# Batch Normalization in Machine Learning

劉晉良

Jinn-Liang Liu

清華大學計算與建模科學研究所

Institute of Computational and Modeling Science

National Tsing Hua University, Taiwan

Oct 16, 2019

In mathematical modeling, statistical modeling and experimental sciences, the values of **dependent variables** depend on the values of **independent variables**. The dependent variables represent the output or outcome whose variation is being studied. The independent variables, also known in a statistical context as **regressors**, represent inputs or causes, that is, potential reasons for variation. Depending on the context, an independent variable is sometimes called a "**predictor** variable", **regressor**, **covariate**, "controlled variable", "manipulated variable", "explanatory variable", exposure variable (see reliability theory), "risk factor" (see medical statistics), "**feature**" (in machine learning and pattern recognition) or "input variable."

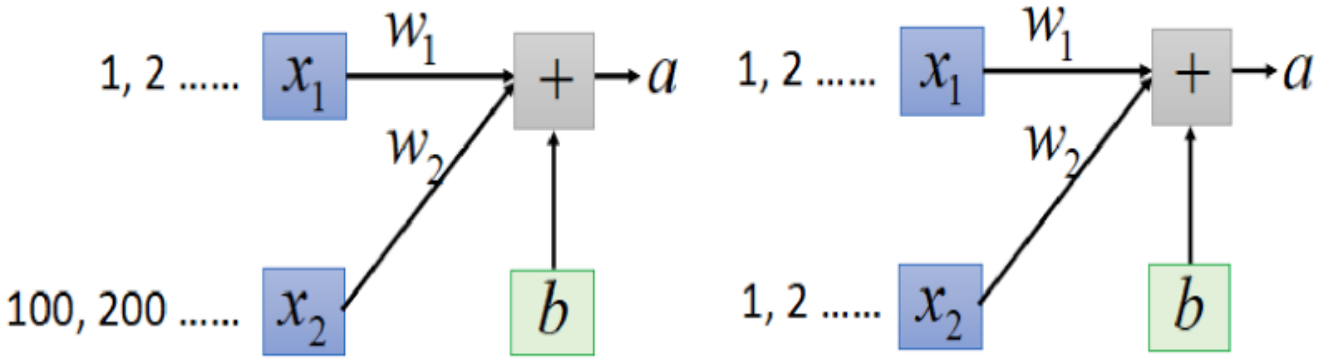
---- Wikipedia

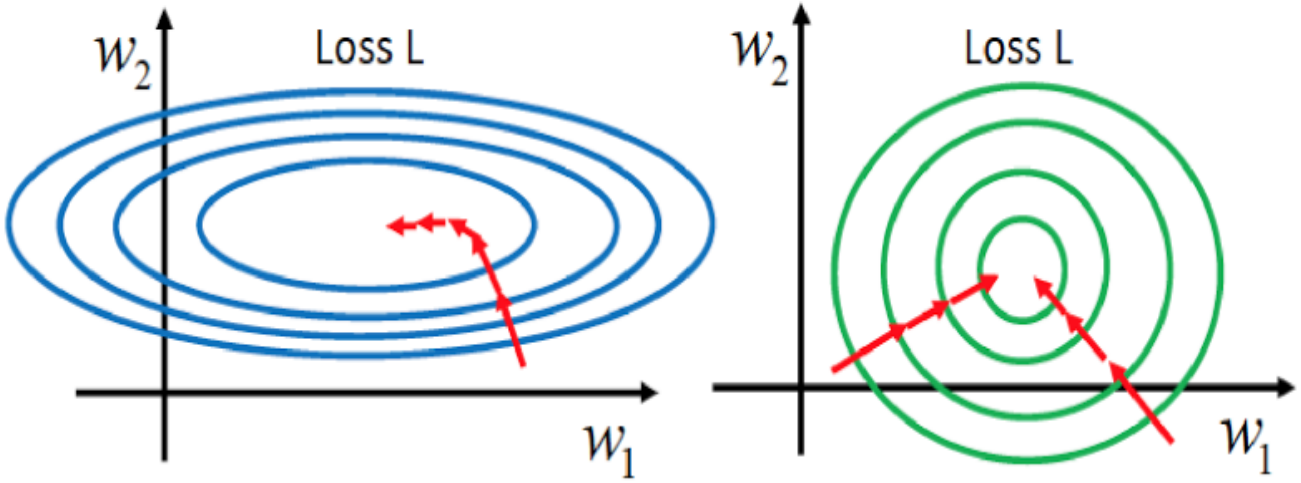
# 台大李宏毅

先從 Feature scaling 或是稱作 Feature Normalization 說起

## Feature Scaling

Make different features have the same scaling




假設 x1 跟 x2 的數值差距很大

x1 值的範圍 1,2,3,4, ...

