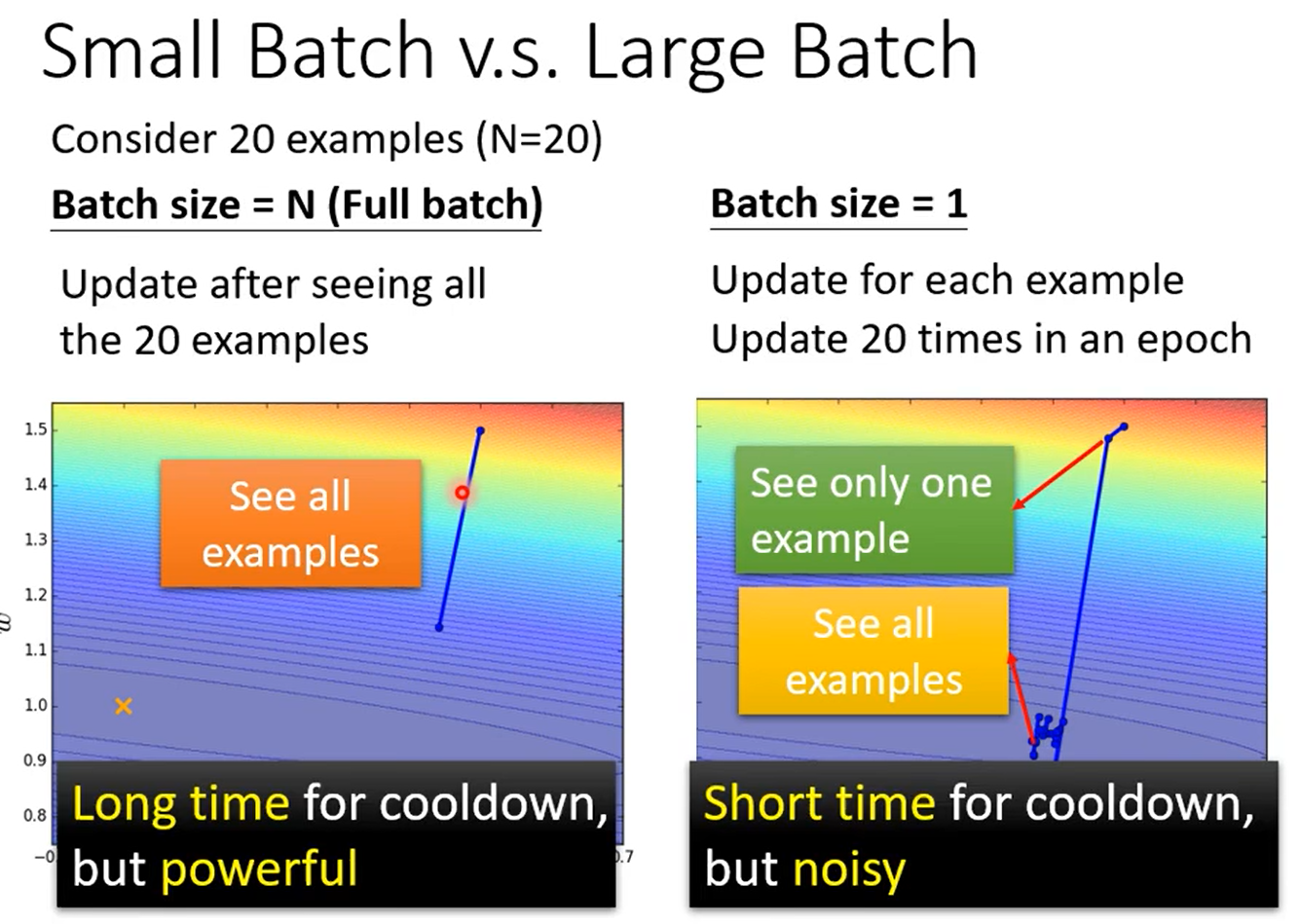
## Batch计算速度的考虑

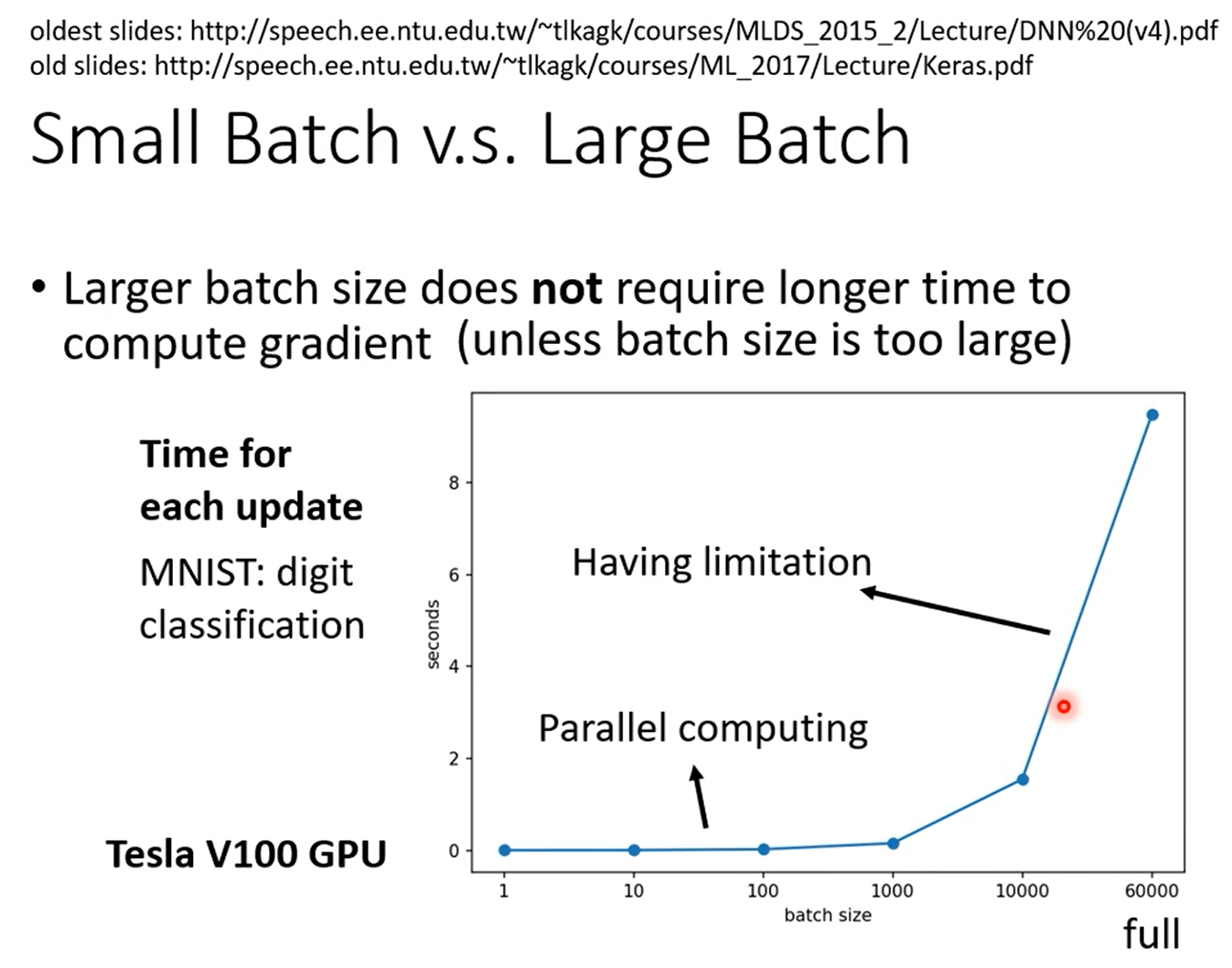
