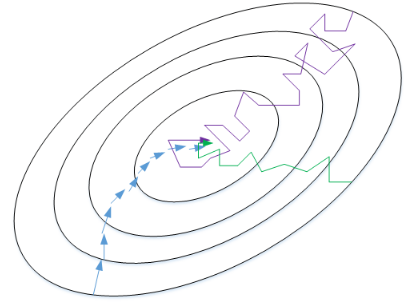
深度学习优化算法

http://blog.csdn.net/red\_stone1/article/details/78348753

Mini-batch gradient，相比batch gradient，能够提高运算速度，兼顾收敛速度，是一种平衡选择

Stachastic-batch，每次训练只有一个样本，虽然运算速度最快（样本处理时间短），但是收敛速度反而比mini-batch底，因为梯度的振荡太大。样本太少，意味着单次训练获取梯度的局部性太强。

每次epoch，最好是将总体训练数据重新打乱、重新分成T组mini-batches，这样有利于训练出最佳的神经网络模型

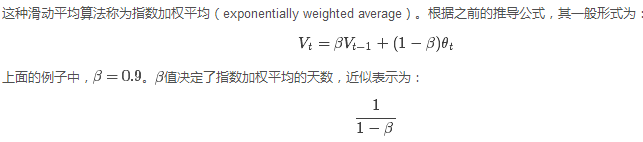


蓝色为batch，运算慢，但是振荡小。

绿色为mini-batch，运算速度和收敛速度均衡。

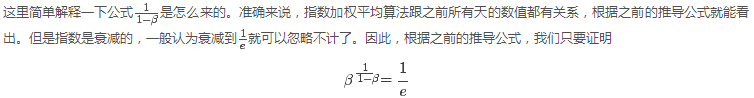
紫色为stachastic-batch，运算速度快，收敛折腾多

指数加权平均（Exponentially weighted averages）



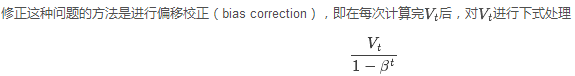
Beta越大，当前数值对于整体平均的影响越小，整体曲线越平滑。反之……

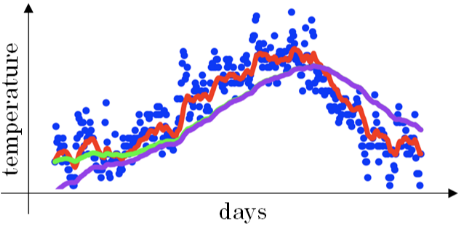
指数加权平均公式的得出过程。关键在于确定多少天之前的数据会衰减到忽略不计



理解指数加权平均。一段时间的数据乘以指数衰减的权重，时间越久，影响越小

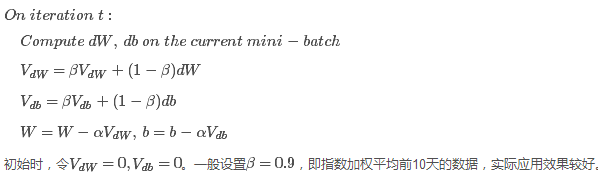
初始阶段，指数加权平均的偏差修正



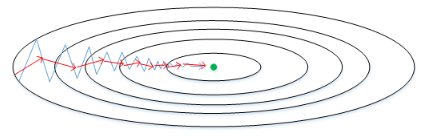


紫色为修正前的，绿色为修正后的

动量梯度下降算法，momentum gradient



理解动量梯度算法。有多种说法，但本质都是某次迭代的梯度变化值，不仅与当次计算的梯度有关，还与之前的梯度有关。梯度是渐变而非瞬变的，保证了梯度下降的平稳和准确，减少了震荡



蓝色为普通梯度下降，红色为动量梯度下降