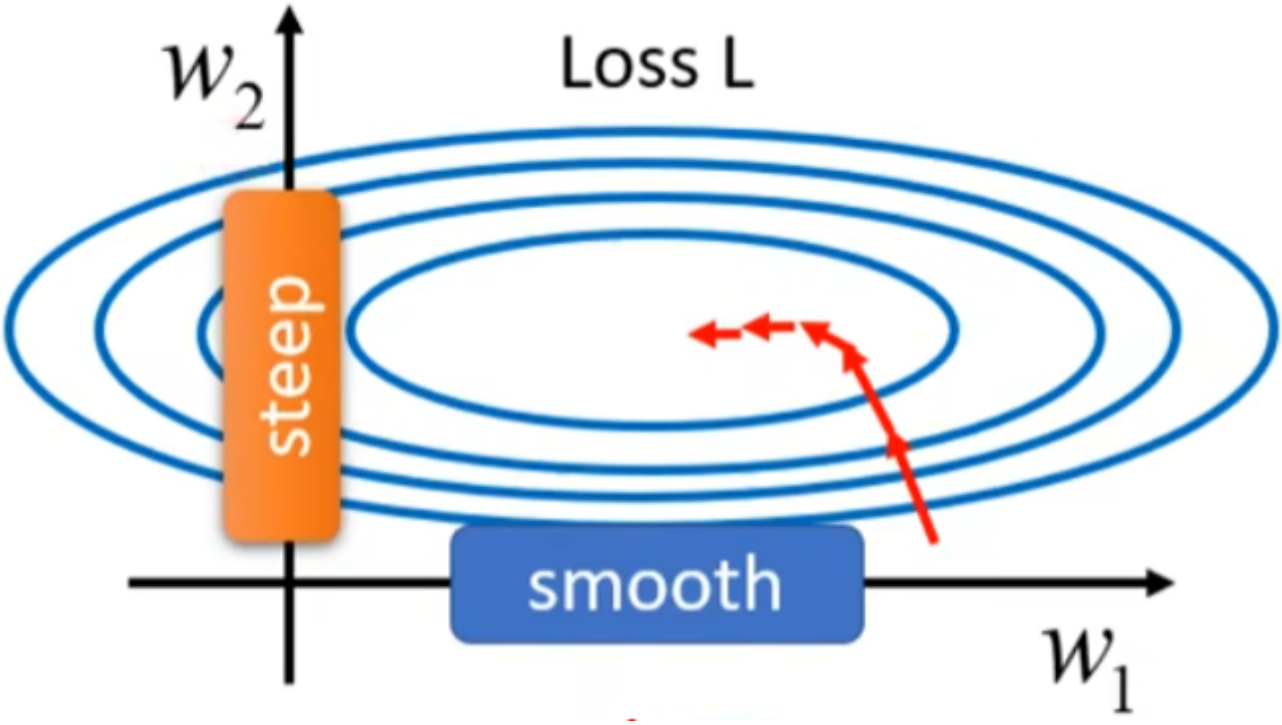
补充二：利用Batch Normalization优化训练

引入

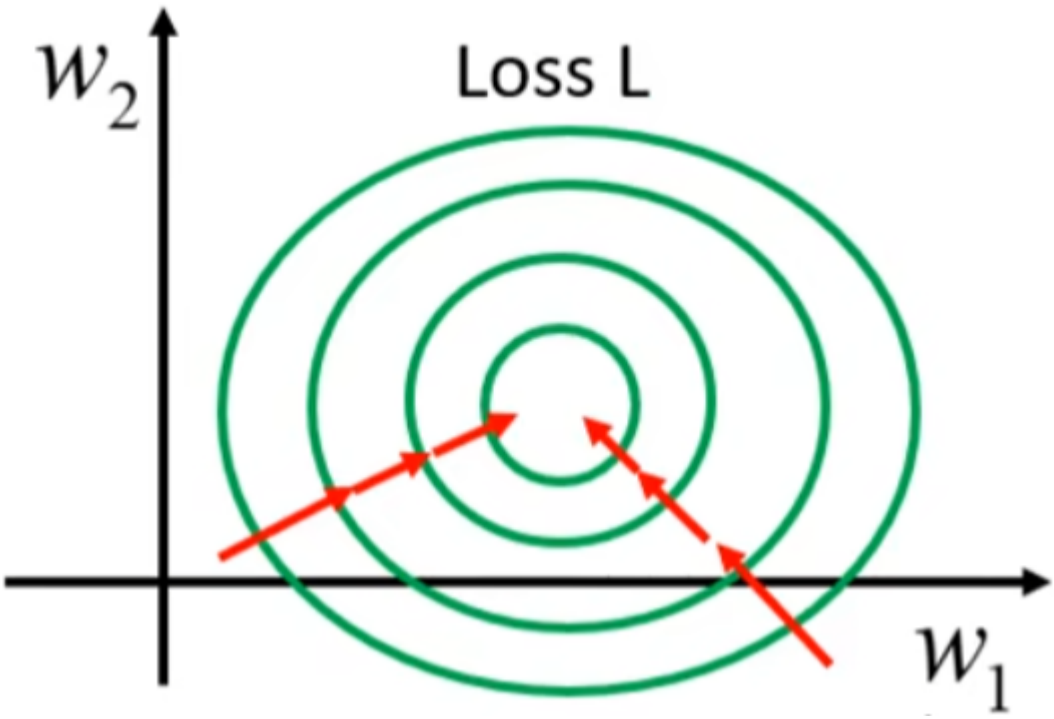
图中，w_2改变一点点，变化就很大，那能不能把error surface“铲平”（这样子就不会存在陡的时候需要慢慢摸索的情况了）