### 不同batch的影响



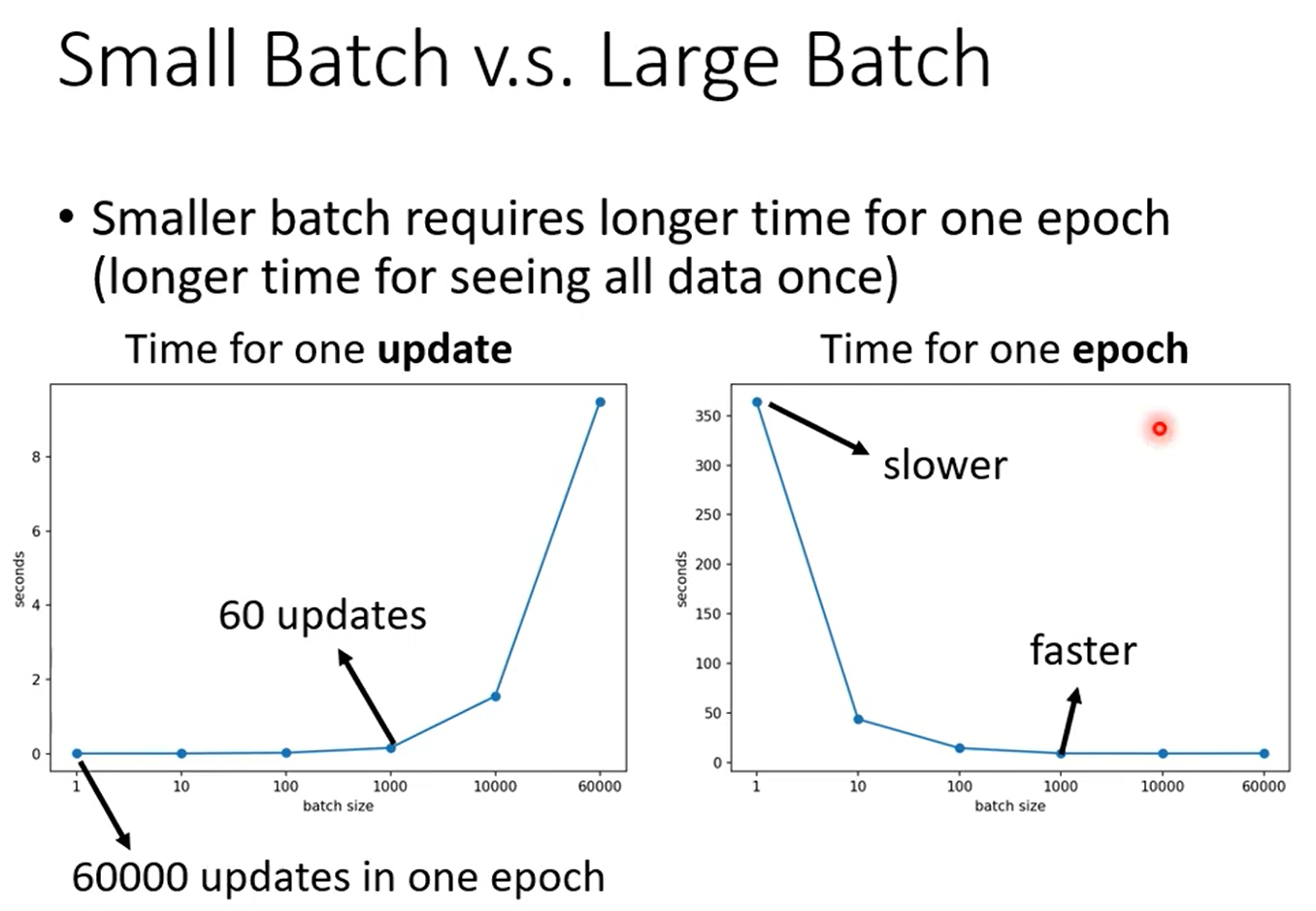
一次用上全部的样本，则速度慢，但是比较稳定

每次只看一个样本，速度很快，但是可能振荡