x2 值的範圍 100, 200, 300, ...

x1 的 weight 是 w1，x2 的 weight 是 w2

w1 前面乘的值比較小，所以他對於結果的影響比較小

w2 前面乘的值比較大，所以他對於結果的影響比較大

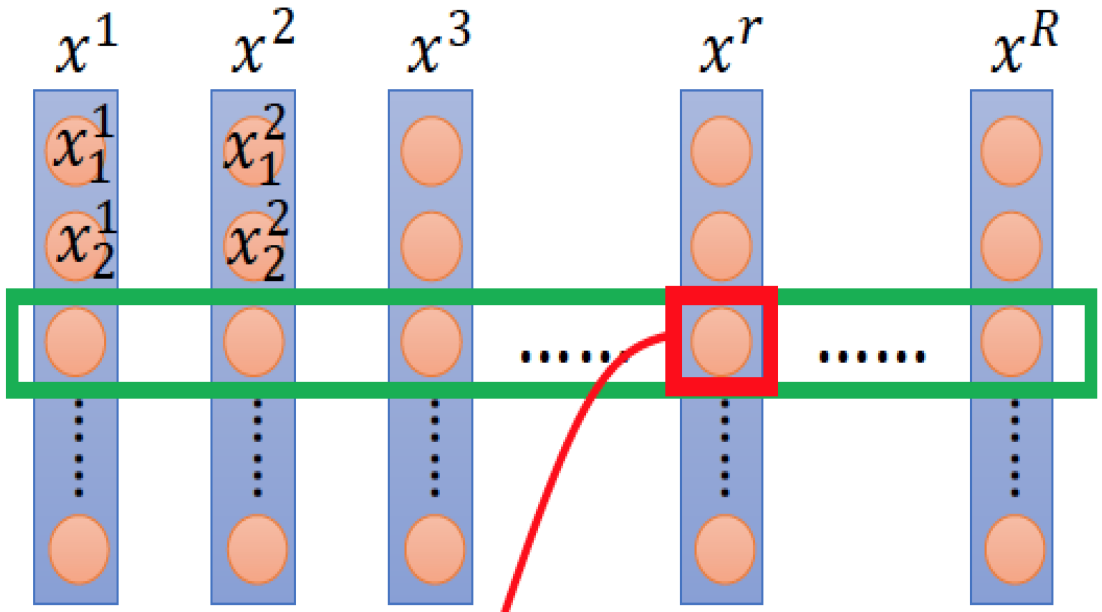
我們把 w1, w2 對於 Loss 的值的影響做圖 (左下圖藍色圈),

在 w1 方向上的變化斜率比較小，Gradient 變化比較小

在 w2 方向上的變化斜率比較小，Gradient 變化比較大

這樣會讓 training 變得比較不容易，要分別在不同方向上給不同的 learning rate

# Feature Scaling




For each dimension i:

mean: $m_i$

standard deviation: $\sigma_i$

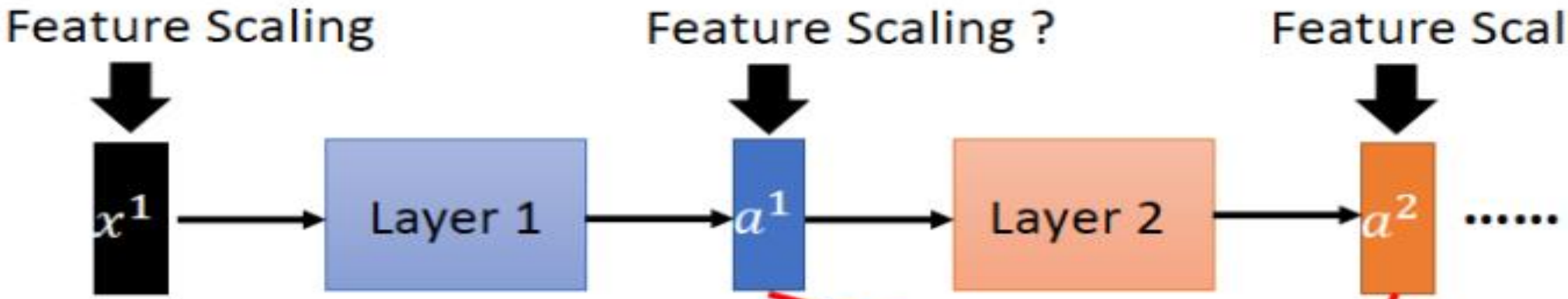
$$x_i^r \leftarrow \frac{x_i^r - m_i}{\sigma_i}$$

