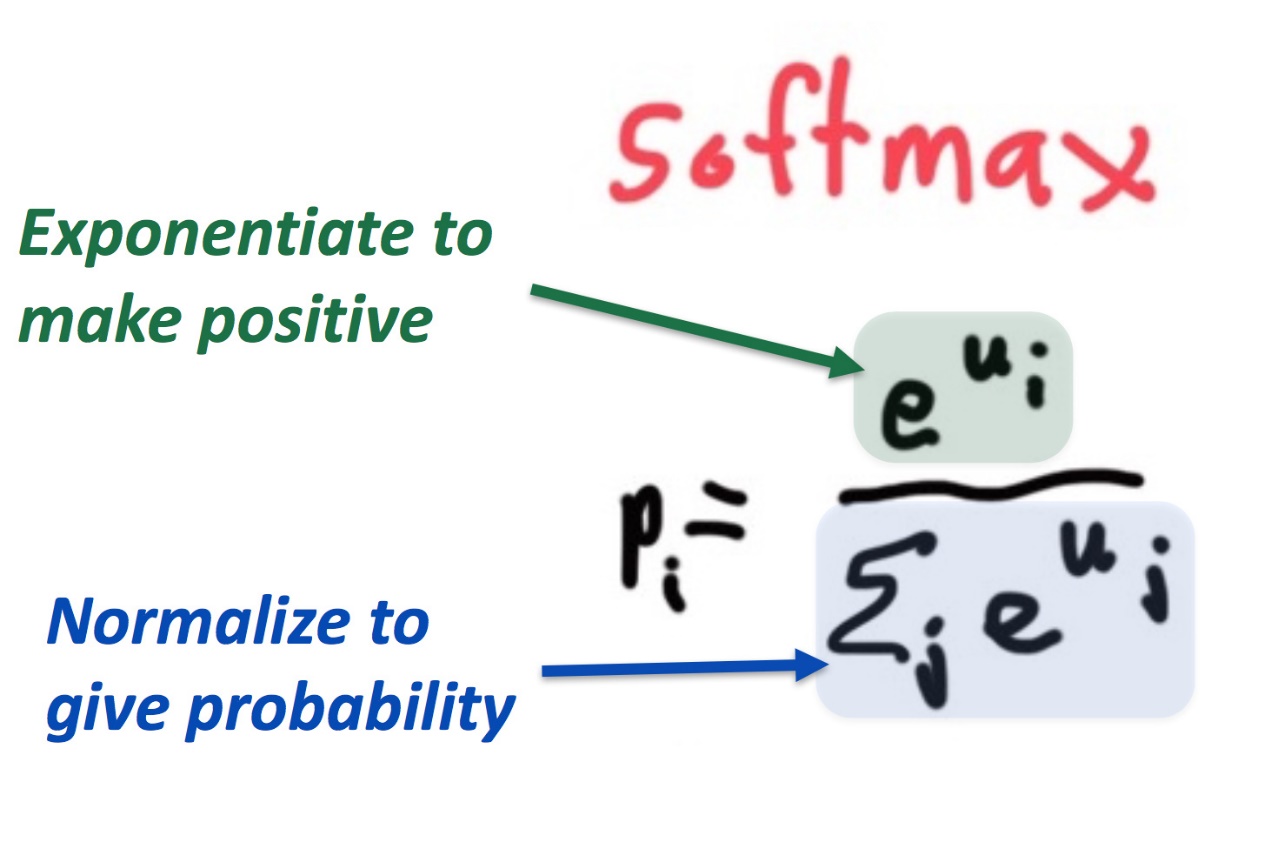
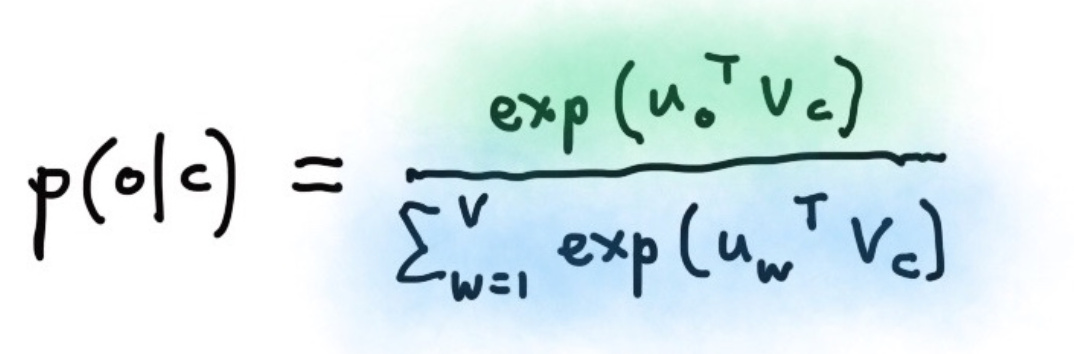
**Softmax function：从实数空间到概率分布的标准映射方法**



