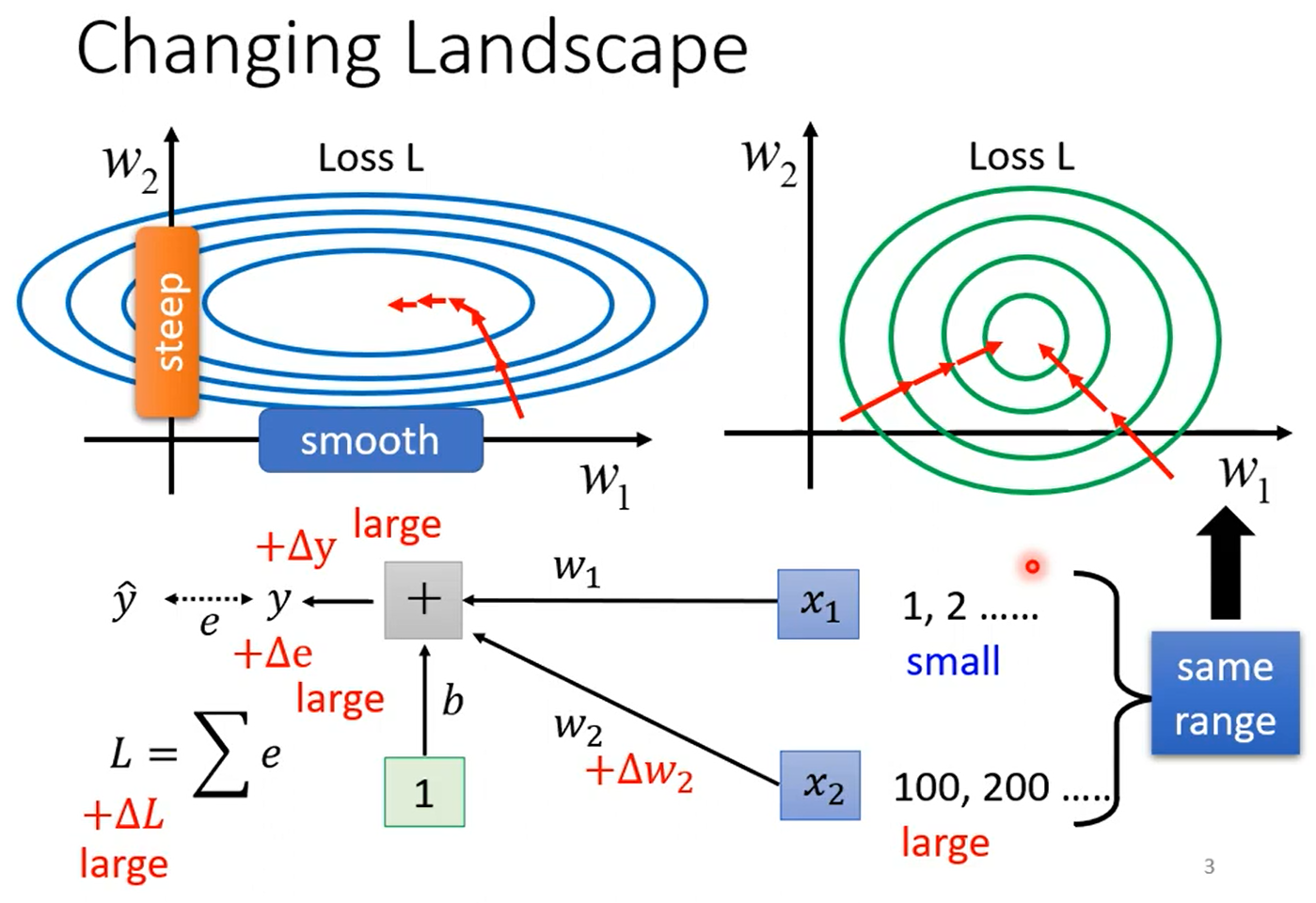