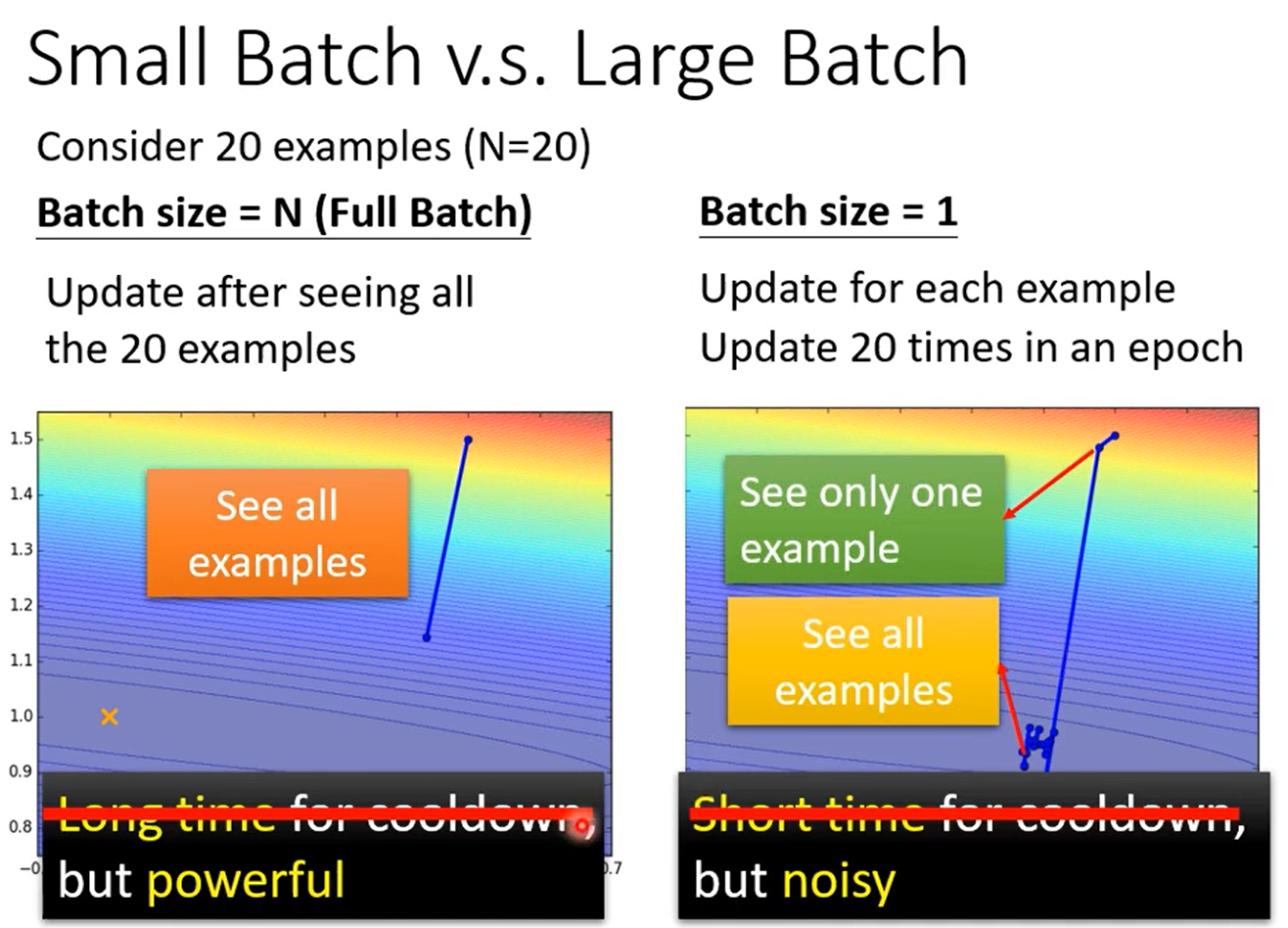
### 1.2 考虑并行计算的时候，其实大批量的收敛速度和小批量也差不多，gpu有并行运算的能力，但是也是有极限的



