# 机器学习笔记

李宏毅教授机器学习(视频地址：https://www.bilibili.com/video/BV1JE411g7XF?p=3&vd\_source=0a27b8c0284e7cd3b72bc0c89478f2d7)

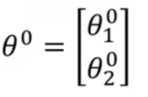
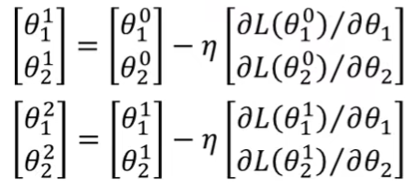
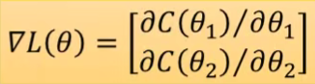
# 第二课：Gradient Descent

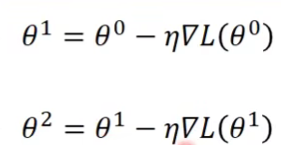
# Tip1：

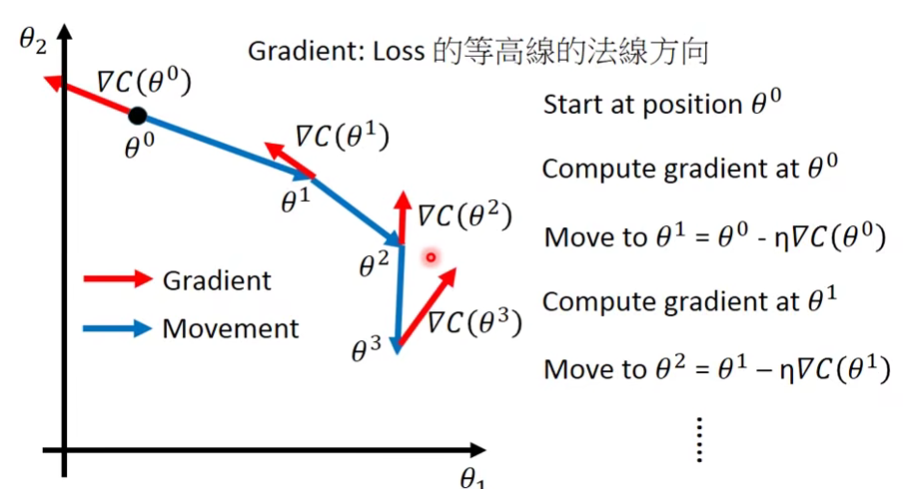
在机器学习的第三步中，我们需要找到一个最好的function，而找这个最好的function是一个最优化的问题。

我们定义了一个**loss function：L**，它是function的function，用来代表一个function的好坏程度。

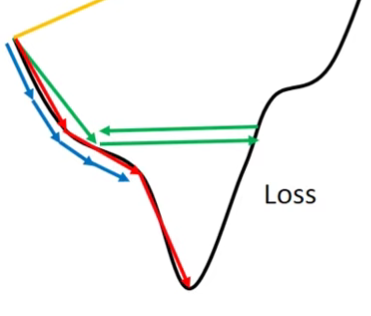
假设θ是我们的参数，它有两个取值和，η是我们的learning rate，

我们的初始值随机选取一个，那么之后的取值就表示为，其中式子后面部分是loss function对θ的偏导数，我们用Gradient(梯度)来表示后面的部分，写作

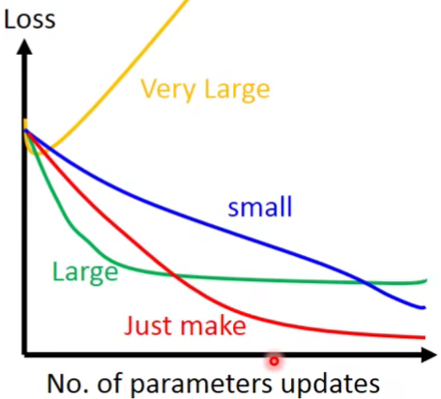
那么我们前面的式子就可以写成，整个求θ最优值的过程可以表示成下图的形式，



我们设置η的时候要比较小心，因为η会影响到我们每一步的跨度，如果跨度过小，那么我们寻找最优值就需要花过多的时间；如果跨度过大，就有可能跳过最优值的那个“坑”，如下图所示，

