# 8 数值稳定性，激活函数和硬件

# 概要

- 数值稳定性
  - 梯度爆炸
  - 梯度消失
- 稳定模型训练
  - 权重初始化
- 激活函数

# 数値稳定性

Wikipedia

# 神经网络的梯度

- 考虑具有 d 层的神经网络

$$\mathbf{h}^t = f_t(\mathbf{h}^{t-1})$$

$$y = \ell \circ f_d \circ \cdots \circ f_1(\mathbf{x})$$

- 计算损失 $\ell$ 的梯度 $\mathbf{W}_t$

$$\frac{\partial \ell}{\partial \mathbf{W}^t} = \frac{\partial \ell}{\partial \mathbf{h}^d} \frac{\partial \mathbf{h}^d}{\partial \mathbf{h}^{d-1}} \cdots \frac{\partial \mathbf{h}^{t+1}}{\partial \mathbf{h}^t} \frac{\partial \mathbf{h}^t}{\partial \mathbf{W}^t}$$

*d-t* 矩阵相乘

# 深度神经网络的两个问题

梯度爆炸

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

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

梯度消失

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

# 例子： MLP

- 假设现有一个MLP (为简单而没有偏差参数)

  $f_t(\mathbf{h}^{t-1}) = \sigma(\mathbf{W}^t\mathbf{h}^{t-1})$ $\sigma$是一个激活函数

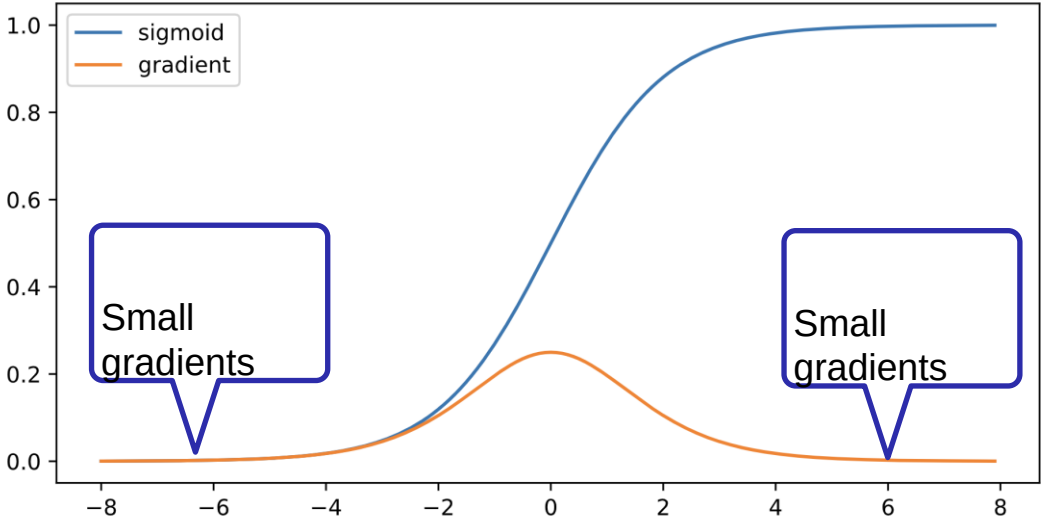
  $\frac{\partial \mathbf{h}^t}{\partial \mathbf{h}^{t-1}} = \mathrm{diag}\left(\sigma'(\mathbf{W}^t\mathbf{h}^{t-1})\right)(W^t)^T$ $\sigma'$ 是 $\sigma$ 的导函数

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

# 梯度消失

- 使用sigmoid作为激活函数

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



# 梯度消失

- 使用sigmoid作为激活函数

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

- $\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^{i}} = \prod_{i=t}^{d-1} \text{diag}\left(\sigma'(\mathbf{W}^{i}\mathbf{h}^{i-1})\right)(W^{i})^{T}$是 t 个较小值的乘积
- 例如 $0.8^{100} \approx 2 \times 10^{-10}$

# 梯度消失的问题

- 梯度值趋近为0的渐变
  - 16位浮点（梯度值小于 $2^{-24} \approx 5.96 \times 10^{-8}$ 即为0）
- 训练没有进展
  - 无论如何选择学习率（LR）
- 底层训练基本无效
  - 只有顶层训练有效
  - 使网络更深可能并没有更好