分析原因，由于$y = wx + b$，真实值差距的e：

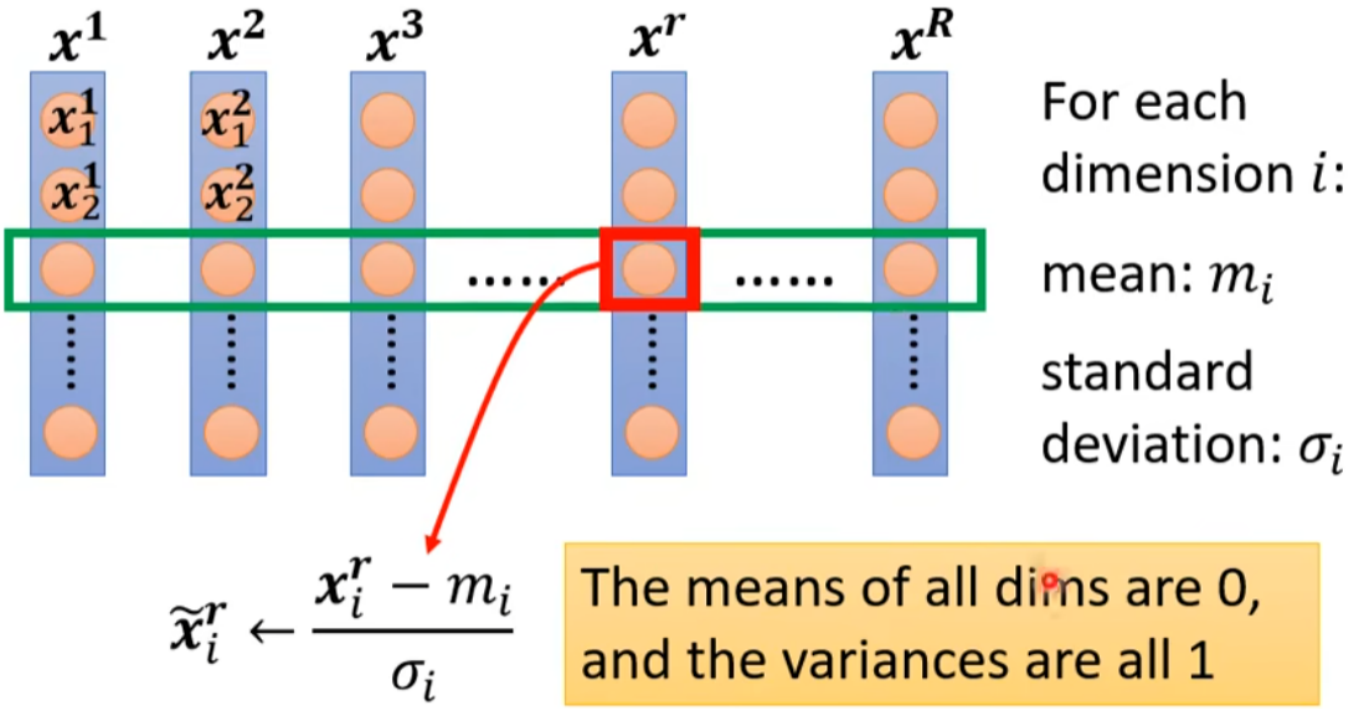
- 当x很小的时候，w的变化对计算值y的变化不大
- 当x很大的时候，即使w有很小的变化，y的值也会变化很大 -> e的变化就很大