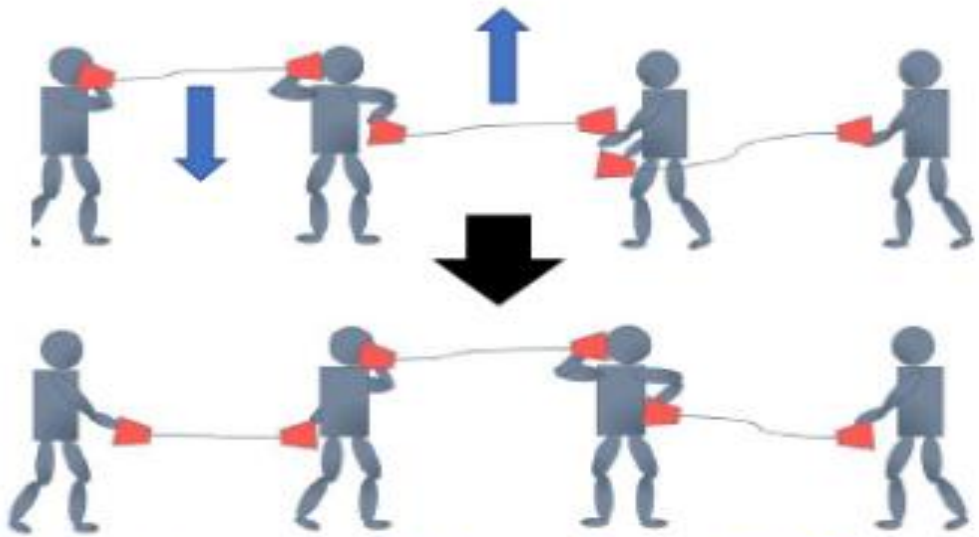
The means of all dimensions are 0, and the variances are all 1