# 梯度爆炸

- 使用 ReLU 作为激活函数

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

- $\prod_{i=t}^{d-1} (W^i)^T$为来自$\prod_{i=t}^{d-1} \frac{\partial \mathbf{h}^{i+1}}{\partial \mathbf{h}^i} = \prod_{i=t}^{d-1} \text{diag}\left(\sigma'(\mathbf{W}^i \mathbf{h}^{i-1})\right)(W^i)^T$的元素
  - 当$d_t$很大时，可产生很大的值$1.5^{100} \approx 4 \times 10^{17}$

# 梯度爆炸的问题

- 梯度值超出范围：无穷大值
  - 严重的使用 16 位浮点
    - 范围：[6e-5 , 6e4]
- 对学习率（LR）敏感
  - 不够小的 LR -> 更大的权重 -> 更大的梯度
  - 太小的 LR -> 模型训练没有进展
  - 可能需要在训练期间大幅改变 LR

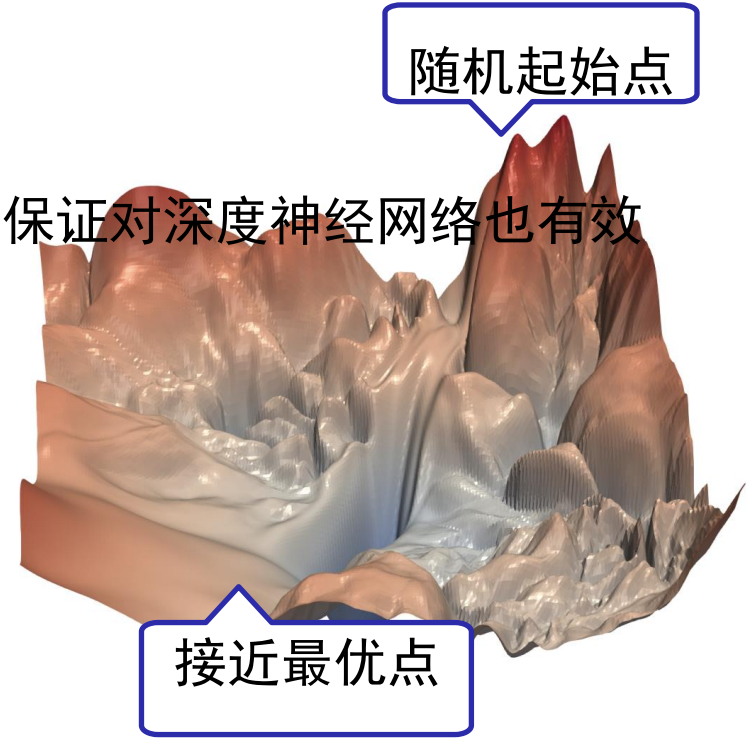
# 稳定模型训练

# 稳定模型训练

- 目标：确保渐变值在适当的范围内
  - 例如 在 [1e-6,1e3] 之间
- 改变神经网络框架结构（乘法 -> 加）
  - ResNet，LSTM
- 归一化
  - 批量归一化，渐变修剪
- 适当的权重初始化和激活函数

# 权重初始化

- 使用适当范围内的随机值初始化权重
- 训练的开始容易受到数值不稳定性的影响
  - 远离最优点的表面可能很复杂
  - 接近最优点的表面可能更平坦
- 根据 $\mathcal{N}(0,0.01)$初始化的小网络很有效，但不保证对深度神经网络也有效



# 每层神经网络的常数方差

- 将每层的输出和梯度看成随机变量
- 使每层的输出的均值和方差相同，类似于梯度

正向

$$\mathbb{E}[h_i^t] = 0$$

$$\mathrm{Var}[h_i^t] = a$$

反向

$$\mathbb{E}\left[\frac{\partial \ell}{\partial h_i^t}\right] = 0 \qquad \mathrm{Var}\left[\frac{\partial \ell}{\partial h_i^t}\right] = b \qquad \forall i, t$$

*a* 和 *b* 都是常数

# 例子： MLP

- 假设：
  - i.i.d $w_{i,j}^t \mathbb{E}\left[w_{i,j}^t\right] = 0, \mathrm{Var}\left[w_{i,j}^t\right] = \gamma_t$
  - $h_i^{t-1}$ 与 $w_{i,j}^t$ 是独立的
  - 激活：用 $\mathbf{h}^t = \mathbf{W}^t \mathbf{h}^{t-1}$, $\mathbf{W}^t \in \mathbb{R}^{n_t \times n_{t-1}}$

$$\mathbb{E}\left[h_i^t\right] = \mathbb{E}\left[\sum_j w_{i,j}^t h_j^{t-1}\right]_0^0 = \sum_j \mathbb{E}\left[w_{i,j}^t\right] \mathbb{E}\left[h_j^{t-1}\right] = 0$$