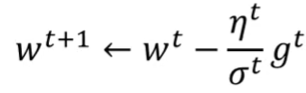
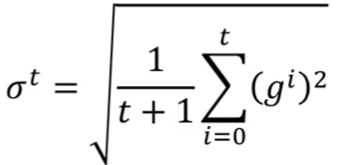


在维度高的时候我们就没法用这种图示来表示这个寻找最优值的过程，所以我们需要用一个来描述参数修改的图了，它的横坐标是参数的修改，纵坐标是对应的loss，如下图所示

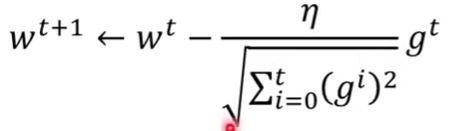


我们想要让机器自动地调整learning rate（η），我们知道，当我们越调整η的时候它下一次调整的量就应该相对应地减少，所以我们可以让η是一个与时间t有关的函数，。

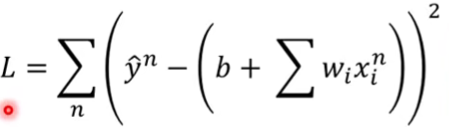
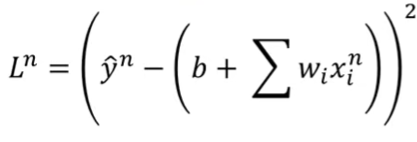
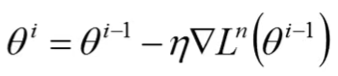
但是对于不同的参数我们的修改量应当是不同的，所以我们就有另外一种自动调整η的方法，叫做Adagrad，它的做法如下图

，其中，而为过去所有微分值的开平方平均值，即

整理之后我们就可以得到参数修改的形式为

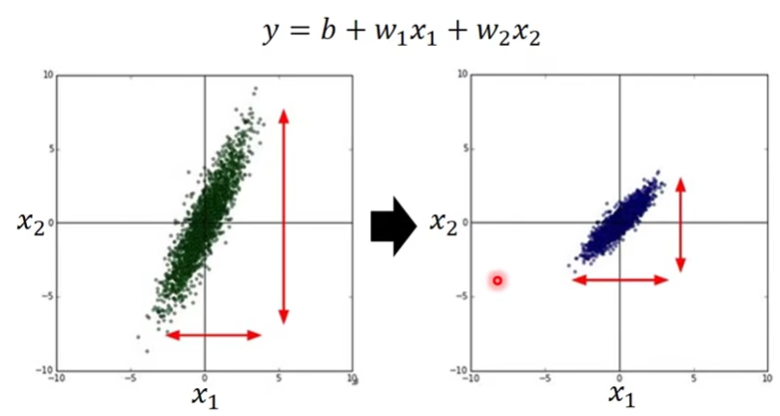


# Tip2：Stochastic Gradient Descent（随机梯度下降法）

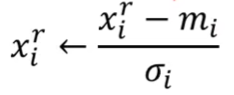
在做回归的时候，我们的loss function的形式是，它需要考虑所有的样例数据，而Stochastic Gradient Descent则是每次只拿一个样例数据来计算loss，表示为，然后参数的修改就表示为，它的优缺点也很明显，优点是可以更快更频繁地修改我们的参数，而缺点就是它的修改比较不平滑，比较散乱。

# Tip3：Feature Scaling（特征缩放）

对于不同的参数，他们的修改比例可能是不同的，而我们想要让他们的变化比例相同，就需要进行缩放，如下图所示



而这个参数的缩放就按照如下所示的方法进行，

，其中表示第i个参数的均值，表示第i个参数的标准差。