数值稳定性

当神经网络变得很深的时候，很可能变得不稳定

神经网络的梯度

- 考虑如下有 d 层的神经网络

t：层 $\mathbf{h}^t = f_t(\mathbf{h}^{t-1})$ and $y = \ell \circ f_d \circ \ldots \circ f_1(\mathbf{x})$ 即x -> f1 -> ... -> fd -> loss ->y

- 计算损失 ℓ 关于参数 $\mathbf{W}_t$ 的梯度

$$\frac{\partial \ell}{\partial \mathbf{W}^t} = \frac{\partial \ell}{\partial \mathbf{h}^d} \underbrace{\frac{\partial \mathbf{h}^d}{\partial \mathbf{h}^{d-1}} \cdots \frac{\partial \mathbf{h}^{t+1}}{\partial \mathbf{h}^t}}_{d-t \text{ 次矩阵乘法}} \frac{\partial \mathbf{h}^t}{\partial \mathbf{W}^t}$$

向量关于向量的导数是矩阵，所以这里实际上是矩阵乘法

矩阵乘法次数过多，会导致。。。

数值稳定性的常见两个问题

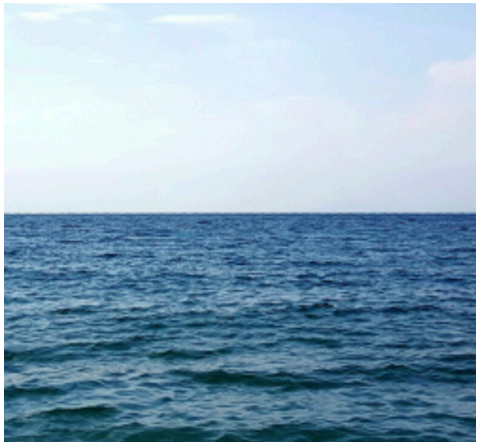
$$\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^{i}}$$

梯度爆炸

$$1.5^{100} \approx 4 \times 10^{17}$$

如果梯度稍微大于1，那累计100次可能会超过浮点数上限

梯度消失



$$0.8^{100} \approx 2 \times 10^{-10}$$

如果梯度稍微小于1，那累计100次可能会超过浮点数下限

例子：MLP

- 加入如下 MLP （为了简单省略了偏移）

$$f_t(\mathbf{h}^{t-1}) = \sigma(\mathbf{W}^t\mathbf{h}^{t-1}) \qquad \sigma \text{ 是激活函数}$$

$$\frac{\partial \mathbf{h}^t}{\partial \mathbf{h}^{t-1}} = \text{diag}\left(\sigma'(\mathbf{W}^t\mathbf{h}^{t-1})\right)(W^t)^T \qquad \sigma' \text{ 是}\sigma\text{的导数函数}$$

对角矩阵，只有主对角线上有元素

为什么是对角矩阵？激活函数是对元素的而不是对向量的，因此只有对应的元素会被涉及，合起来就是对角阵

$$\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^i} = \prod_{i=t}^{d-1} \text{diag}\left(\sigma'(\mathbf{W}^i\mathbf{h}^{i-1})\right)(W^i)^T$$

梯度爆炸

- 使用 ReLU 作为激活函数

$$\sigma(x) = \max(0,x) \quad \text{and} \quad \sigma'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- $\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^{i}} = \prod_{i=t}^{d-1} \text{diag}\left(\sigma'(\mathbf{W}^i \mathbf{h}^{i-1})\right)(W^i)^T$ 的一些元素会来自于 $\prod_{i=t}^{d-1} (W^i)^T$
  - 如果 d-t 很大，值将会很大

由于ReLU的导数是0 or 1，只有当输入大于0的时候才会有梯度传播。但是每一次的权重Wi也会对梯度的大小产生巨大影响权重矩阵的值如果很大，梯度也会被放大。

梯度爆炸的问题

- 值超出值域（infinity）
  - 对于16位浮点数尤为严重（数值区间 6e-5 - 6e4）

Nvidia GPU 对于16位浮点数计算比32位快一倍

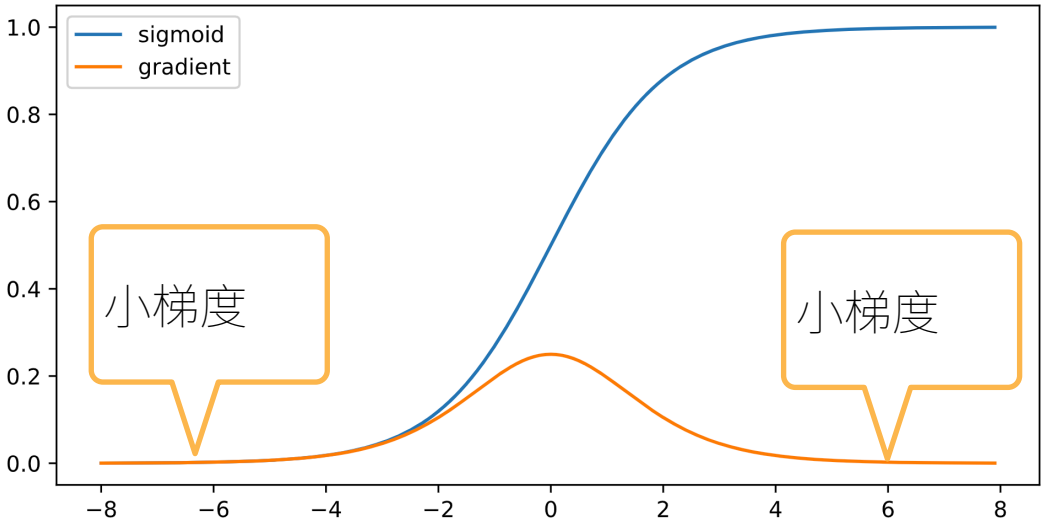
- 对学习率敏感
  - 如果学习率太大 -> 大参数值 -> 更大的梯度
  - 如果学习率太小 -> 训练无进展
  - 我们可能需要在训练过程不断调整学习率

恶性循环

梯度消失

- 使用 sigmoid 作为激活函数

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad \sigma'(x) = \sigma(x)(1 - \sigma(x))$$



梯度消失

- 使用 sigmoid 作为激活函数

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \sigma'(x) = \sigma(x)(1 - \sigma(x))$$

- $\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^{i}} = \prod_{i=t}^{d-1} \text{diag}\left(\sigma'(\mathbf{W}^i \mathbf{h}^{i-1})\right)(W^i)^T$ 的元素值是 d-t 个小数值的乘积

$$0.8^{100} \approx 2 \times 10^{-10}$$

梯度消失的问题

- 梯度值变成 0
  - 对 16 位浮点数尤为严重
- 训练没有进展
  - 不管如何选择学习率
- 对于底部层尤为严重
  - 仅仅顶部层训练的较好
  - 无法让神经网络更深

梯度越来越小

总结

- 当数值过大或者过小时会导致数值问题
- 常发生在深度模型中，因为其会对 n 个数累乘