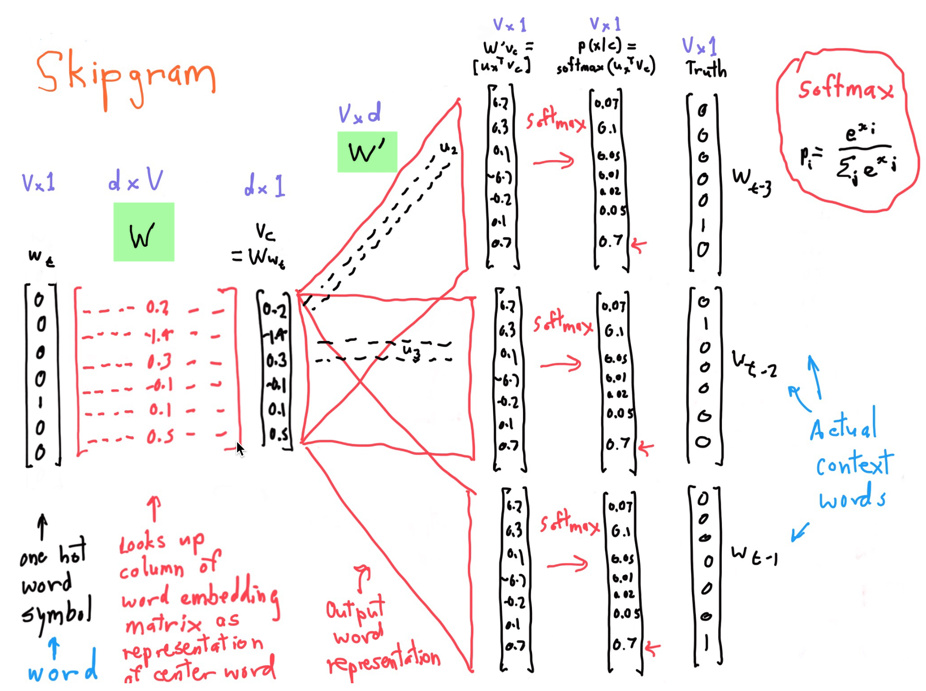
指数函数可以把实数映射成正数，然后归一化得到概率。

softmax之所叫softmax，是因为指数函数会导致较大的数变得更大，小数变得微不足道；这种选择作用类似于max函数。

预测到的某个上下文条件概率p(wt+j|wt)p(wt+j|wt)可由softmax得到：

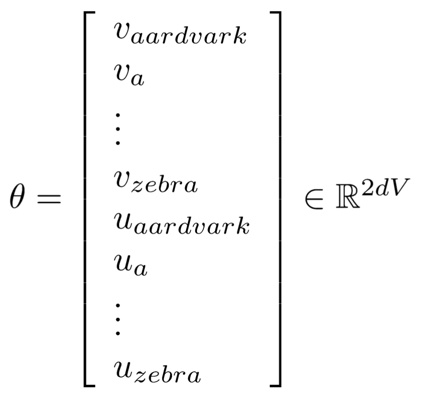


o是输出的上下文词语中的确切某一个，c是中间的词语。u是对应的上下文词向量，v是词向量。

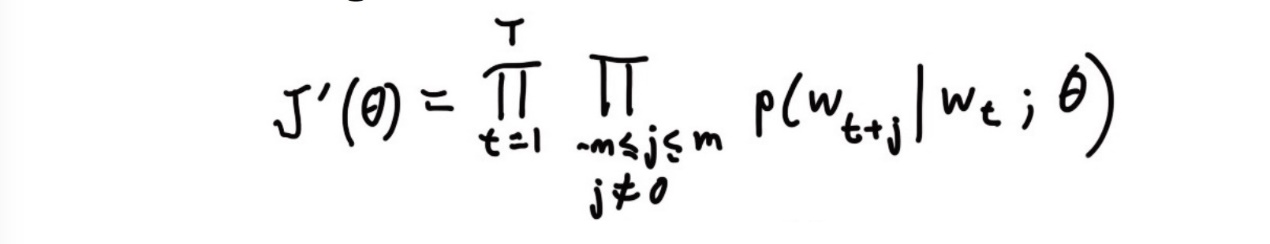


别看这张图有点乱，但其实条理很清晰，基本一图流地说明了问题。从左到右是one-hot向量，乘以center word的W于是找到词向量，乘以另一个context word的矩阵W'得到对每个词语的“相似度”，对相似度取softmax得到概率，与答案对比计算损失。