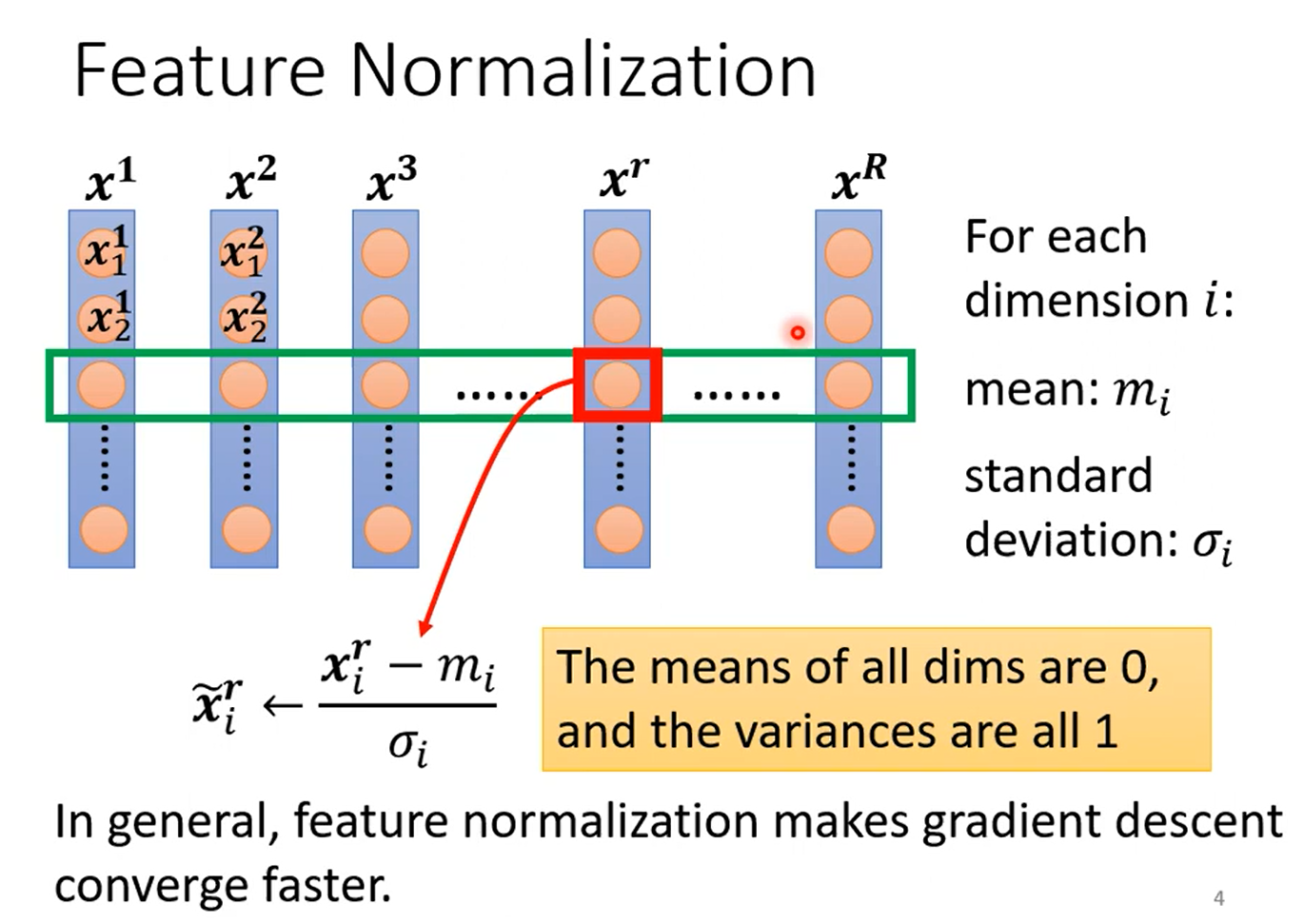
BN可以直接改变error surface的形状