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

Feature Scaling 做法：
對於第 i 個維度的每一筆 data
都減掉該維度的 mean
再除以該維度的 standard deviation

# How about Hidden Layer?



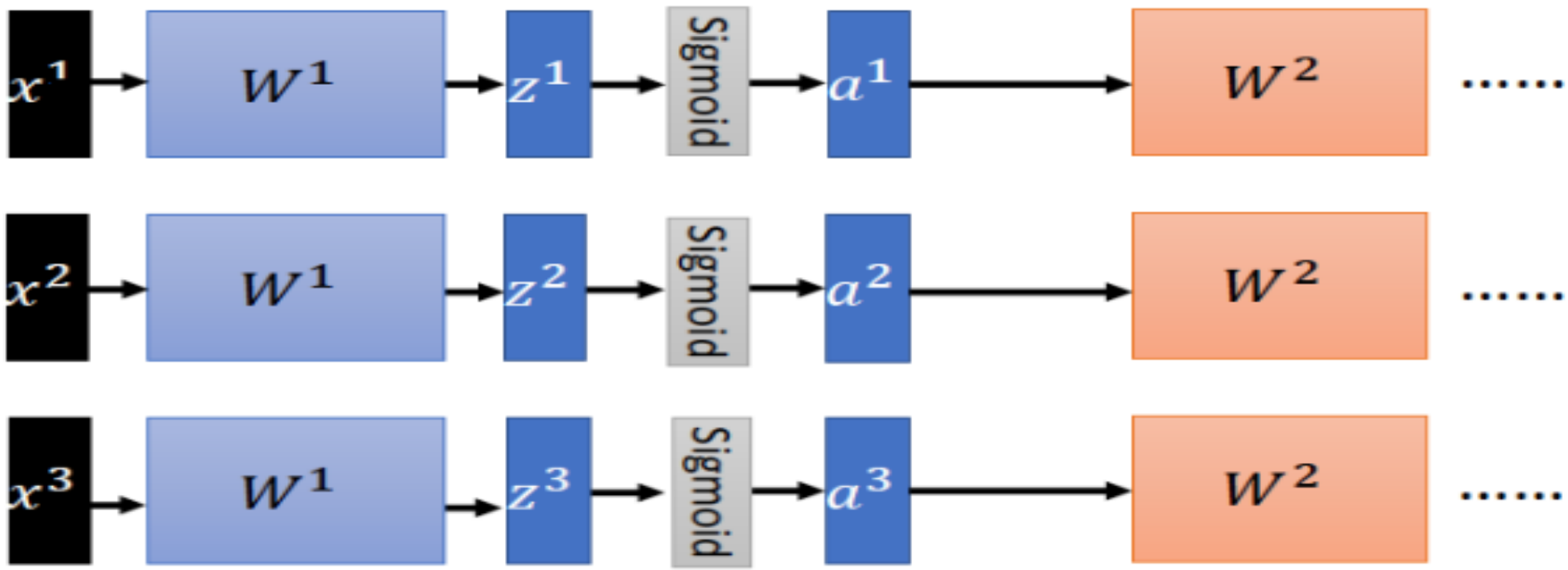
Internal Covariate Shift

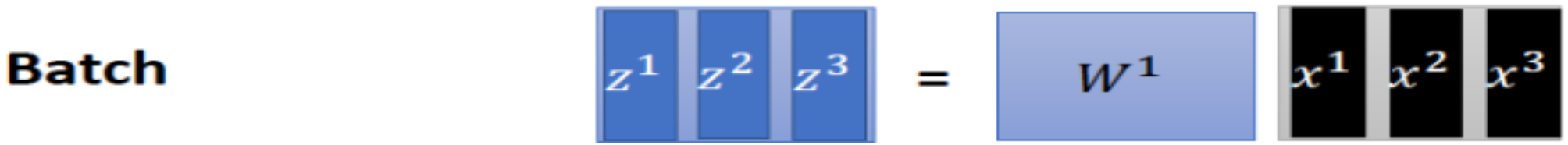
Difficulty: their statistics change during the training ...

