指数加权移动平均

（Exponential Weighted Moving Average）

1. EMA 简介：

EMA：是以指数式递减加权的移动平均，各数值的胶圈影响力随时间呈指数式递减，时间越靠近当前时刻的数据加权影响力（权重）越大。

发展过程：

算术平均（权重相等）——>加权平均（权重不等）——>移动平均（指定时间段，对时间序列数据进行移动计算平均值）——>批量归一化（BN）及各种优化算法的基础

2. 推导过程及理解

给定超参数$0 \le \gamma < 1$，当前时间步 t 的变量y_t是上一时间步 t-1 的变量y_{t-1}和当前时间步另一变量x_t的线性组合：

$$y_t = \gamma y_{t-1} + (1-\gamma)x_t.$$

对y_t展开：

$$\begin{aligned} y_t &= (1-\gamma)x_t + \gamma y_{t-1} \\ &= (1-\gamma)x_t + (1-\gamma)\cdot\gamma x_{t-1} + \gamma^2 y_{t-2} \\ &= (1-\gamma)x_t + (1-\gamma)\cdot\gamma x_{t-1} + (1-\gamma)\cdot\gamma^2 x_{t-2} + \gamma^3 y_{t-3} \\ &\quad \cdots \end{aligned}$$

令$n = 1/(1-\gamma)$，则：

$$(1-1/n)^n = \gamma^{1/(1-\gamma)}$$

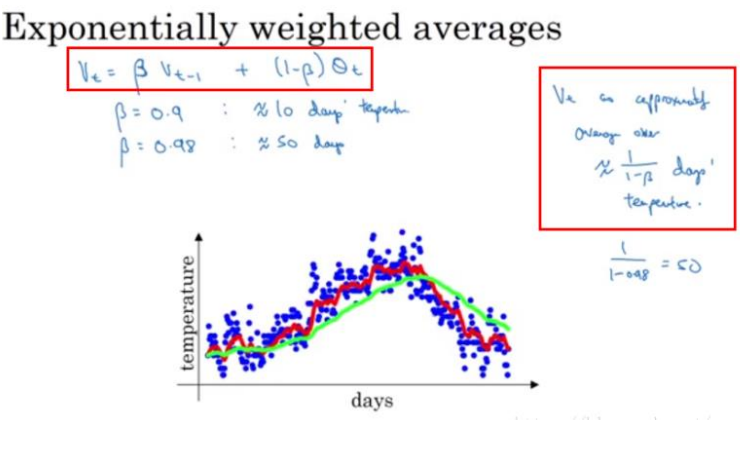
$$\because \lim_{n\to\infty}(1-\frac{1}{n})^n = \exp(-1)$$

$$\therefore \gamma \to 1 \quad \gamma^{1/(1-\gamma)} = \exp(-1)$$

若把 exp(-1)当作一个比较小的数，则在近似中忽略所有含$\gamma^{1/(1-\gamma)}$和比$\gamma^{1/(1-\gamma)}$高阶的系数的项。

所以在实际过程中常常将y_t看作是对最近$1/(1-\gamma)$个时间步的x_t值的加权平均，并且当前时间步 t 越接近的x_t值获得的权重越大（越接近 1）。