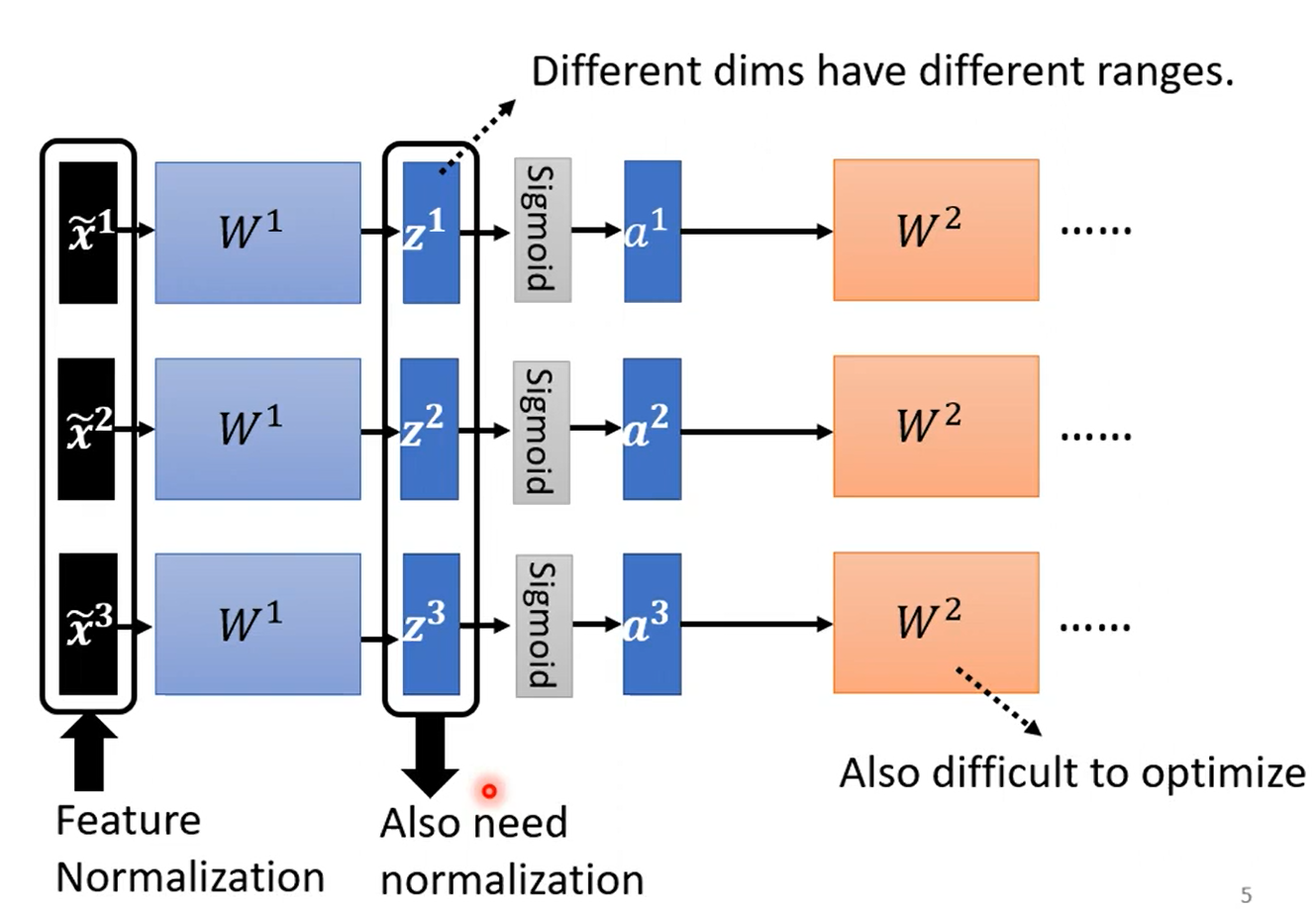
## 1.为什么出现椭圆形的损失函数，因为没有进行特征标准化（feature normalization）