### 1.3 更新一个epoch，其实小batch所需的时间更长

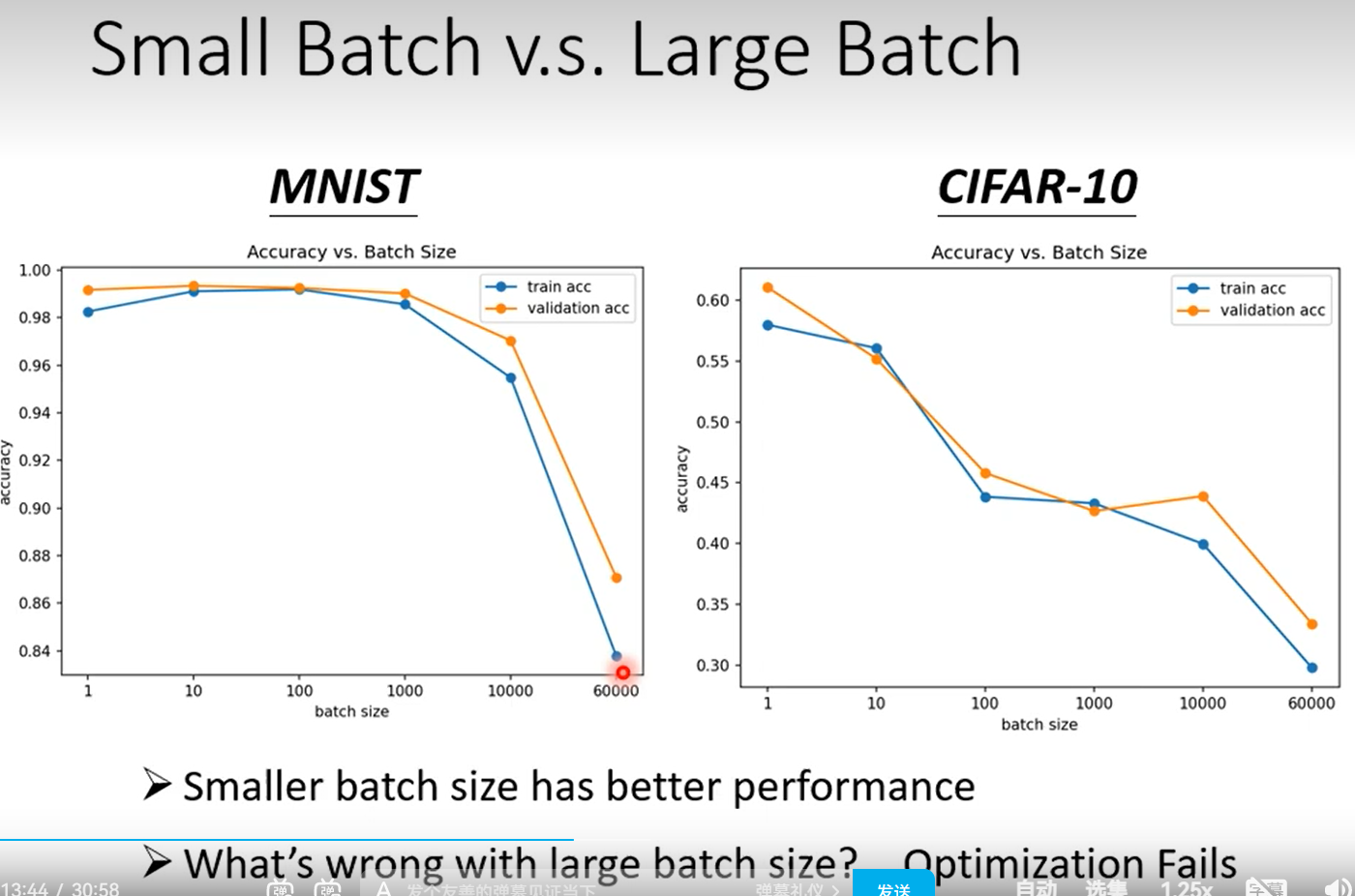


### 1.4 