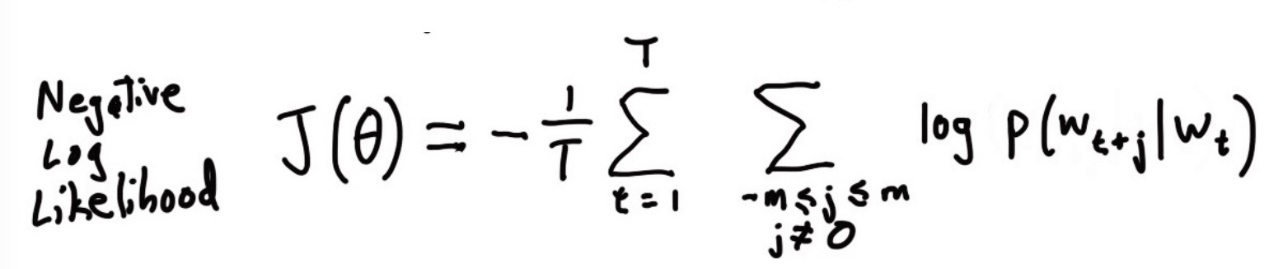
优化的目标是W和W’，两个矩阵中都含有V个词向量，，也就是说同一个词有两个词向量，哪个作为最终的、提供给其他应用使用的embeddings呢？有两种策略，要么加起来，要么拼接起来。在CS224n的编程练习中，采取的是拼接起来的策略