## 2.特征标准化有助于加速训练



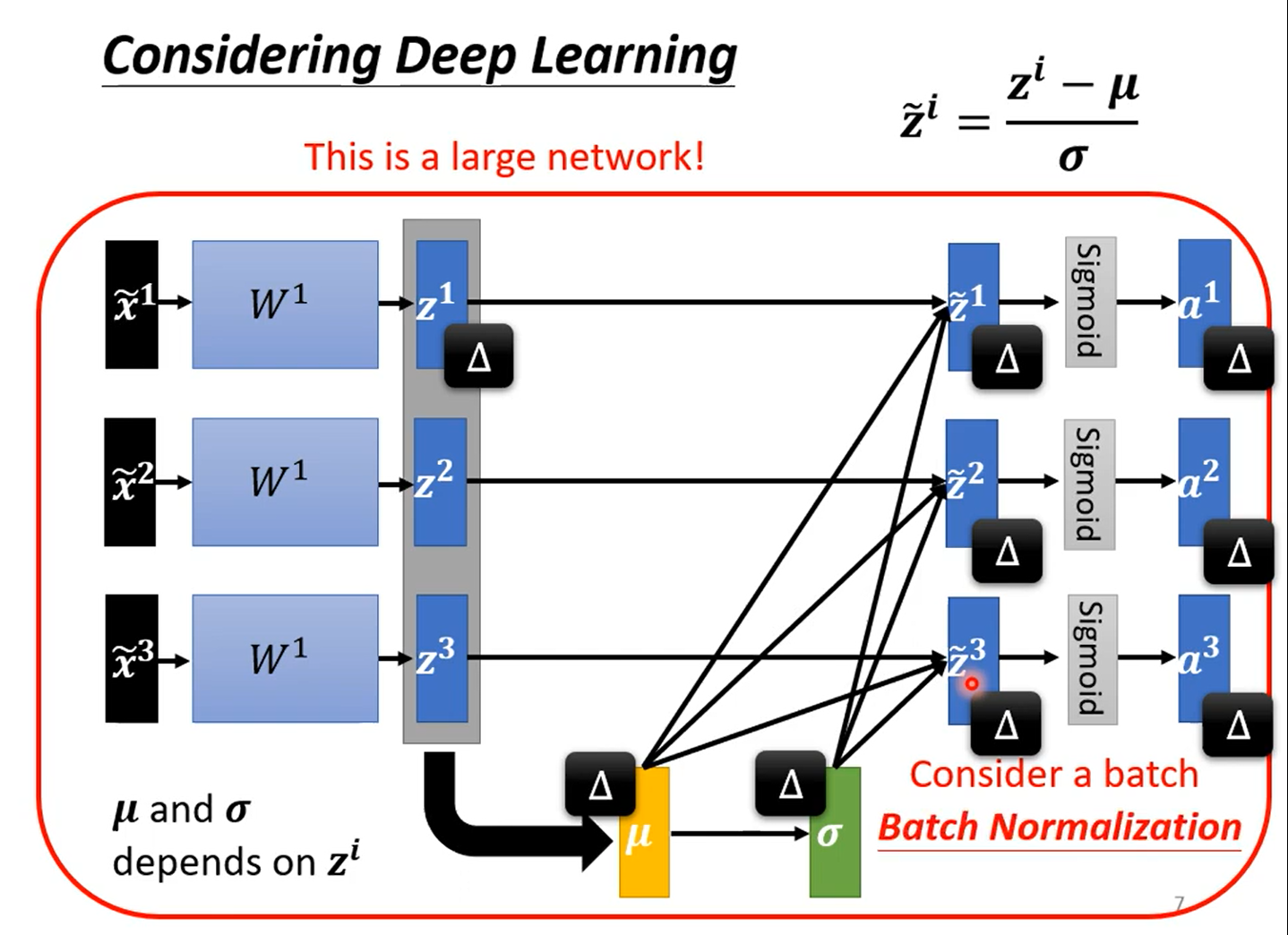
## 3.但是只要经过一层之后，又会造成数据分布出现很大的差异，那么又需要进行标准化，因此需要BN