结论就是在gpu下，cooldown的时间不是很关键

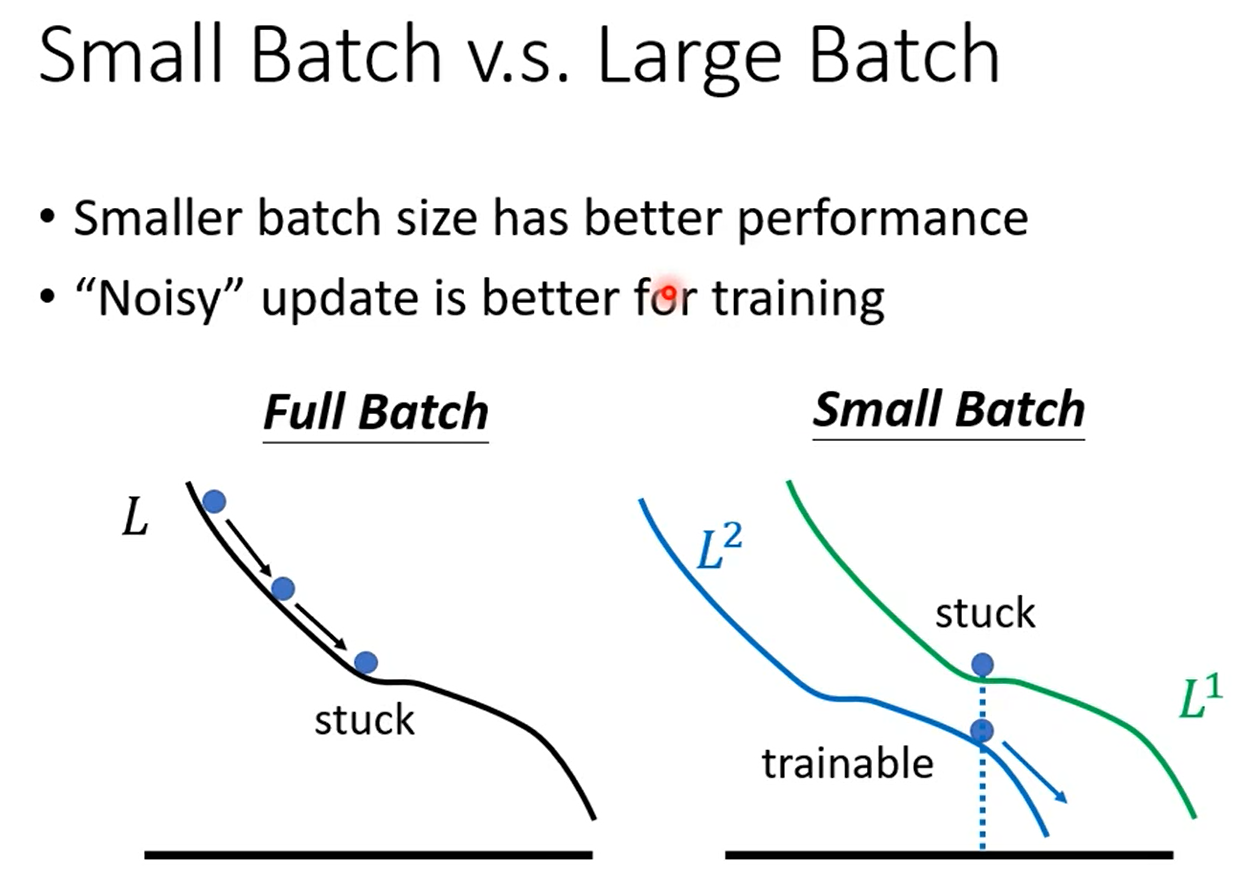


## 2.batch对损失函数的考虑

