

¹**RMSProp** 算法 (Hinton, 2012) ¹修改 AdaGrad 以在非凸设定下效果更好，改变梯度积累为指数加权的移动均值。²AdaGrad 旨在应用于凸问题时快速收敛。当应用于非凸函数训练神经网络时，学习轨迹可能穿过了很多不同的结构，最终到达一个局部是凸碗的区域。³AdaGrad 根据平方梯度的整个历史收缩学习速率，可能使得学习速率在达到这样的凸结构前就变得太小了。⁴RMSProp 使用指数衰减平均以丢弃遥远过去的历史，使其能够在找到凸碗状结构后快速收敛，它就像一个初始化于