### 3.1 一个trirck，是在激活函数之前对z进行标准化，还是在激活函数之后对a进行标准化？

一般而言差不多，但是如果用的是sigmoid激活函数的话，推荐是对z进行标准化，因为将z处理到0附近，则sigmoid的梯度比价大一点

### 3.2 BN作为一个layer，算一个batch的均值与方差，然后进行批标准化（适用于batchsize比较大的时候）



### 3.3 实际的BN，为什么还需要加上γ和β？有什么后果？

#### 3.3.1为什么要加？

本来BN之后，batch的mean就是0，是对输出的一种限制，有时候我们希望输出不能均值为0

#### 3.3.2 不会破坏BN的作用吗？

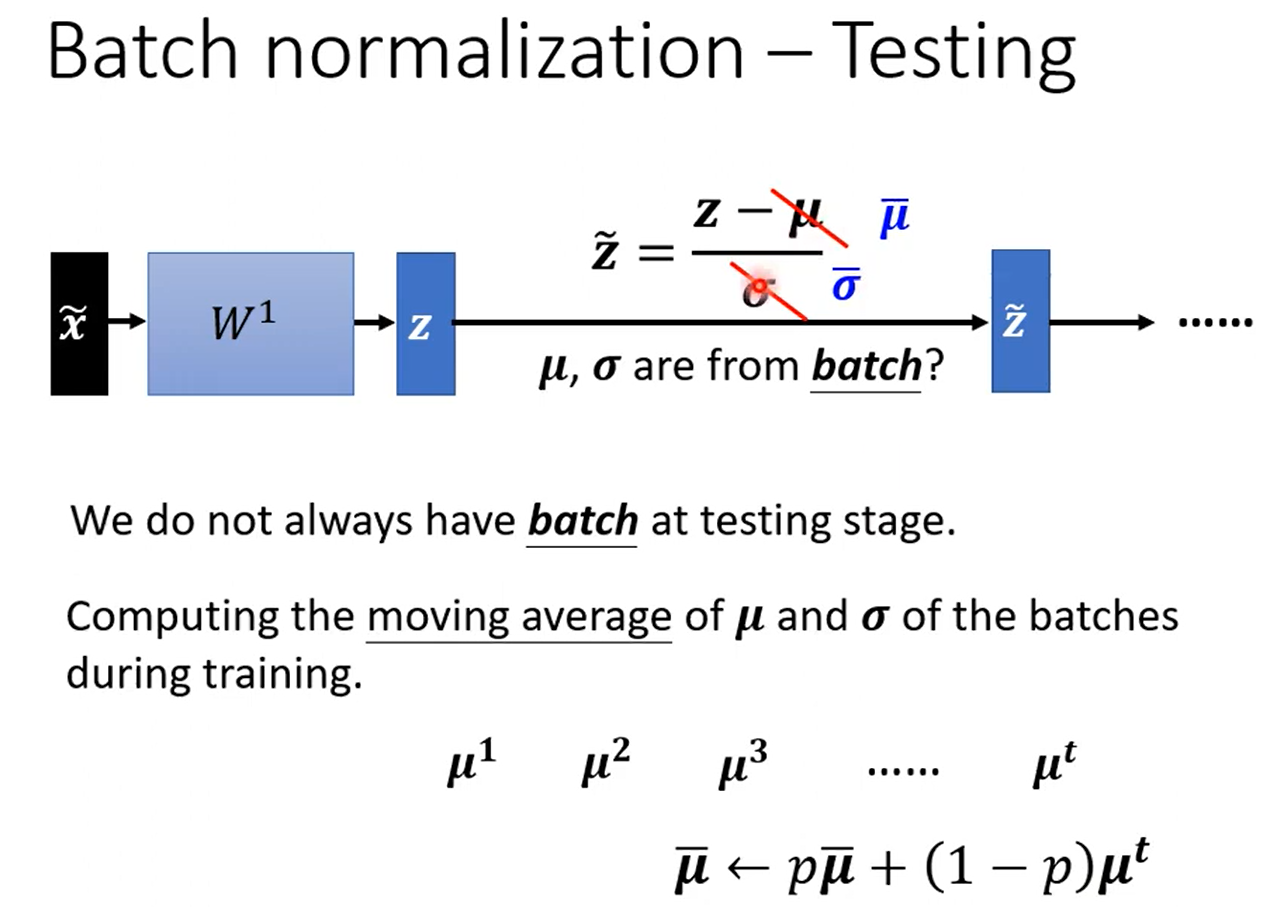
初始设置γ全是1，β全是0，也就是保留BN的作用

## 4.测试时候的BN

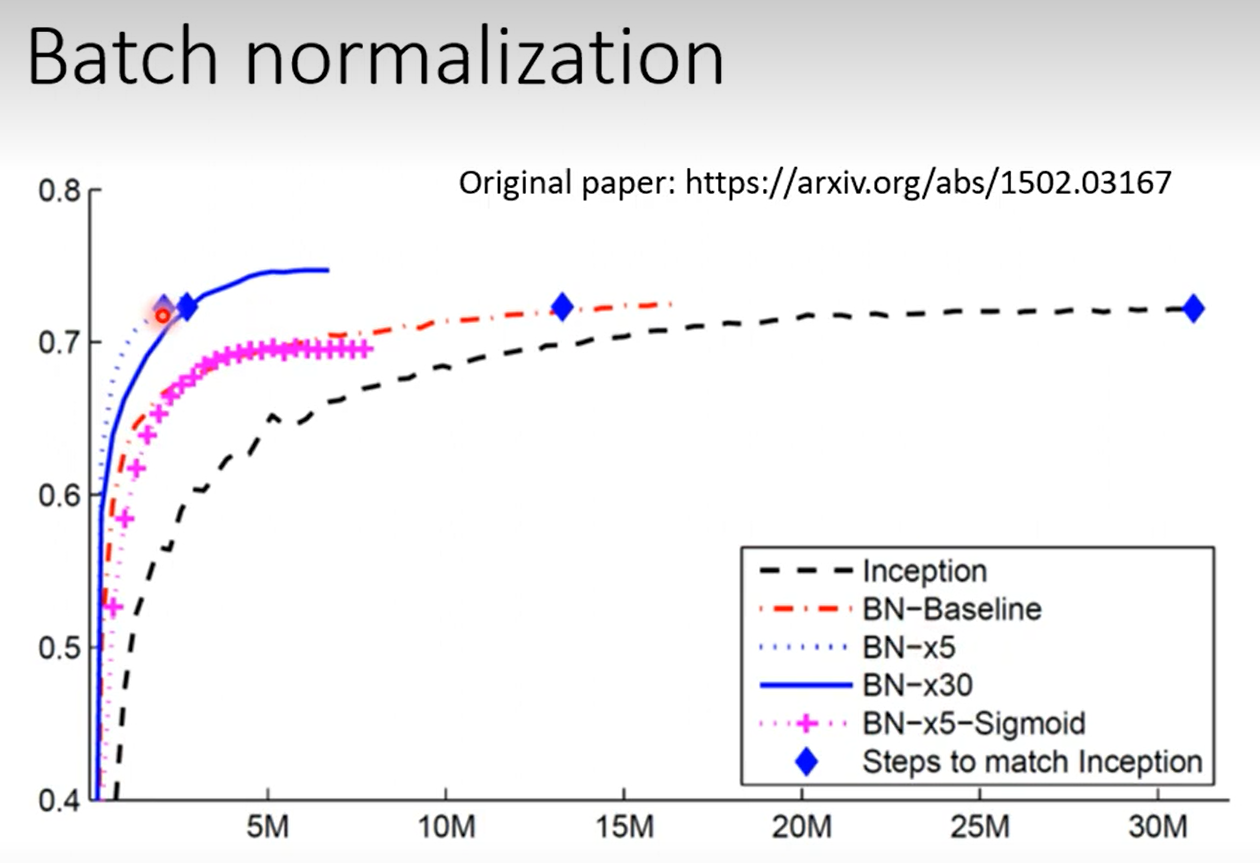
