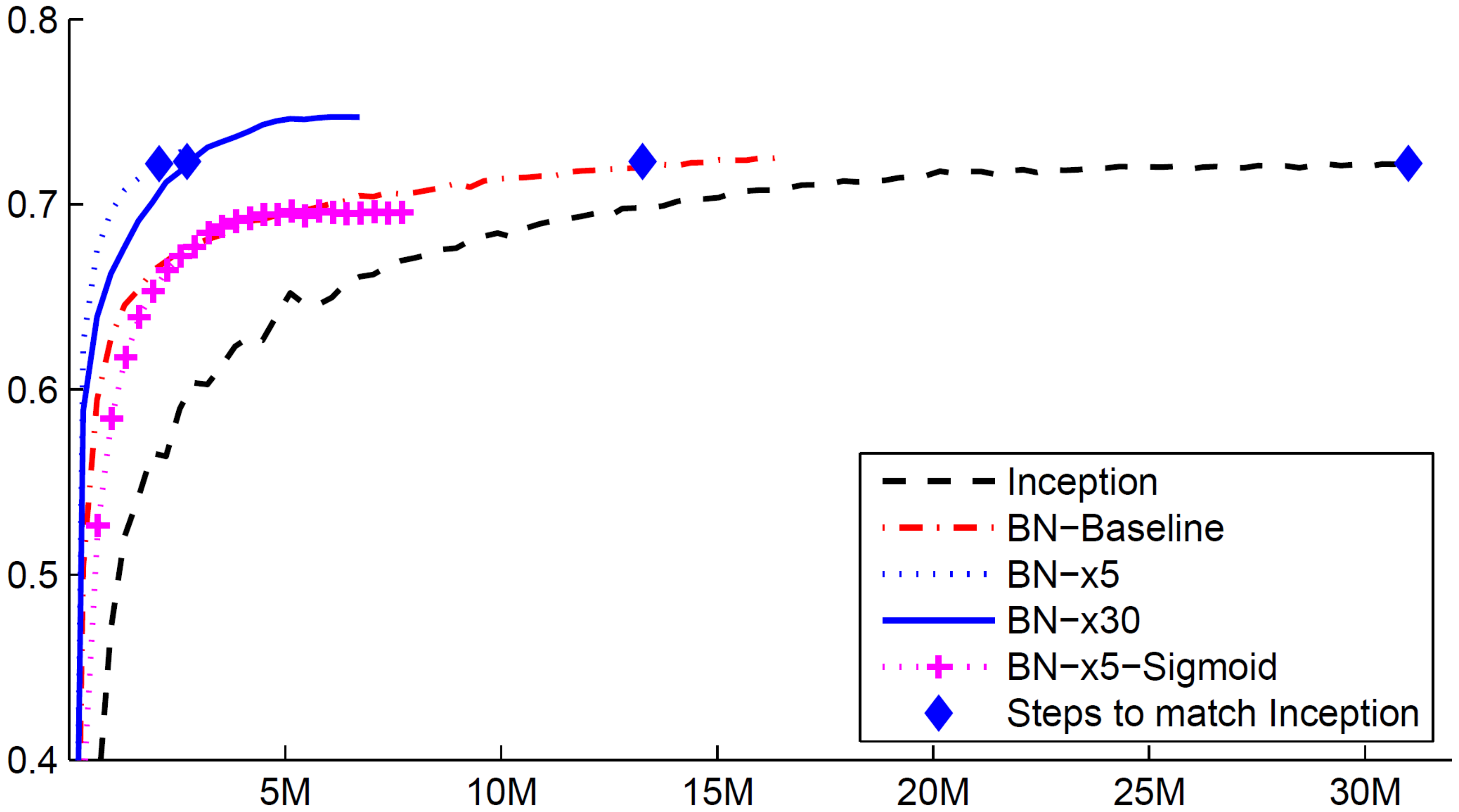


Figure 2: *Single crop validation accuracy of Inception and its batch-normalized variants, vs. the number of training steps.*