# 正向方差

$$\mathrm{Var}\left[h_i^t\right] = \mathbb{E}\left[\left(h_i^t\right)^2\right] - \mathbb{E}\left[h_i^t\right]^2 = \mathbb{E}\left[\left(\sum_j w_{i,j}^t h_j^{t-1}\right)^2\right]$$

$$= \mathbb{E}\left[\sum_j \left(w_{i,j}^t\right)^2 \left(h_j^{t-1}\right)^2 + \sum_{j\neq k} w_{i,j}^t w_{i,k}^t \, {}_j^{t-1} h_k^{t-1}\right]$$

$$= \sum_j \mathbb{E}\left[\left(w_{i,j}^t\right)^2\right] \mathbb{E}\left[\left(h_j^{t-1}\right)^2\right] \quad \Longrightarrow \quad \eta_{t-1}\gamma_t = 1$$

$$= \sum_j \mathrm{Var}\left[w_{i,j}^t\right] \mathrm{Var}\left[h_j^{t-1}\right] = n_{t-1}\gamma_t \mathrm{Var}\left[h_j^{t-1}\right]$$

# 反向均值和方差

- 同正向分析：

$$\frac{\partial \ell}{\partial \mathbf{h}^{t-1}} = \frac{\partial \ell}{\partial \mathbf{h}^{t}} \mathbf{W}^{t} \quad \Longrightarrow \quad \left(\frac{\partial \ell}{\partial \mathbf{h}^{t-1}}\right)^{T} = (W^{t})^{T} \left(\frac{\partial \ell}{\partial \mathbf{h}^{t}}\right)^{T}$$

$$\mathbb{E}\left[\frac{\partial \ell}{\partial h_i^{t-1}}\right] = 0$$

$$\mathrm{Var}\left[\frac{\partial \ell}{\partial h_i^{t-1}}\right] = n_t \gamma_t \mathrm{Var}\left[\frac{\partial \ell}{\partial h_j^{t}}\right] \quad \Longrightarrow \quad n_t \gamma_t = 1$$

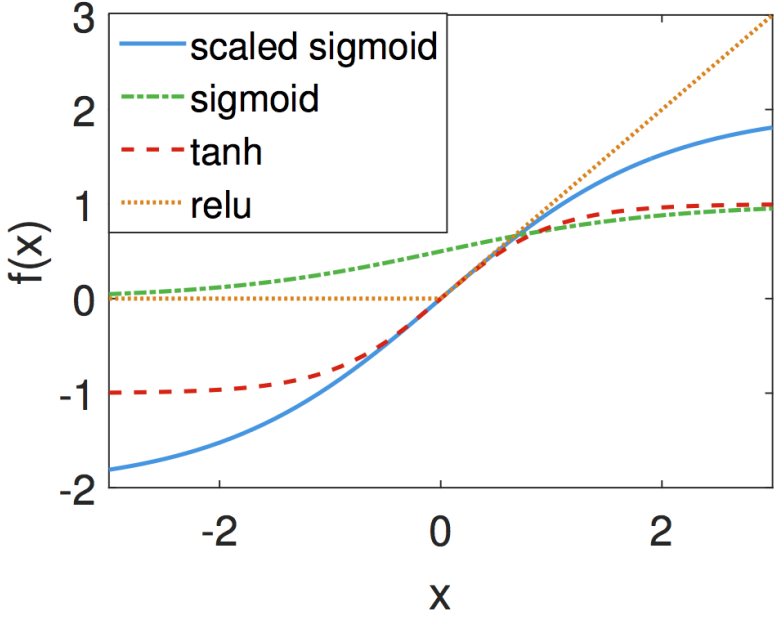
# Xavier 初始化

- 满足 $n_{t-1}\gamma_t = 1$ 和 $n_t\gamma_t = 1$ 是冲突的
- Xavier 初始化

$$\gamma_t(n_{t-1} + n_t)/2 = 1 \rightarrow \gamma_t = 2/(n_{t-1} + n_t)$$

  - 正态分布 $\mathcal{N}\left(0, \sqrt{2/(n_{t-1} + n_t)}\right)$
  - 统一分配 $u\left(-\sqrt{6/(n_{t-1} + n_t)}, \sqrt{6/(n_{t-1} + n_t)}\right)$
    - $U[-a, a]$ 方差是 $a^2/3$
- 适应权重形状，特别是在 $n_t$ 变化时