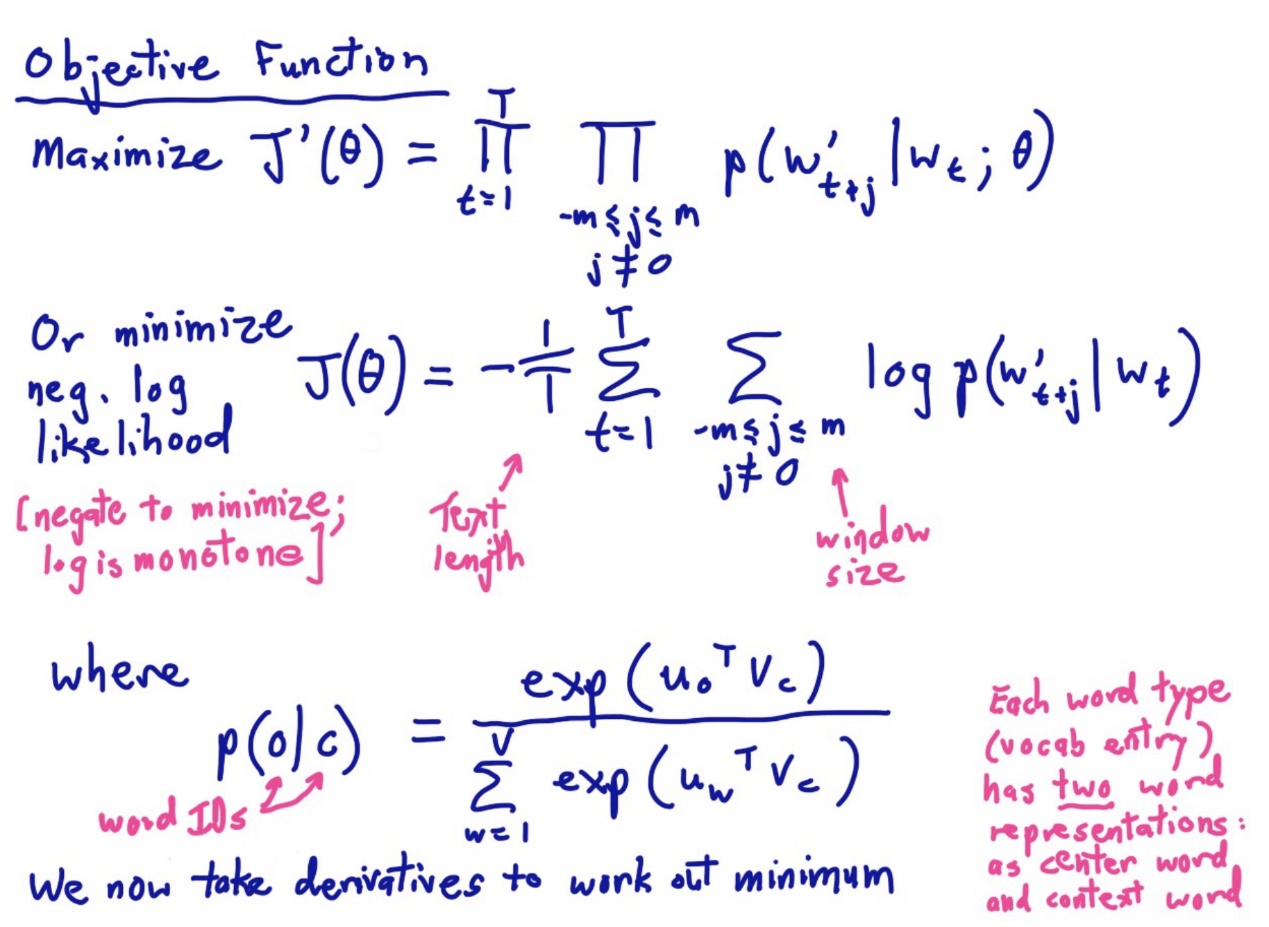


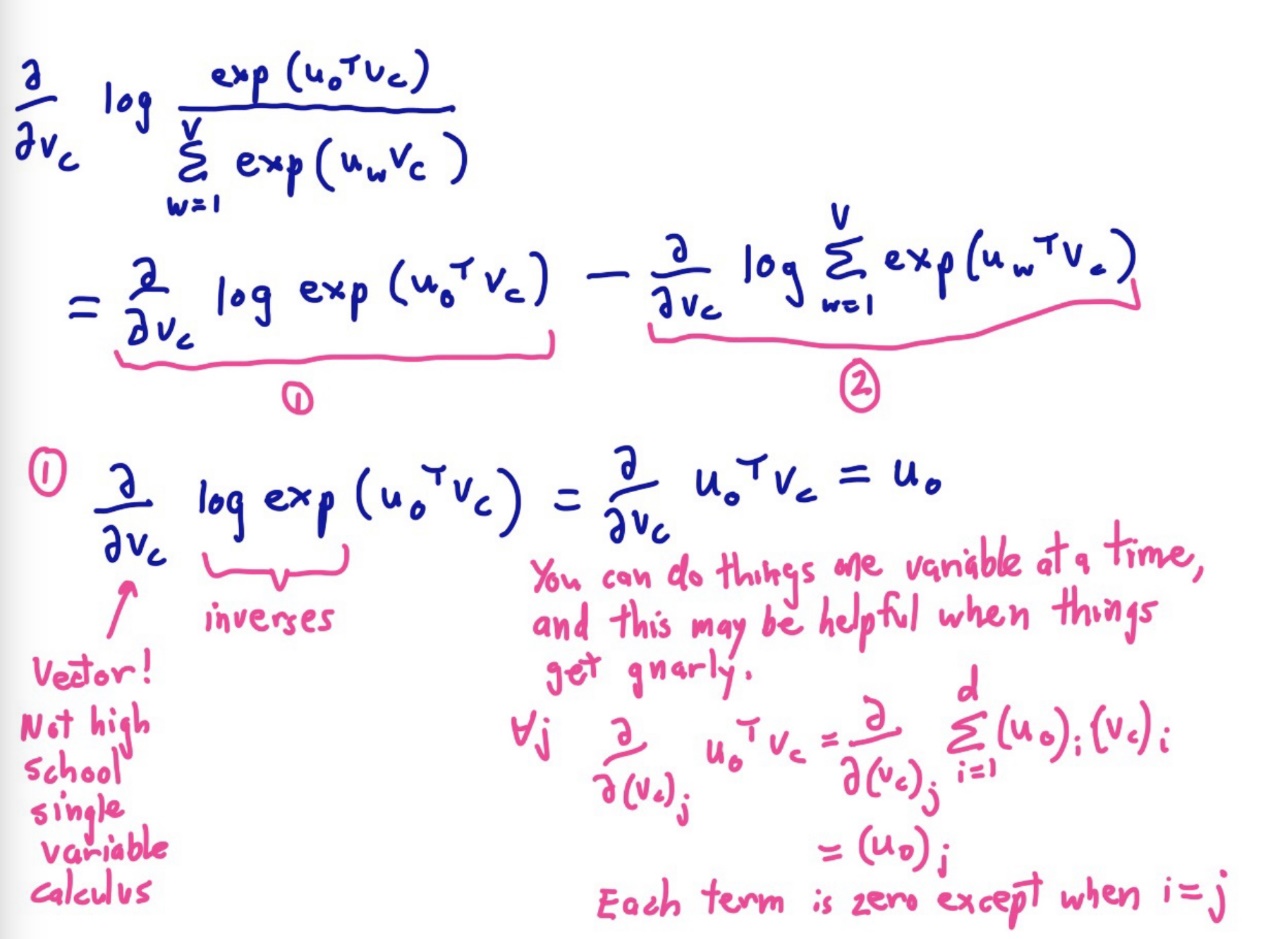
目标函数定义为所有位置的预测结果的乘积：