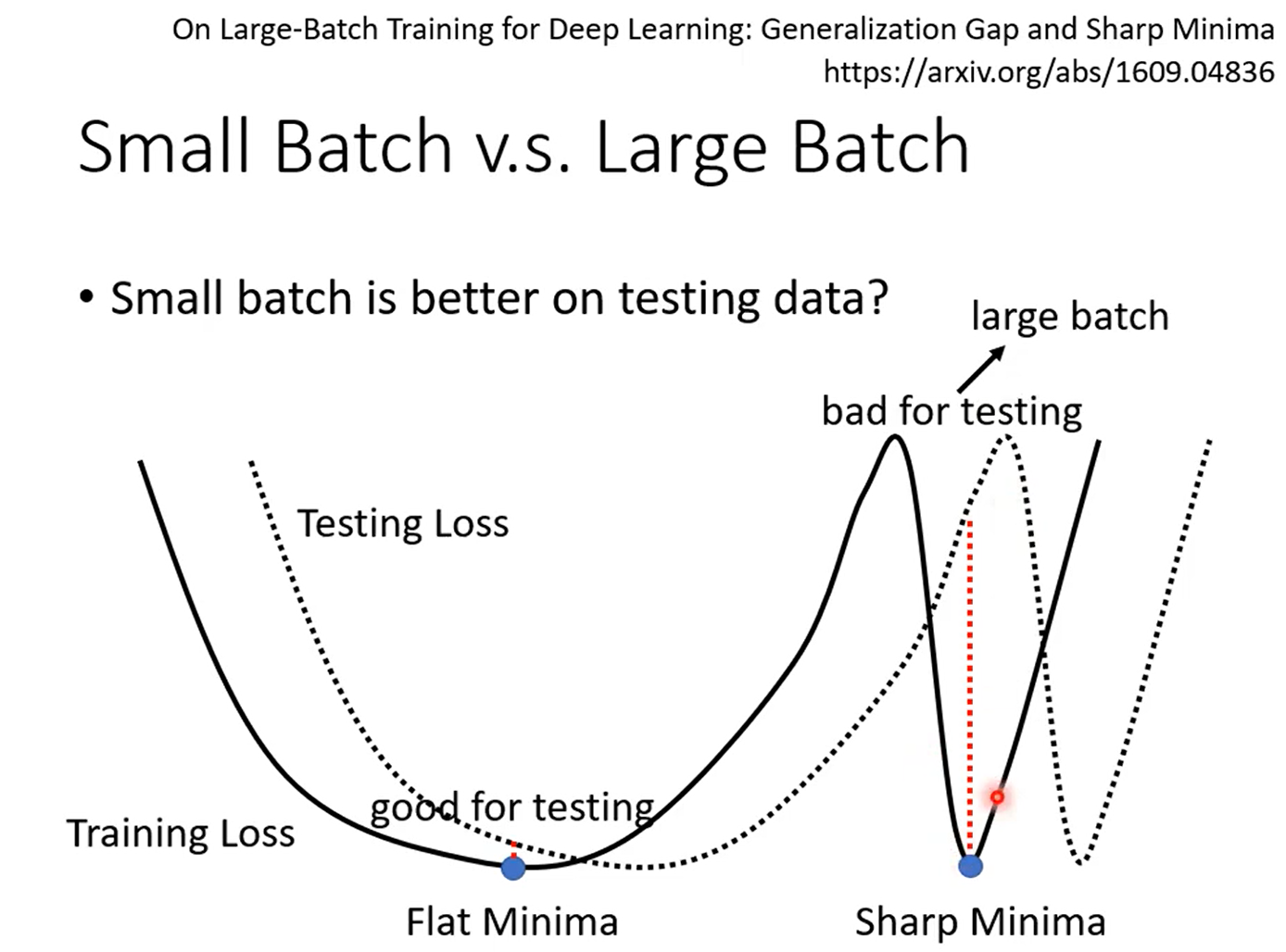
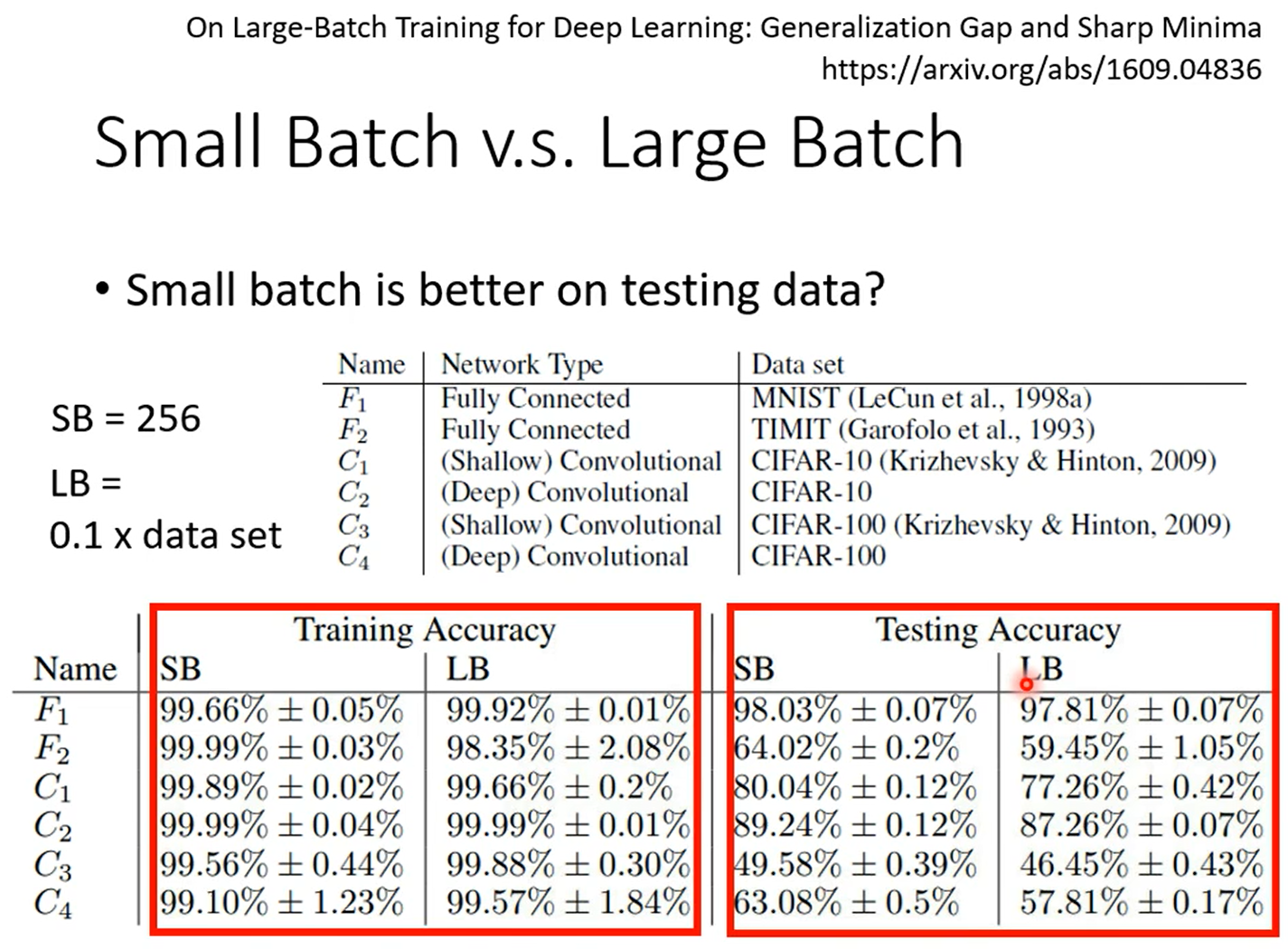
### 2.1 小的batch可能有噪声，对训练的效果更好