### 4.1 存在的问题与解决的方法

因为测试的时候经常是一个一个输入的，一个输入的时候就没有batch，那怎么算均值和标准差呢？

Pytorch中是对训练阶段的均值和方差进行移动平均

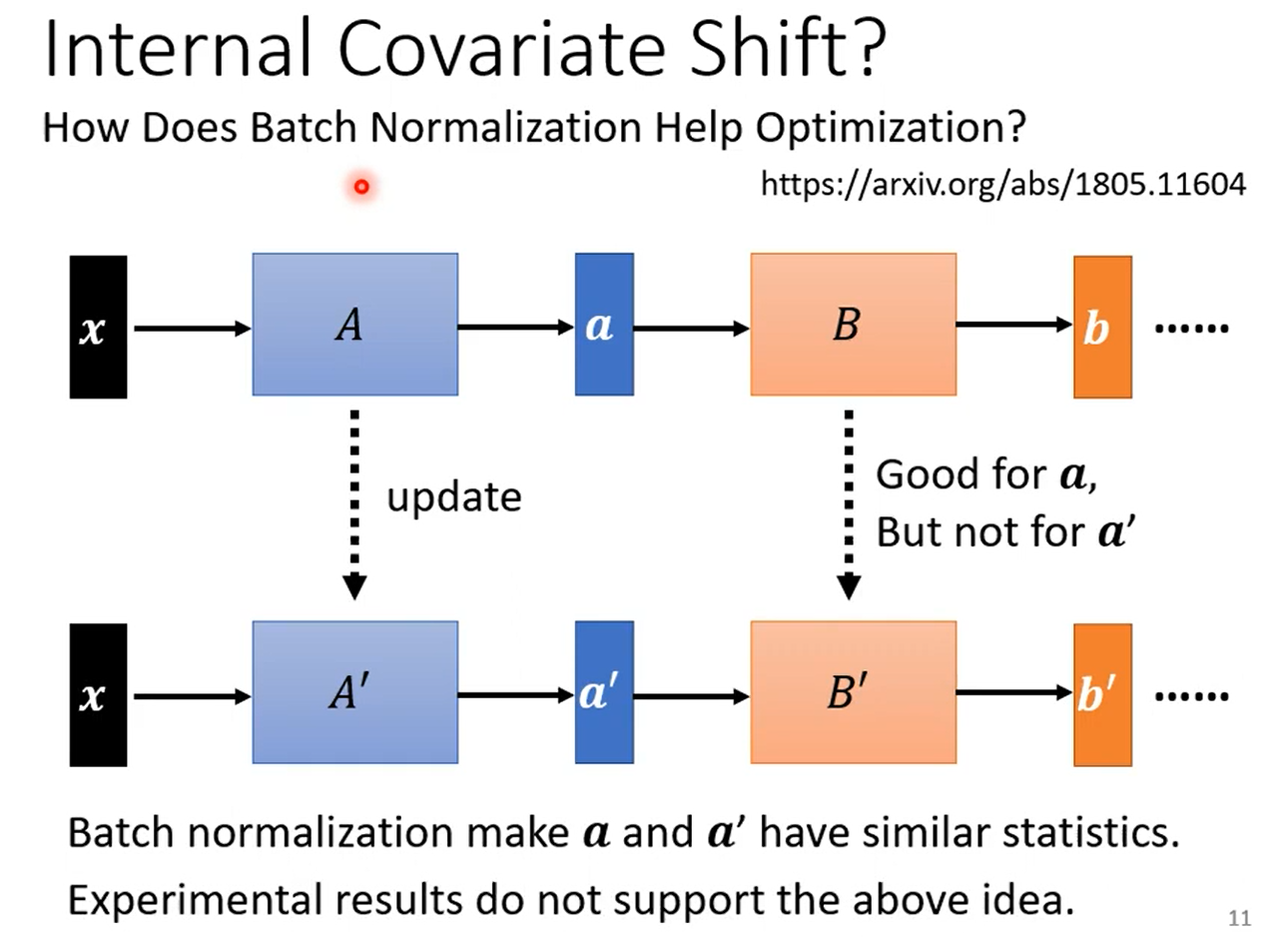


## 5.BN的效果（可以加速收敛，另外由于error surface比较平滑，可以使用比较大的学习率）