Batch normalization

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

# Batch



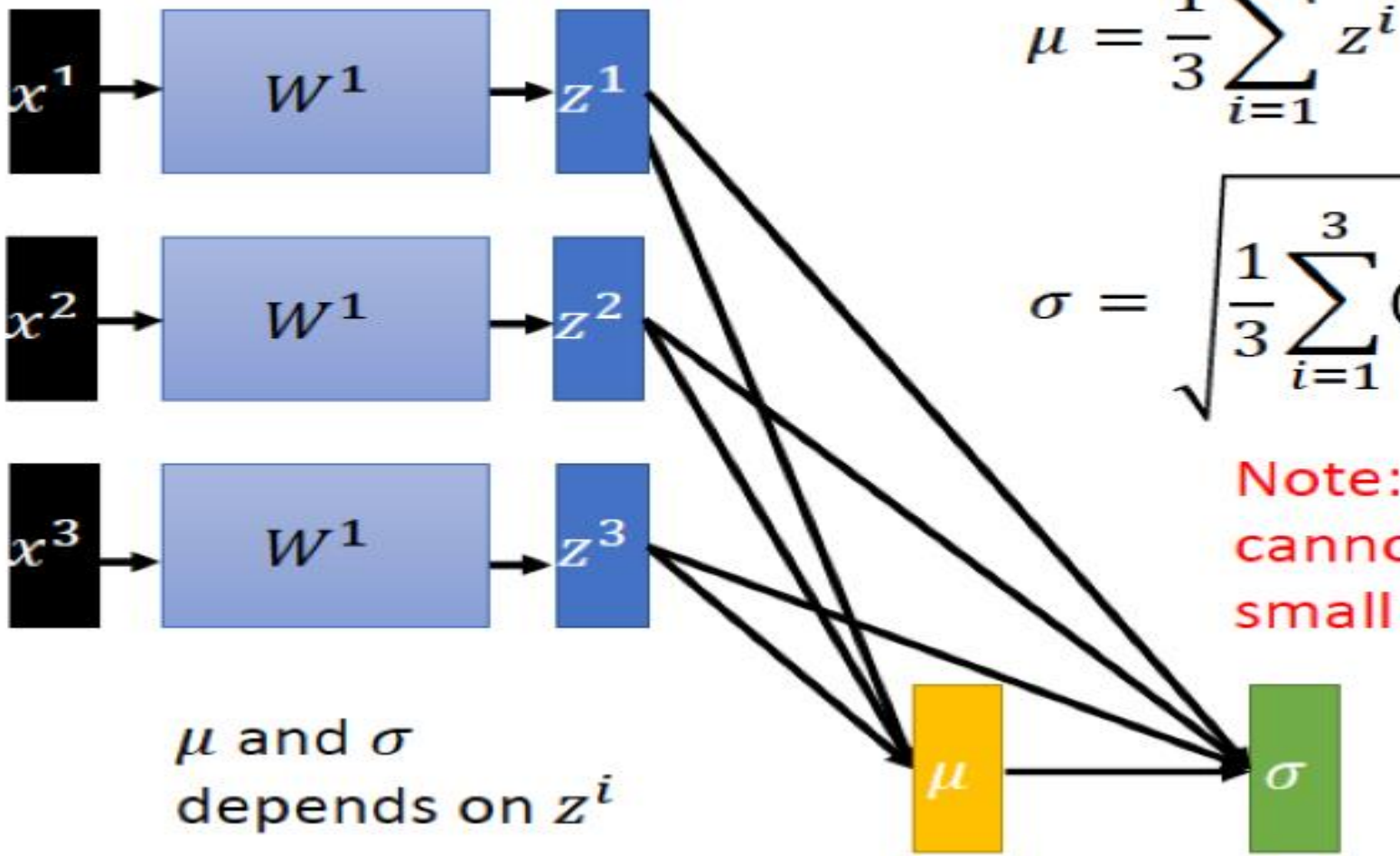




使用 GPU 加速運算時，假設我們的 Batch = 3

x1, x2, x3 這三筆資料會是平行運算的，我們把 x1, x2, x3 排在一起變成一個 input matrix X

對 weight matrix W 運算，得到一個 output matrix Z

# Batch normalization




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

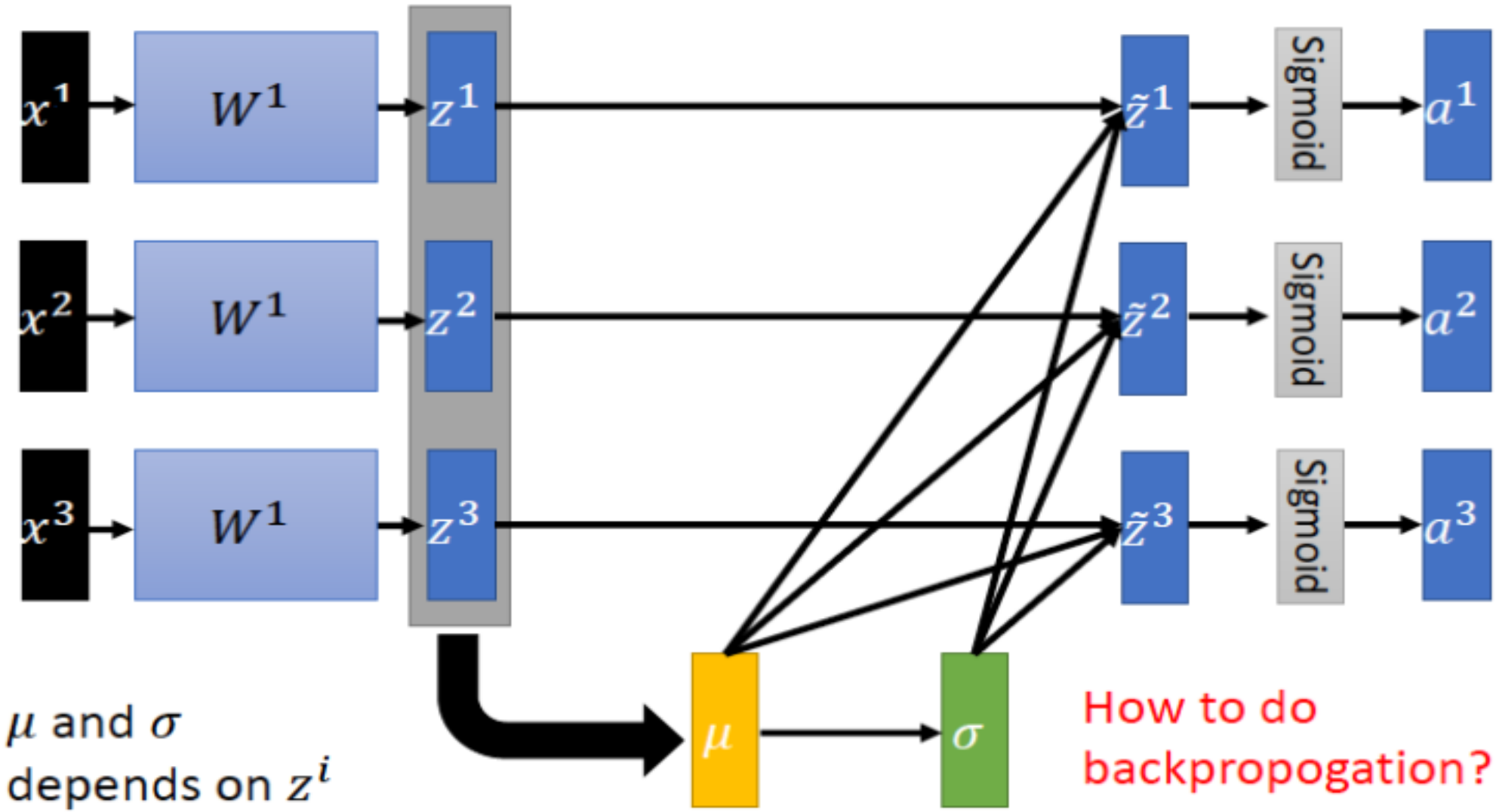
$$\sigma = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(z^i - \mu)^2}$$

Note: Batch normalization cannot be applied on small batch.

$\mu$ and $\sigma$ depends on $z^i$

# Batch normalization

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




$\mu$ and $\sigma$ depends on $z^i$

How to do backpropogation?