能不能让x有相同的数据范围，让梯度下降尽量是一个“圆”：从任意一个点下降都很容易到谷底



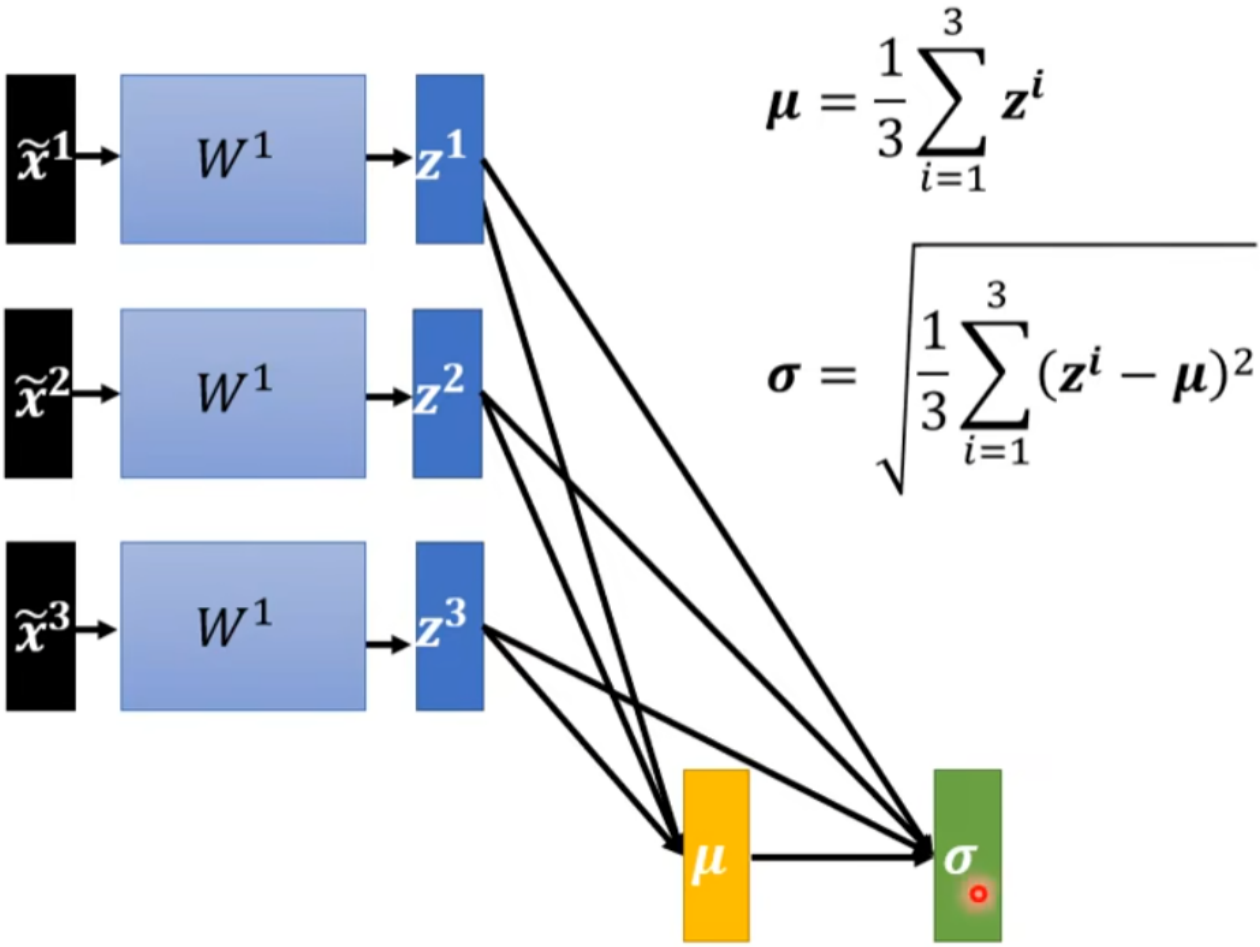
Feature Normalization

（1）直接对输入值x归一化：让不同输入的相同位置做归一化，计算平均值和标准差。

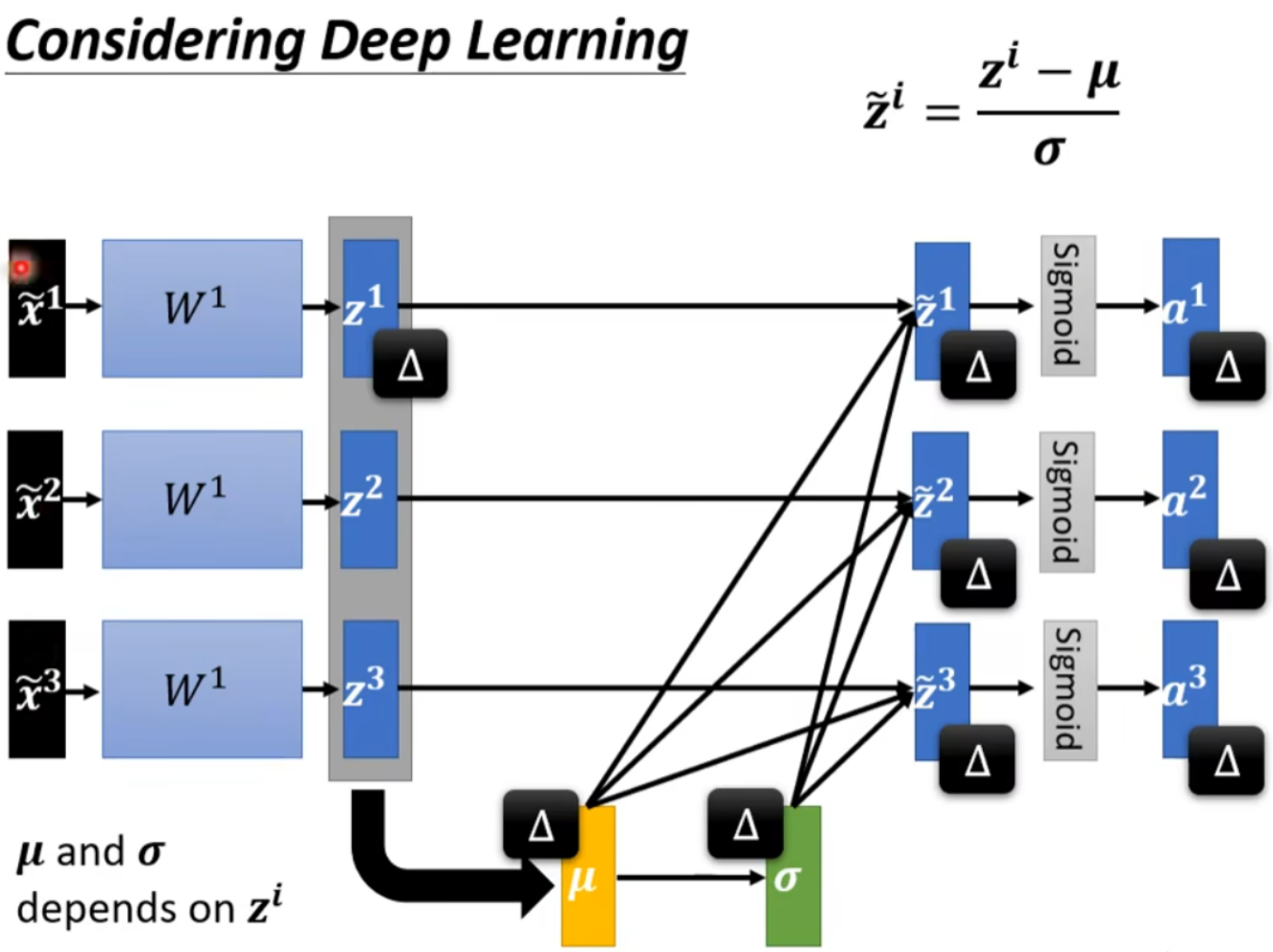