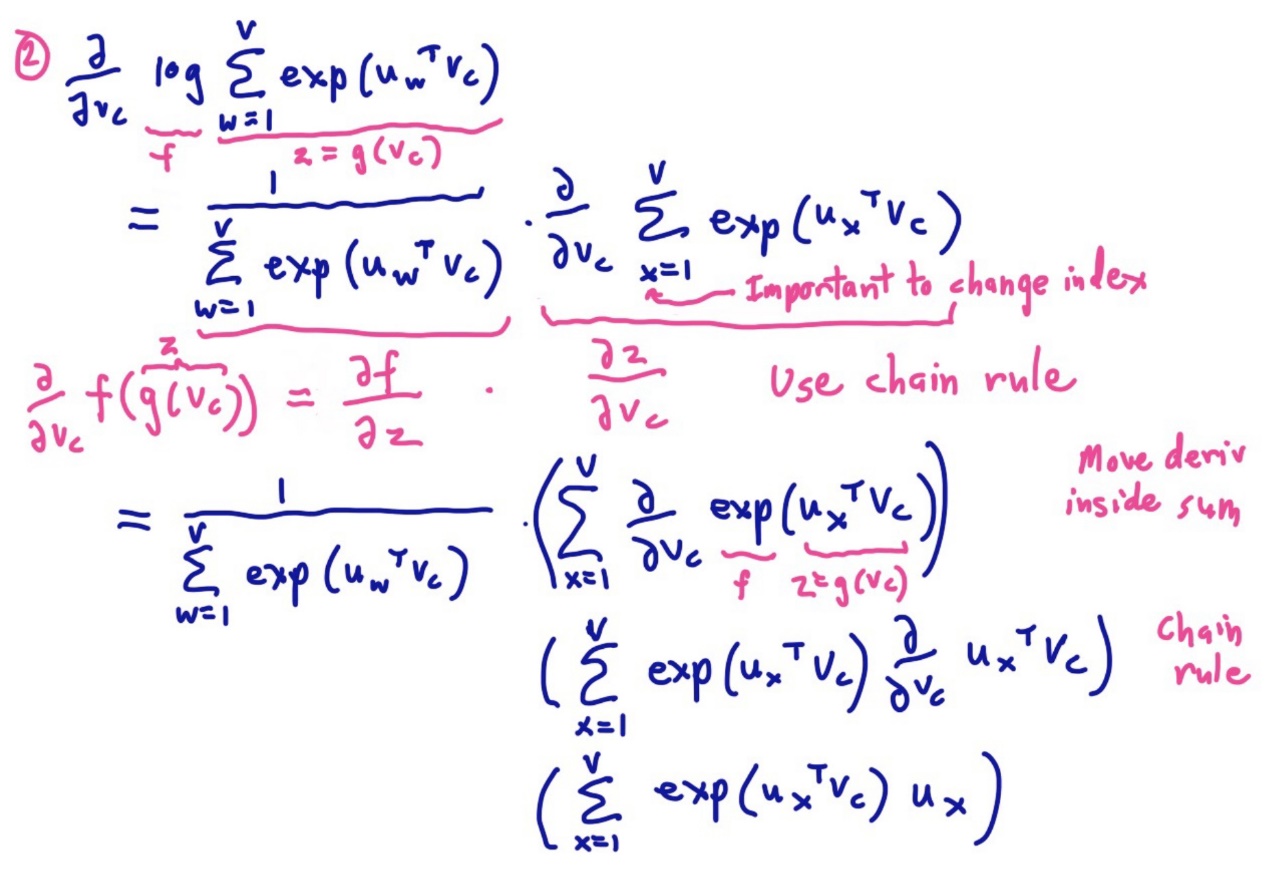


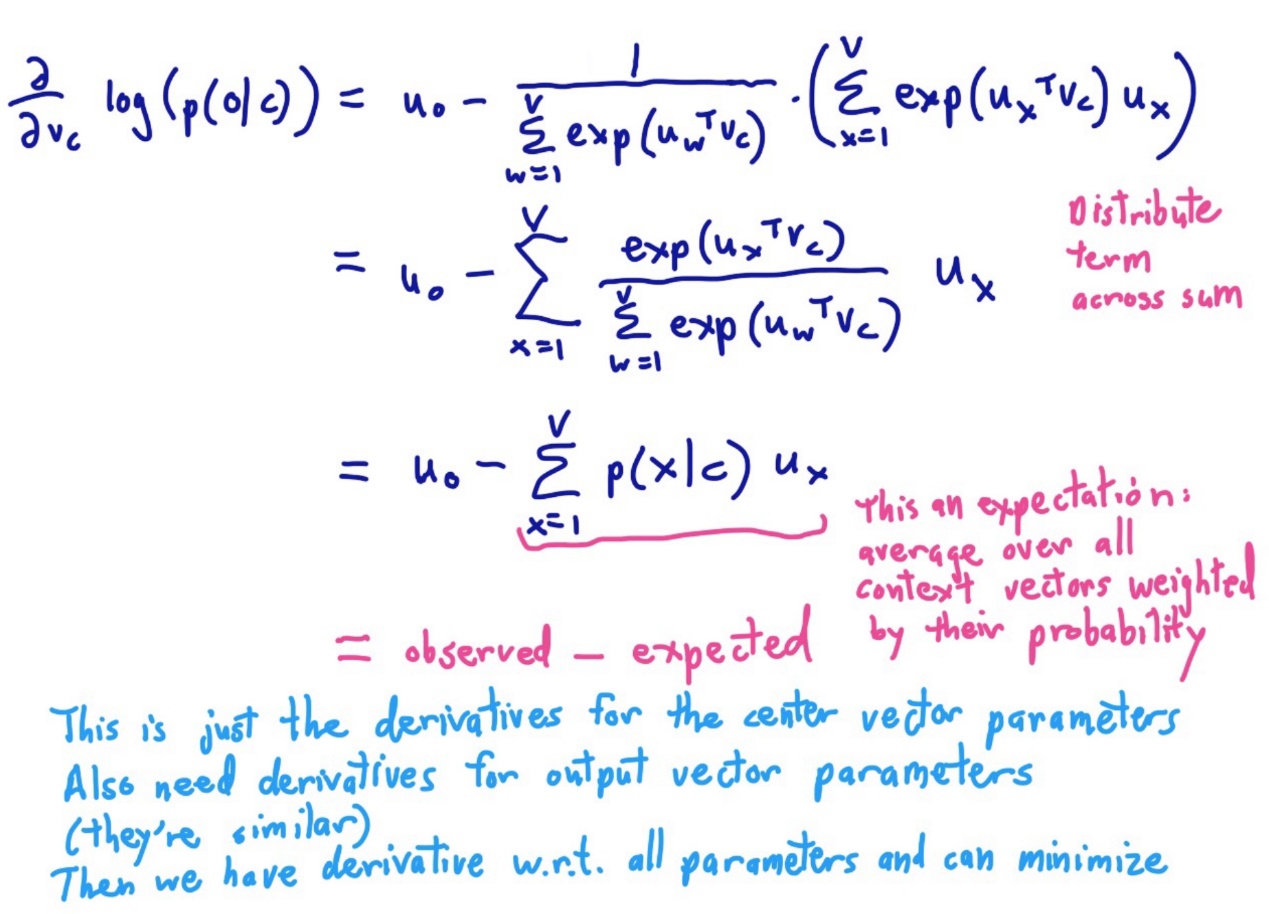
要最大化目标函数。对其取个负对数，得到损失函数——对数似然的相反数：











导数的后一项刚好就是均值（期望）。p为概率，乘以u。然后加和，就是均值的定义；

**梯度下降、SGD**