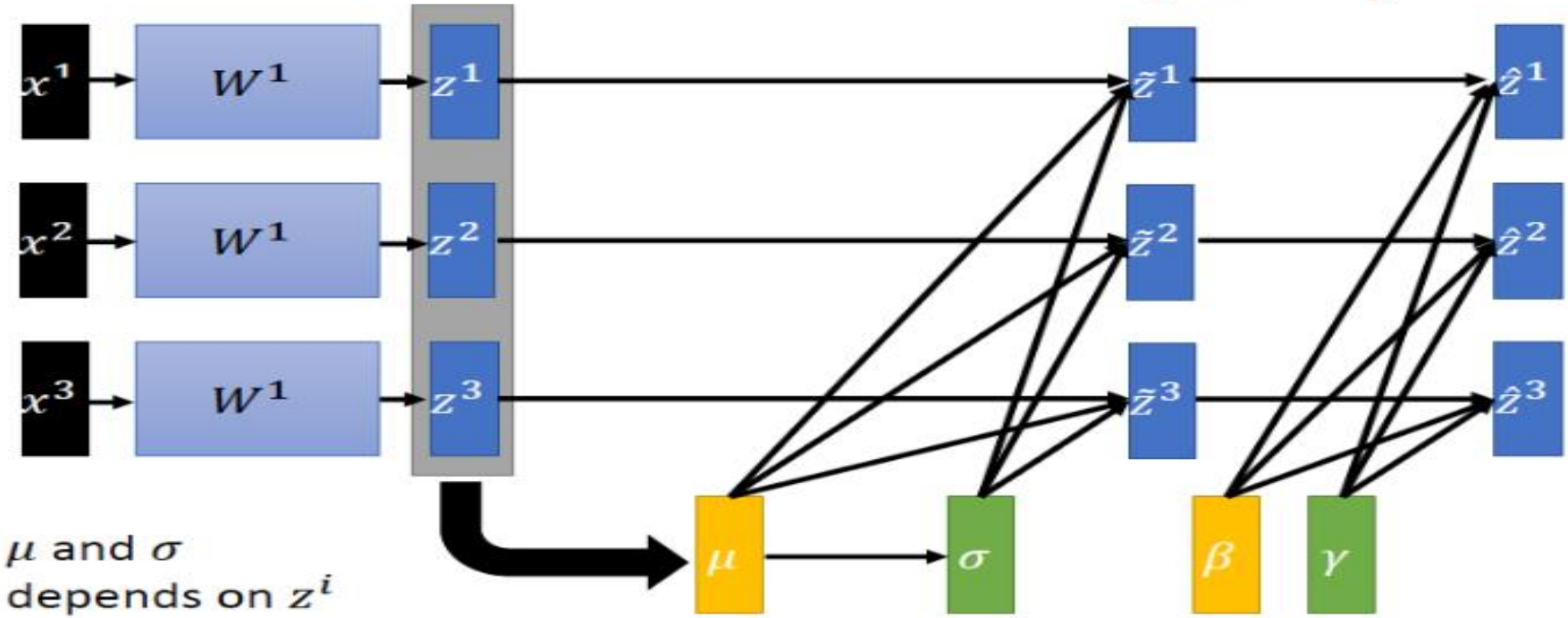
我們把 Z Normalize 後，可以直接丟進 activation function

但是有時候我們不希望丟進 activation function 的資料 mean 都是 0，標準差都是 1

# Batch normalization

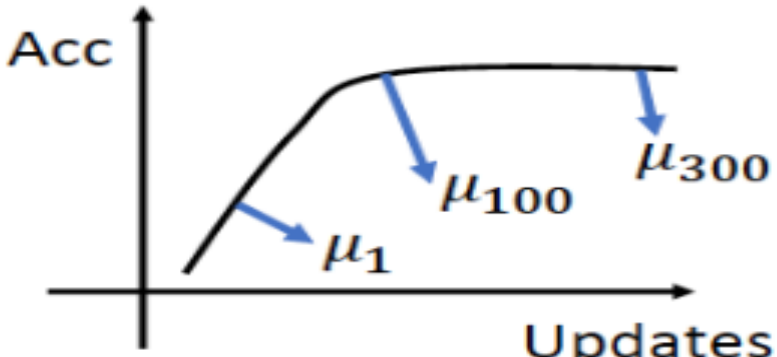
$$\tilde{z}^i = \frac{z^i - \mu}{\sigma}$$