# 激活函数



# 简单的线性激活函数

- 假设$\sigma(x) = \alpha x + \beta$ $\mathbf{h}' = \mathbf{W}^t\mathbf{h}^{t-1}$ and $\mathbf{h}^t = \sigma(\mathbf{h}')$

$$\mathbb{E}\left[h_i^t\right] = \mathbb{E}[\alpha h_i' + \beta] = \beta \quad \Longrightarrow \quad \beta = 0$$

$$\begin{aligned}\mathrm{Var}\left[h_i^t\right] &= \mathbb{E}\left[\left(h_i^t\right)^2\right] - \mathbb{E}\left[h_i^t\right]^2 \\ &= \mathbb{E}[(\alpha h_i' + \beta)^2] - \beta^2 \\ &= \mathbb{E}[\alpha^2 (h_i')^2 + 2\alpha\beta h_i' + \beta^2] - \beta^2 \\ &= \alpha^2 \mathrm{Var}[h_i']\end{aligned} \quad \Longrightarrow \quad \alpha = 1$$

# 反向

- 假设

$$\sigma(x) = \alpha x + \beta, \quad \frac{\partial \ell}{\partial \mathbf{h}'} = \frac{\partial \ell}{\partial \mathbf{h}^t}(W^t)^T \quad \text{and} \quad \frac{\partial \ell}{\partial \mathbf{h}^{t-1}} = \alpha \frac{\partial \ell}{\partial \mathbf{h}'}$$

- $\mathbb{E}\left[\frac{\partial \ell}{\partial h_i^{t-1}}\right] = 0 \implies \beta = 0$

- $\mathrm{Var}\left[\frac{\partial \ell}{\partial h_i^{t-1}}\right] = \alpha^2 \mathrm{Var}\left[\frac{\partial \ell}{\partial h_j'}\right] \implies \alpha = 1$

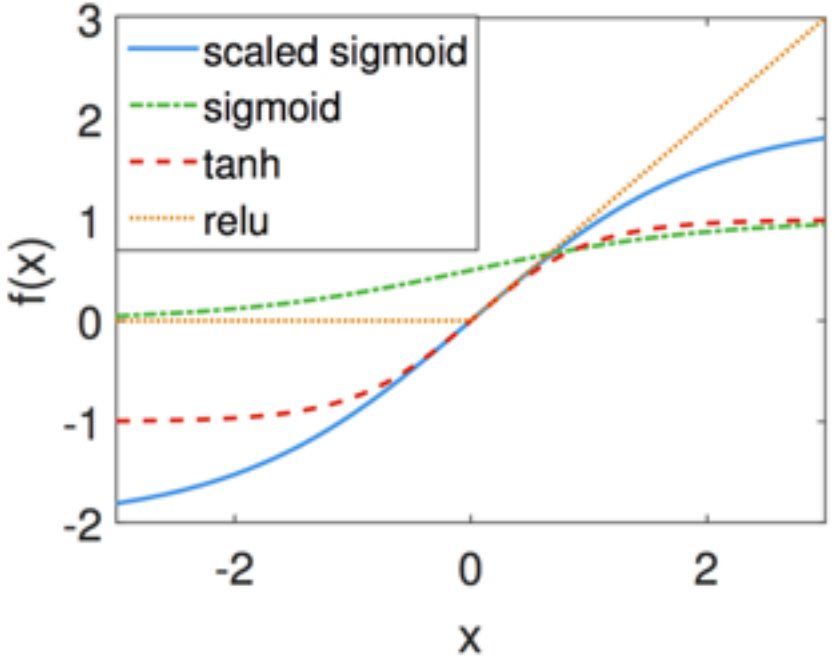
# 回顾激活函数

- 通过 Taylor 展开式

$$\text{sigmoid}(x) = \frac{1}{2}$$

$$\tanh(x) = 0$$

$$\text{relu}(x) =$$




- 用以下方法“纠正”sigmoid

$$4 \times \text{sigmo}$$

# 总结

- 数值稳定性
  - 梯度爆炸
  - 梯度消失
- 稳定模型训练
  - 权重初始化
- 激活函数