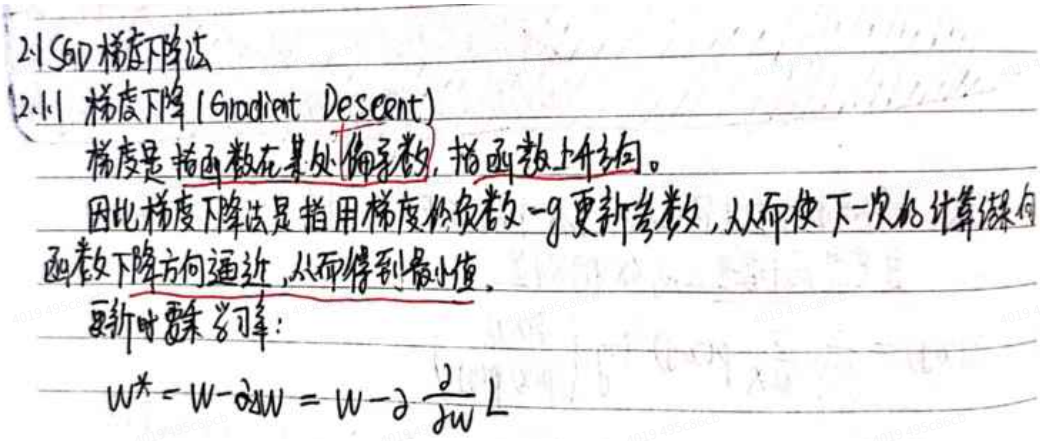
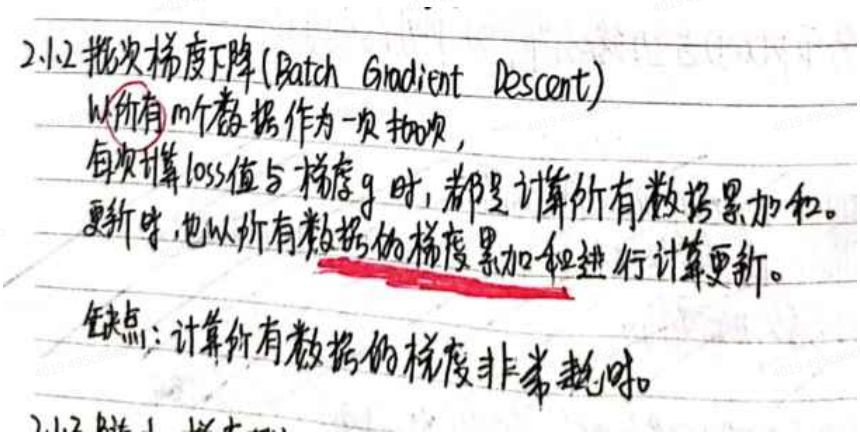
5.梯度下降法的改进过程（SGD->Adam)

1.梯度下降法

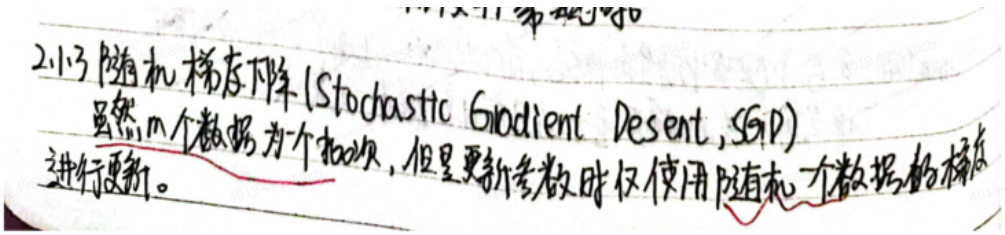
1.1梯度下降

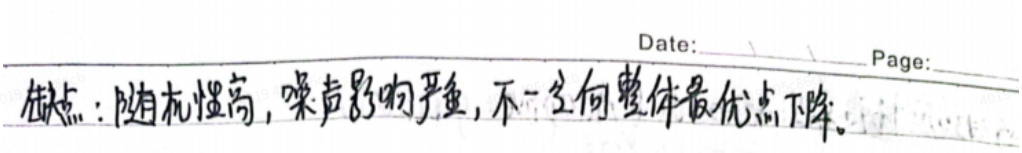


1.2批次梯度下降



1.3随机梯度下降





1.4小批次梯度下降

2.1.4 小批次梯度下降 (Mini-batch GD (MBGD)

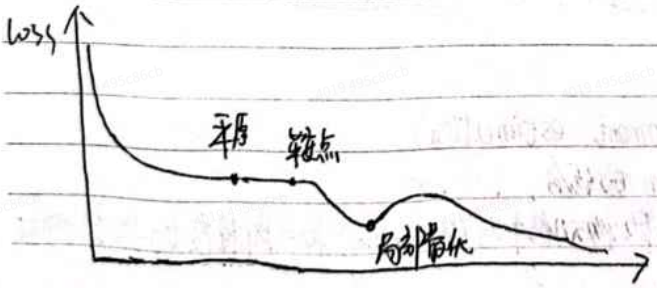
把所有样本分为n个batch (一般是随机的), 每次计算损失的梯度使用一个batch的数据进行计算, 并更新参数, 从而避免了唯一随机性和全局计算的耗时性。

优点: 得到的梯度下降方向是局部最优的, 整体速度快.

(一般说的SGD就是指 Mini-batch GD)

2.动量梯度下降

梯度下降法可能会停在平坦, 鞍点, 局部最优点 (这三处梯度均为0), 因此带动量的梯度下降法能依靠前的梯度值, 冲过平坦, 鞍点和局部最优点, 提高泛化性。



$W = W - \alpha V_{dw}$

V: 历史参考的更新值大小, 相当于把现在与过去的梯度的一个加权。

可以视为当前一个动量。

3.自适应梯度下降

2.3 自适应梯度算法 Adagard (Adaptive gradient)

针对不同的变量, 提供不同的学习率。

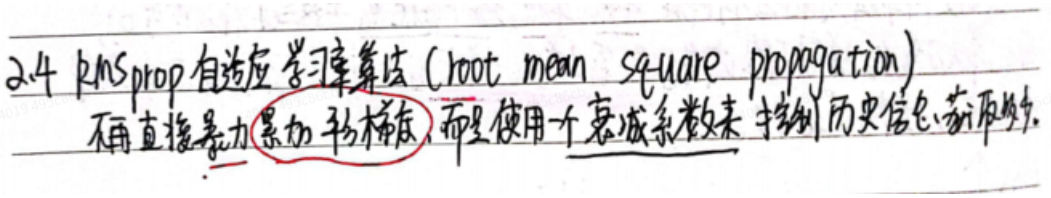
解决方法:

为每一参数建立历史累计梯度值, 利用历史累计梯度作为分母, 从而使各个参数在训练后期被除于不同的除数, 得到自适应参数值.

为什么要不同学习率?

当一些变量被优化到最优点时, 另外一些并没有, 使用统一的学习率就会影响优化过程, 大大减小都不好。

4.RMSprop自适应学习率算法



5.Adam（Adaptive moment estimation）