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

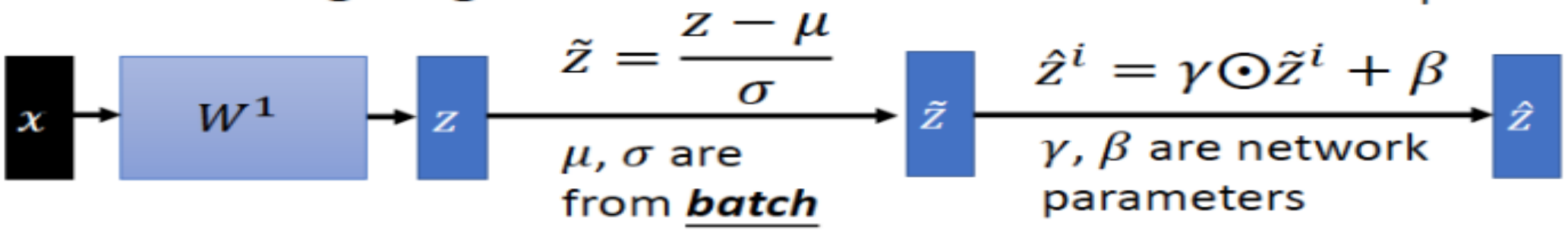


接下來看一下 testing 時怎麼做：

# Batch normalization




- At testing stage:




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

Ideal solution:

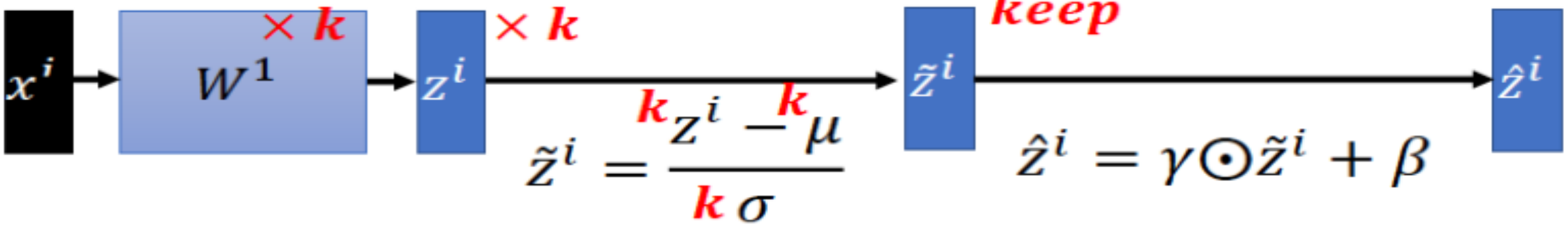
Computing $\mu$ and $\sigma$ using the whole training dataset.

Practical solution:

Computing the moving average of $\mu$ and $\sigma$ 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.




- BN reduces the demand for regularization.