### 2.2 为什么有噪声的data（小batch）会有更好的结果？



### 2.3 大的batch可能造成过拟合，因此在测试数据上性能差一点

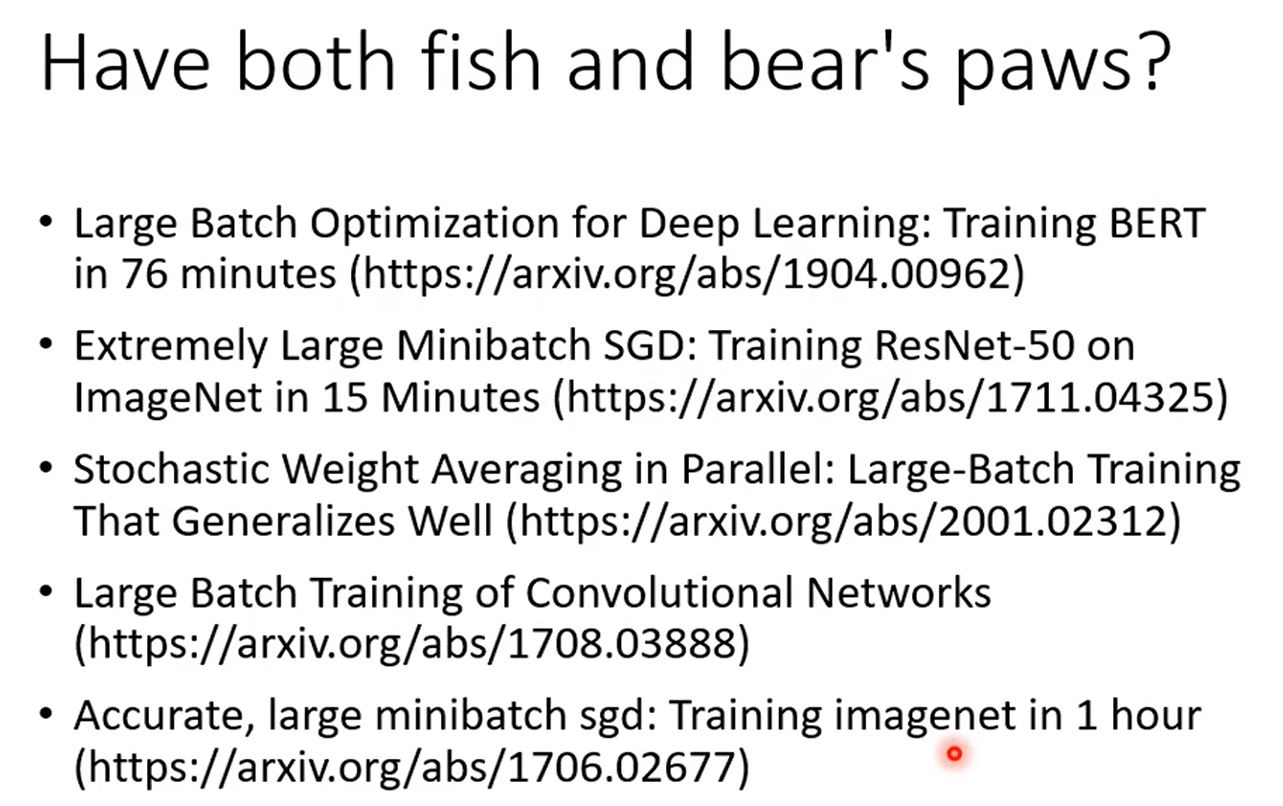


小的batch会走到平坦的极小值点，大的batch会走到很陡峭的极小值点

## 3.batch的总结

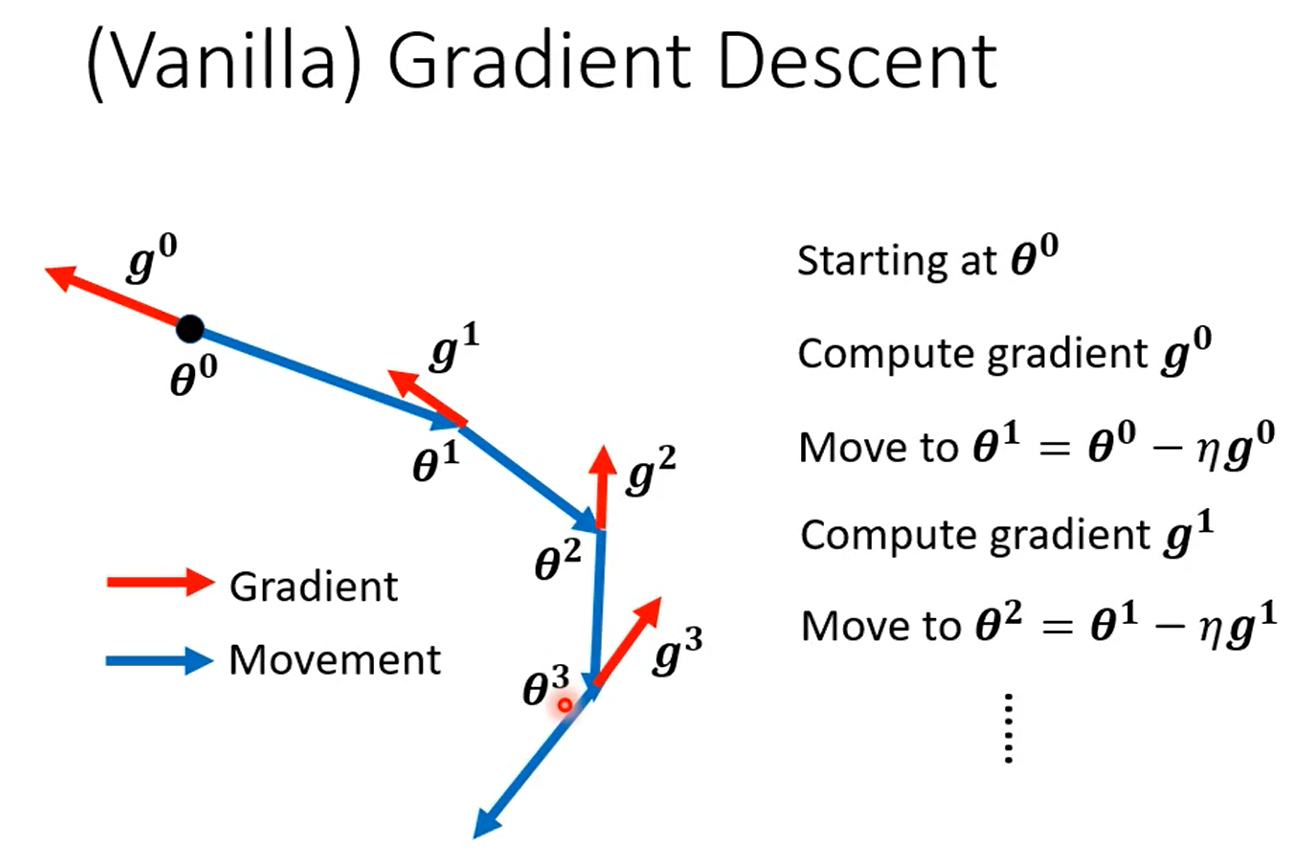


