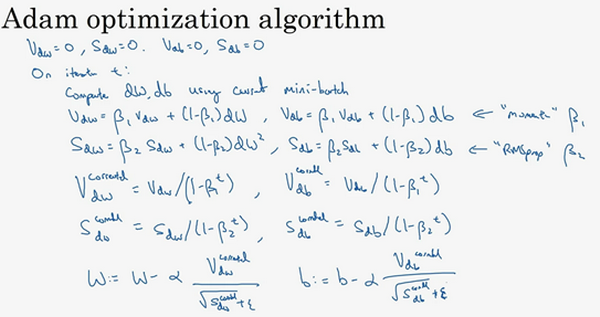
### 2.8 Adam 优化算法(Adam optimization algorithm)

在深度学习的历史上，包括许多知名研究者在内，提出了优化算法，并很好地解决了一些问题，但随后这些优化算法被指出并不能一般化，并不适用于多种神经网络，时间久了，深度学习圈子里的人开始多少有些质疑全新的优化算法，很多人都觉得动量（**Momentum**）梯度下降法很好用，很难再想出更好的优化算法。所以**RMSprop**以及**Adam**优化算法（**Adam**优化算法也是本视频的内容），就是少有的经受住人们考验的两种算法，已被证明适用于不同的深度学习结构，这个算法我会毫不犹豫地推荐给你，因为很多人都试过，并且用它很好地解决了许多问题。

**Adam**优化算法基本上就是将**Momentum**和**RMSprop**结合在一起，那么来看看如何使用**Adam**算法。



使用**Adam**算法，首先你要初始化，，，，，在第次迭代中，你要计算微分，用当前的**mini-batch**计算，，一般你会用**mini-batch**梯度下降法。接下来计算**Momentum**指数加权平均数，所以（使用，这样就不会跟超参数混淆，因为后面**RMSprop**要用到），使用**Momentum**时我们肯定会用这个公式，但现在不叫它，而叫它。同样。

接着你用**RMSprop**进行更新，即用不同的超参数，，再说一次，这里是对整个微分进行平方处理，。

相当于**Momentum**更新了超参数，**RMSprop**更新了超参数。一般使用**Adam**算法的时候，要计算偏差修正，，修正也就是在偏差修正之后，

，

同样，

也使用偏差修正，也就是，。

最后更新权重，所以更新后是（如果你只是用**Momentum**，使用或者修正后的，但现在我们加入了**RMSprop**的部分，所以我们要除以修正后的平方根加上）。

根据类似的公式更新值，。

所以**Adam**算法结合了**Momentum**和**RMSprop**梯度下降法，并且是一种极其常用的学习算法，被证明能有效适用于不同神经网络，适用于广泛的结构。

图片包含 文字, 白板

描述已自动生成

本算法中有很多超参数，超参数学习率很重要，也经常需要调试，你可以尝试一系列值，然后看哪个有效。常用的缺省值为0.9，这是dW的移动平均数，也就是的加权平均数，这是**Momentum**涉及的项。至于超参数，**Adam**论文作者，也就是**Adam**算法的发明者，推荐使用0.999，这是在计算以及的移动加权平均值，关于的选择其实没那么重要，**Adam**论文的作者建议为，但你并不需要设置它，因为它并不会影响算法表现。但是在使用**Adam**的时候，人们往往使用缺省值即可，，和都是如此，我觉得没人会去调整，然后尝试不同的值，看看哪个效果最好。你也可以调整和，但我认识的业内人士很少这么干。

为什么这个算法叫做**Adam**？**Adam**代表的是**Adaptive Moment Estimation**，用于计算这个微分（），叫做第一矩，用来计算平方数的指数加权平均数（），叫做第二矩，所以**Adam**的名字由此而来，但是大家都简称**Adam**权威算法。