让训练更加稳定

让训练更加稳定

- 目标：让梯度值在合理的范围内
  - 例如 [1e-6, 1e3]
- 将乘法变加法
  - ResNet, LSTM
- 归一化
  - 梯度归一化，梯度裁剪
- 合理的权重初始和激活函数

让每层的方差是一个常数

- 将每层的输出和梯度都看做随机变量
- 让它们的均值和方差都保持一致

正向

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

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

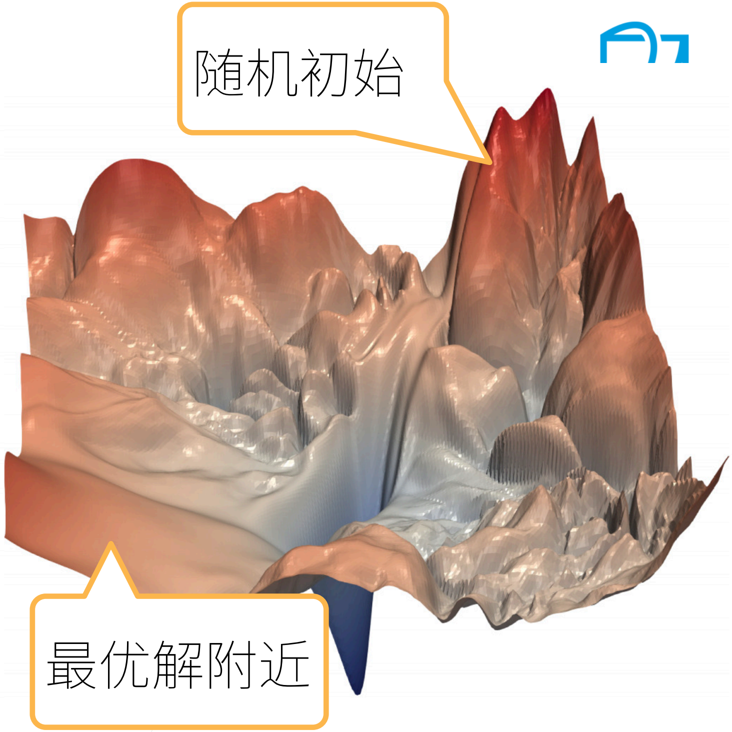
反向

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

a 和 b 都是常数

权重初始化

- 在合理值区间里随机初始参数
- 训练开始的时候更容易有数值不稳定
  - 远离最优解的地方损失函数表面可能很复杂
  - 最优解附近表面会比较平
- 使用 $\mathcal{N}(0, 0.01)$ 来初始可能对小网络没问题，但不能保证深度神经网络



例子：MLP

- 假设
  - $w_{i,j}^t$ 是 i.i.d，那么$\mathbb{E}[w_{i,j}^t] = 0,\ \mathrm{Var}[w_{i,j}^t] = \gamma_t$
  - h_i^{t-1} 独立于 $w_{i,j}^t$
- 假设没有激活函数 $\mathbf{h}^t = \mathbf{W}^t \mathbf{h}^{t-1}$，这里 $\mathbf{W}^t \in \mathbb{R}^{n_t \times n_{t-1}}$

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

正向方差

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

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

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

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

$$n_{t-1}\gamma_t = 1$$

反向均值和方差

- 跟正向情况类似

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

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

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

Xavier 初始

- 难以需要满足 $n_{t-1}\gamma_t = 1$和 $n_t\gamma_t = 1$
- Xavier 使得 $\gamma_t(n_{t-1} + n_t)/2 = 1 \quad \rightarrow \gamma_t = 2/(n_{t-1} + n_t)$
  - 正态分布 $\mathcal{N}\left(0, \sqrt{2/(n_{t-1} + n_t)}\right)$
  - 均匀分布 $\mathcal{U}\left(-\sqrt{6/(n_{t-1} + n_t)}, \sqrt{6/(n_{t-1} + n_t)}\right)$
    - 分布 $\mathcal{U}[-a, a]$ 和方差是 $a^2/3$
- 适配权重形状变换，特别是 n_t

假设线性的激活函数

- 假设 $\sigma(x) = \alpha x + \beta$

$$\mathbf{h}' = \mathbf{W}^t \mathbf{h}^{t-1} \quad \text{and} \quad \mathbf{h}^t = \sigma(\mathbf{h}')$$

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

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

反向

- 假设 $\sigma(x) = \alpha x + \beta$

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

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

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

检查常用激活函数

- 使用泰勒展开

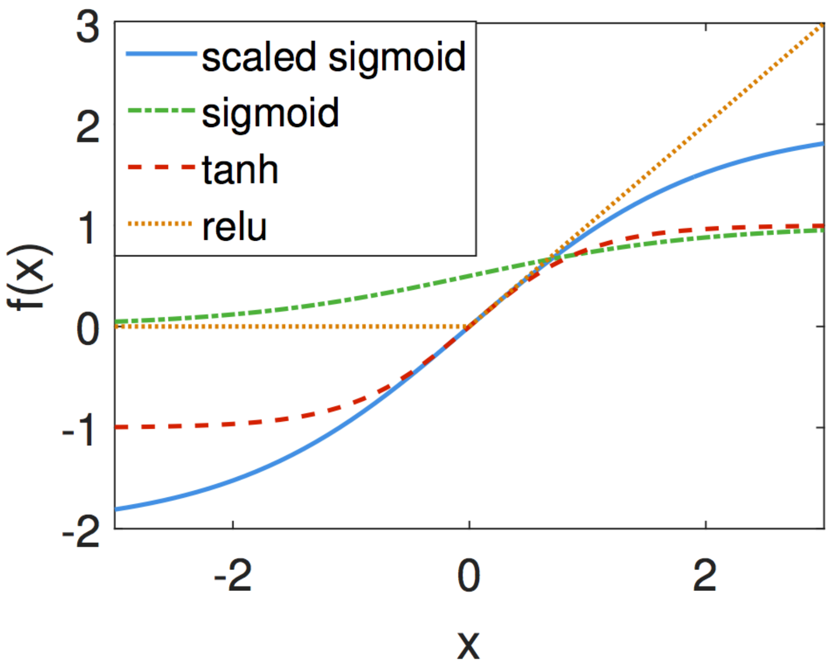
$$\text{sigmoid}(x) = \frac{1}{2} + \frac{x}{4} - \frac{x^3}{48} + O(x^5)$$

$$\tanh(x) = 0 + x - \frac{x^3}{3} + O(x^5)$$

$$\text{relu}(x) = 0 + x \quad \text{for } x \geq 0$$

- 调整 sigmoid：

$$4 \times \text{sigmoid}(x) - 2$$



总结

- 合理的权重初始值和激活函数的选取可以提升数值稳定性