此时，这一行的所有数值在0上下，偏差为1。

（2）对中间的输出结果也需要归一化



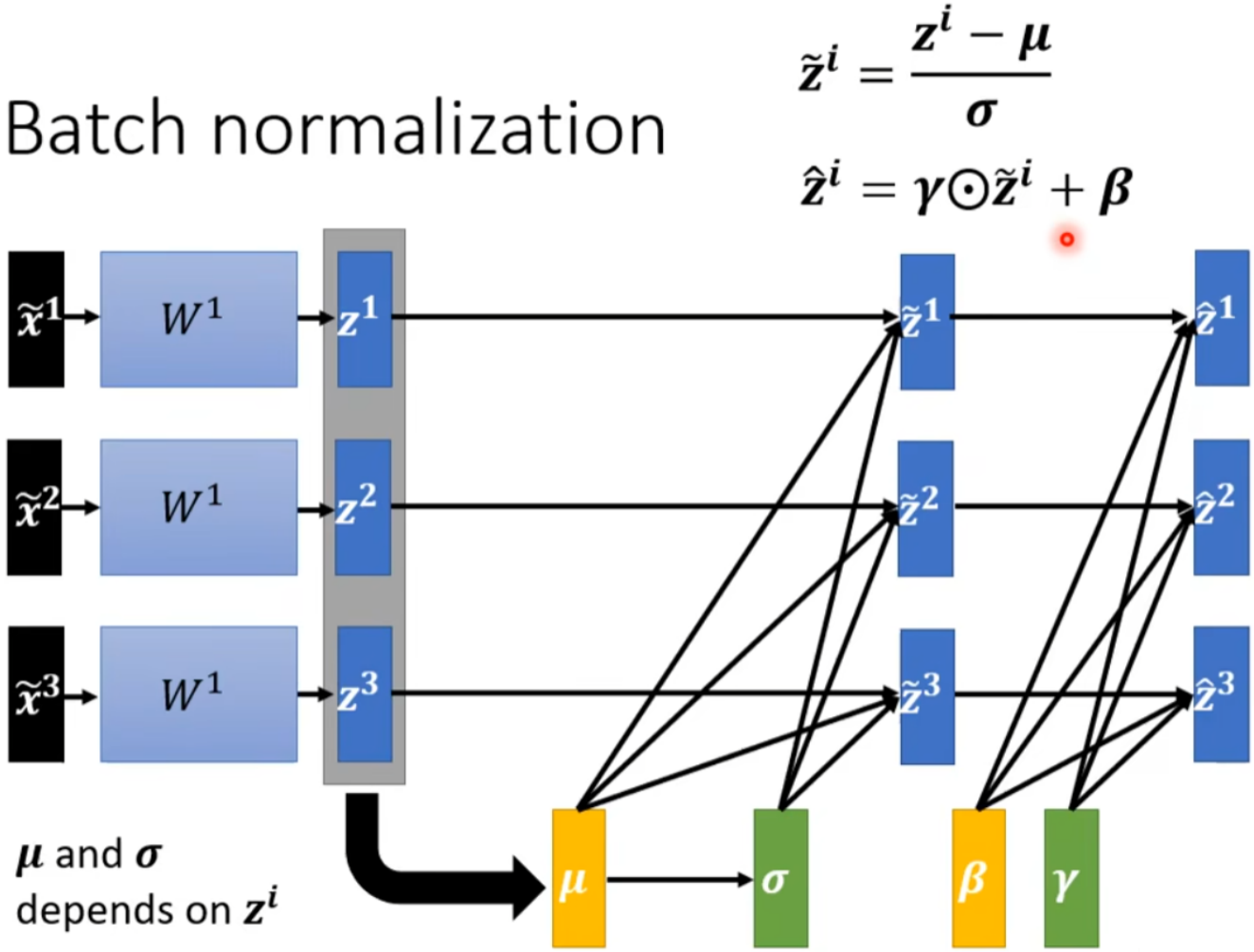
但此时，一个数值的改变会影响μ, σ,从而到所有的输出变化：



但在实际中，由于显存容量，无法算出整个数据集的归一化，因此只是对每个batch计算，即batch normalization

Batch Normalization

但是在实际的训练过程中，需要引入γ, β，让输出也存在一定的差异



同时，在测试集中不一定存在batch，此时过程中之前固定batch数计算参数μ, σ的方法影响比较大。那么就可以利用平均数替代$\bar{\mu} = p\bar{\mu} + (1-p)p^t$