一些文章：



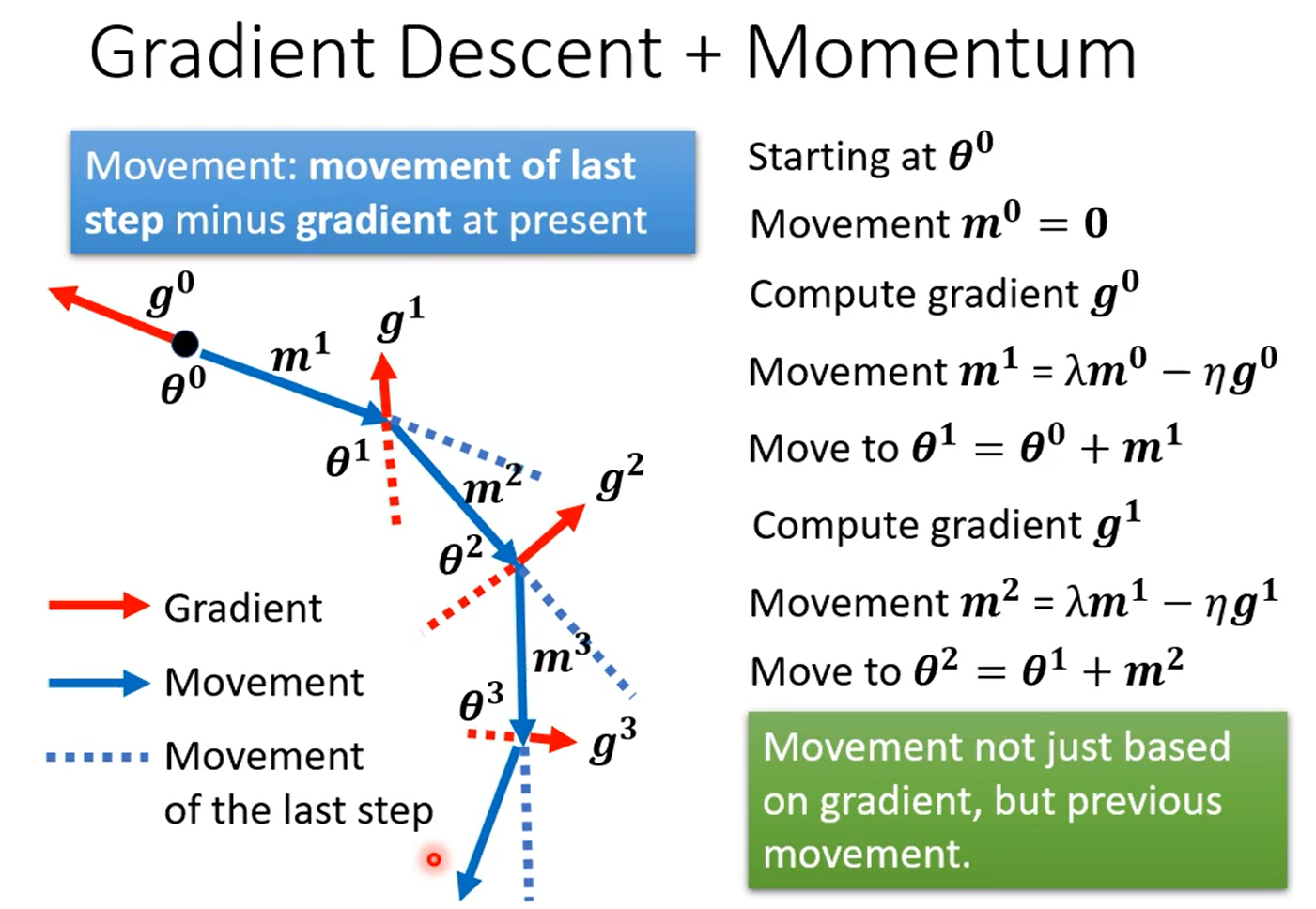
## 4.动量

### 4.1 普通梯度下降

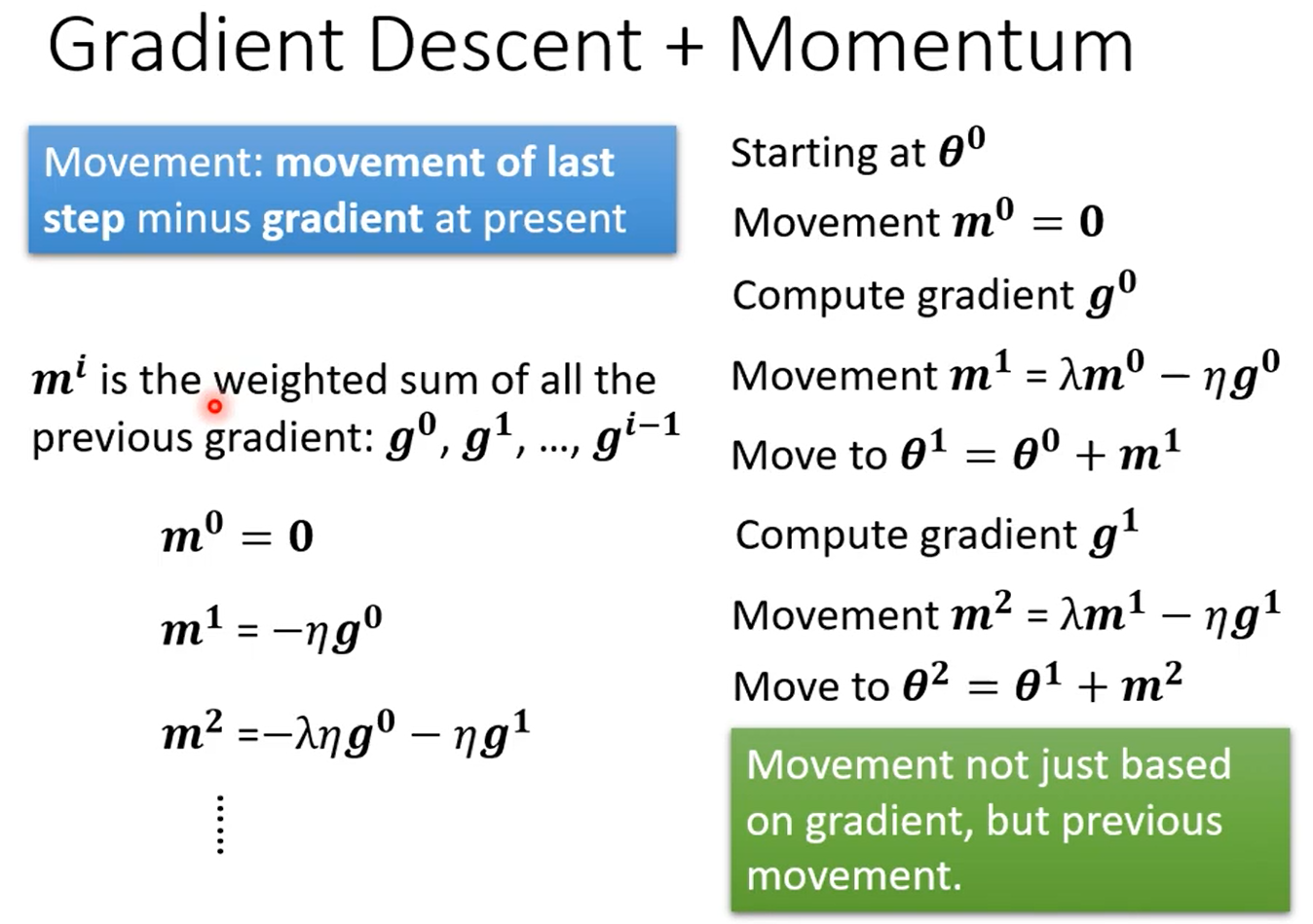


### 4.2 梯度下降+动量

不是向梯度的反方向移动，而是考虑梯度和之前的梯度：



### 4.3 动量就是之前所有梯度的加权和



### 4.4 例子