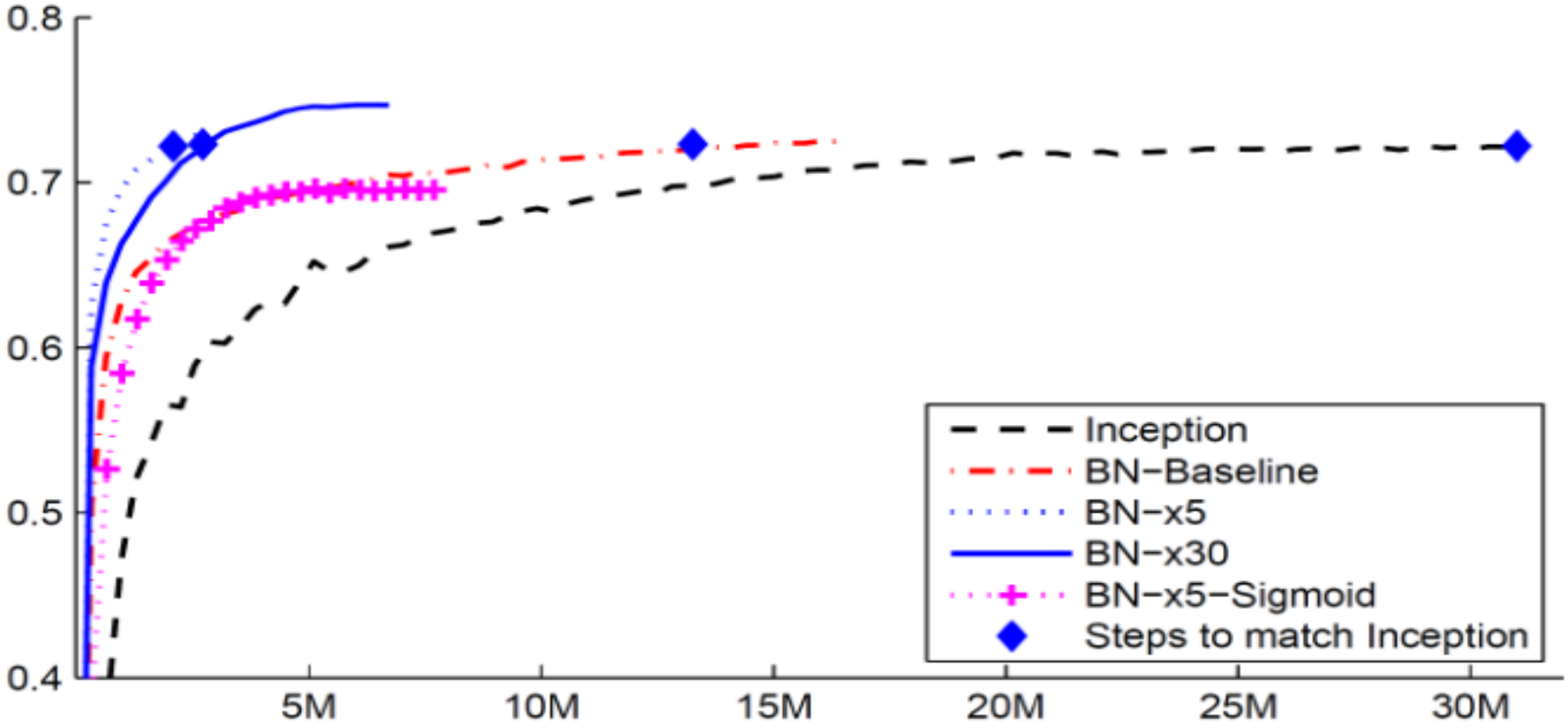



Figure 2: *Single crop validation accuracy of Inception and its batch-normalized variants, vs. the number of training steps.*

# Dataset Shift in Classification:

# Approaches and Problems

**Francisco Herrera**

University of Granada, Spain

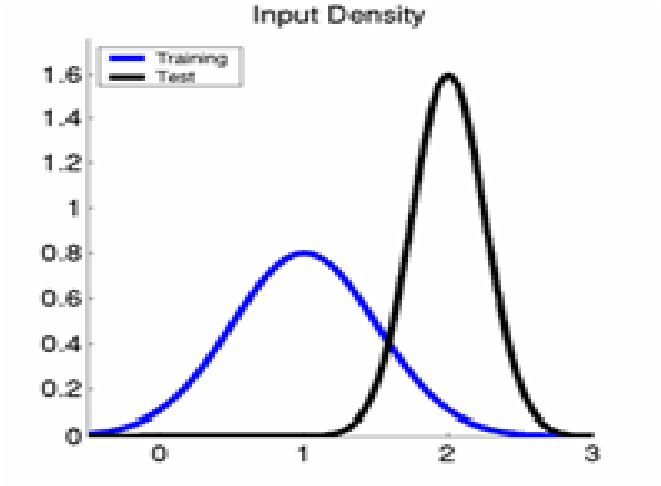
**Definition 1.** ***Dataset shift appears when training and test joint distributions are different.*** That is, when $P_{tra}(y, x) \neq P_{tst}(y, x)$

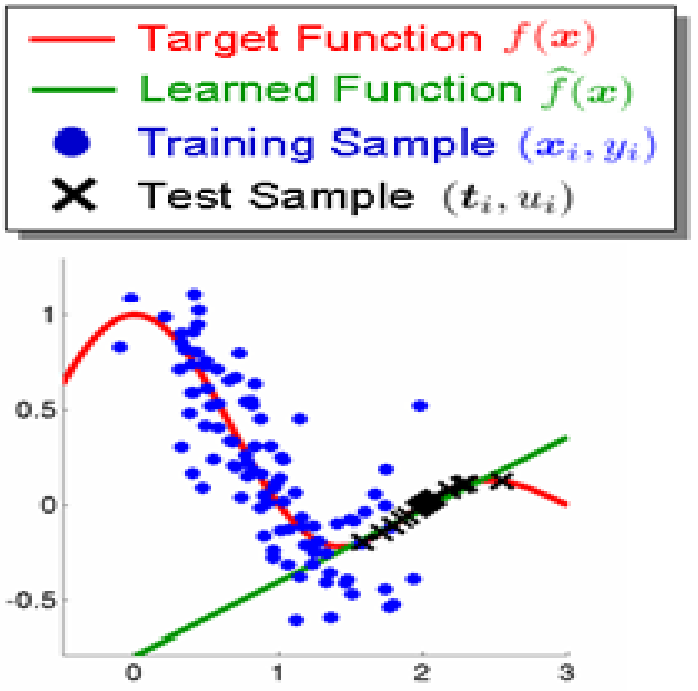
**Definition 2.** Covariate shift appears only in $X \rightarrow Y$ problems, and is defined as the case where $P_{tra}(y|x) = P_{tst}(y|x)$ and $P_{tra}(x) \neq P_{tst}(x)$.

# Characterizing the change. Types of Dataset Shift

## Covariate Shift

Training and test input follow different distributions, but functional relation remains unchanged.




Goal: Estimate test output from $\{(x_i, y_i)\}_{i=1}^{n}$