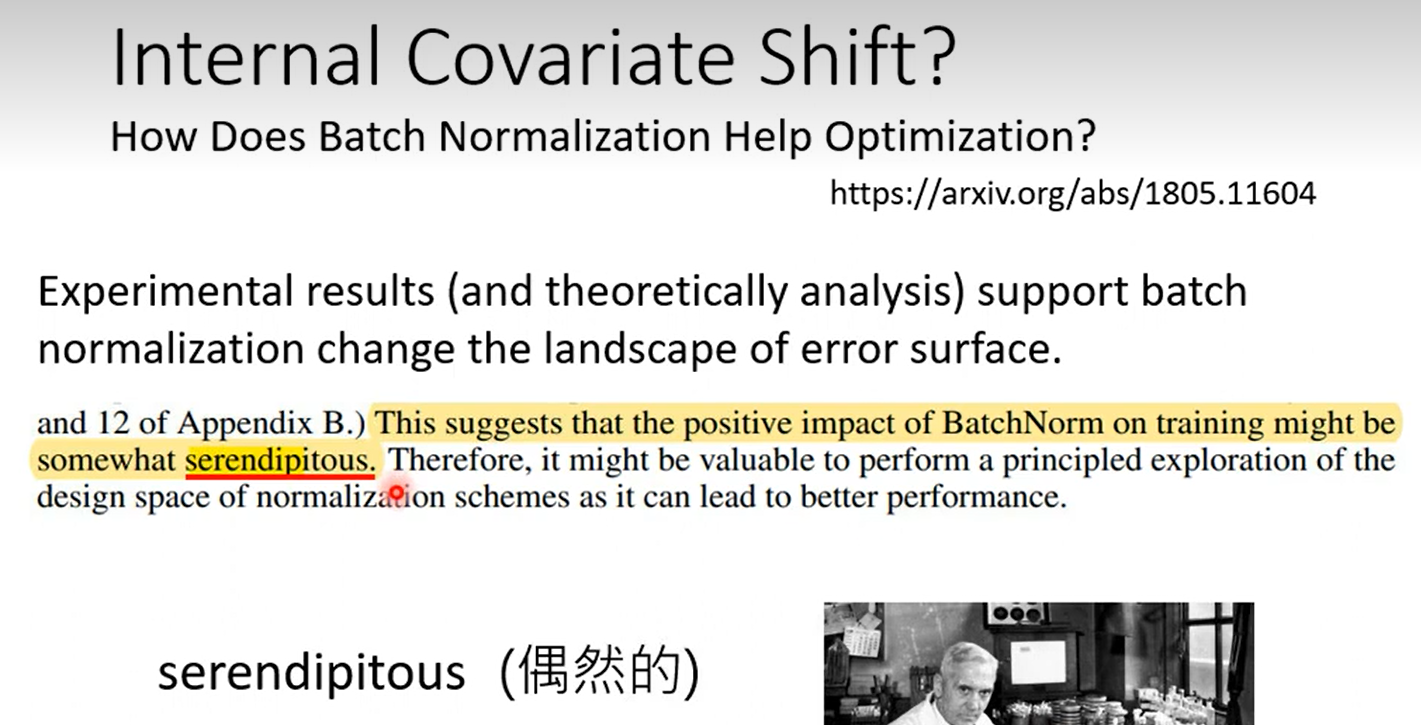
## 6.BN为什么会有效果？

Internal covariate shift



原作者的意思是BN可以让a与a’比较接近

How dose BN help optimation从理论和实践上证明BN可以让error surface变平坦



## 7.其它的normalization