举例：（Andrew Ng deep learning coursea）



例举对于温度的的预测：

a) $v_t = \beta v_{t-1} + (1-\beta)\theta_t$ ，公式中的θ_t为 t 时刻的实际温度，系数β表示加权下降的

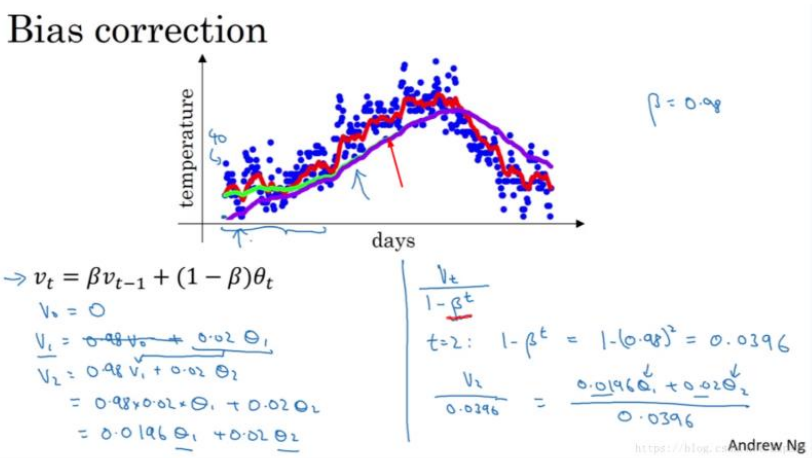
快慢，值越小权重下降的越快，v_t为 t 时刻的 EMA 值。

b) 当$v_0 = 0$时，可得：$v_t = (1-\beta)(\theta_t + \beta\theta_{t-1} + \beta^2\theta_{t-2} + \ldots + \beta^{t-1}\theta_1)$，从公式中可以看出：每天温度（θ）的权重以指数等比形式缩小，时间越靠近当前时刻的数据加权权重越大。

c) 在优化算法中，一般取$\beta >= 0.9$，而$1 + \beta + \beta^2 + \ldots + \beta^{t-1} = \frac{1-\beta^t}{1-\beta}$，所以当 t 足够大时，$\beta^t \approx 0$，此时便是严格意义上的指数加权移动平均。

d) 关于移动平均的解释：取$\beta >= 0.9$，此时有$\beta^{\frac{1}{1-\beta}} \approx \frac{1}{e}$，$N = \frac{1}{1-\beta}$天后，曲线高度下降到了约原来的三分之一，由于时间越往前推移，权重θ越来越小，所以只考虑最近的$N = \frac{1}{1-\beta}$天的数据计算当前时刻的 EMA，也就是移动平均的来源。

EMA 偏差修正：



在$\beta = 0.98$时，理想情况应为绿色曲线，而实际是紫色曲线，起点比真实值低，不能很好地预测起始位置的温度，此问题称为：冷启动问题。是由于$v_0 = 0$造成的。
解决方案：将所有时刻的 EMA 除以$1-\beta^t$后作为修正后的 EMA。当 t 很小时，可以在起始阶段的预测更加准确，当 t 很大时，偏差修正几乎不起作用，对原来试自几乎没有影响。注意：计算 t 时刻修正后的 EMA 时，使用 t-1 时刻修正前的 EMA。

3. EMA 优点

a) 占用内存较少：计算指数加权平均数只占用单行数字的存储和内存，然后把最新数据带入公式，不断覆盖即可。

b) 移动平均线能较好地反应时间序列的变化趋势，权重的大小不同起到的作用不同，时间交久远的变量值权重较低，其影响力也相对较低，时间较近的变量值权重较高，影响力也相对较高。

4. 动量法优化

在梯度下降中，给定学习率，梯度下降迭代自变量会使自变量在不同方向下降速度不同，此时需要较小的学习率避免在某方向上超过目标函数的最优解，而同时，会导致自变量在其他方向上朝最优解移动速度减慢。
为解决上述问题，使用动量法，同小批量随机梯度下降（mini batch）中时间步 t 的小批量随机梯度g_t的定义，设时间步 t 的自变量为x_t，学习率为η_t。 在时间步为 0，动量

法创建速度变量v_0，并将其元素初始化为 0。在时间步 t>0，动量法对每步迭代做以下操作：

$$\begin{aligned}\boldsymbol{v}_t &\leftarrow \gamma \boldsymbol{v}_{t-1} + \eta_t \boldsymbol{g}_t,\\ \boldsymbol{x}_t &\leftarrow \boldsymbol{x}_{t-1} - \boldsymbol{v}_t,\end{aligned}$$

其中动量超参数γ满足：$0 \le \gamma < 1$，当$\gamma = 0$时，动量法等价于小批量随机梯度下降。

结合 EMA：

对动量法中的速度变量进行变形：

$$\boldsymbol{v}_t \leftarrow \gamma \boldsymbol{v}_{t-1} + (1-\gamma)(\frac{\eta_t}{1-\gamma}\boldsymbol{g}_t)$$

由指数加权移动平均的形式可知：速度变量v_t实际上是对序列$\{\eta_{t-i}\boldsymbol{g}_{t-i}/(1-\gamma): i = 0, \dots, 1/(1-\gamma)-1\}$做了指数加权移动平均，相当于小批量随机梯度下降。所以，在动量法中，自变量在各个方向上的移动幅度不仅取决当前梯度，还取决于过去的各个梯度在各个方向上是否一致。

在迭代后期，由于随机噪声问题，经常会在收敛值附近震荡，动量法会起到减速作用，增加稳定性。