基于平均场理论对 He 初始化的推导

1 引理

假设损失函数为 J，我们有以下 back propagation 公式：

$$\frac{\partial J}{\partial W_{ij}^l} = \delta_i^l \cdot \phi(z_j^{l-1}),\ where\ \delta_i^l = \phi'(z_i^l)\sum_j \delta_j^{l+1} W_{ji}^{l+1} \tag{1}$$

其中，δ_i^l 表示第 l 层第 i 个神经元在反向传播中的误差项，定义为 $\partial J/\partial z_i^l$。

Theorem 1.1

在平均场理论中，某一层权重的梯度模长（也即梯度的波动尺度）$\left\|\nabla_{W_{ij}^l} J\right\|^2$ 与该层误差项 δ_i^l 的方差 $\mathbb{E}\left[(\delta_i^l)^2\right]$ 成正比。

Proof：显然：

$$\left\|\nabla_{W_{ij}^l} J\right\|^2 = \sum_{ij}\left(\frac{\partial J}{\partial W_{ij}^l}\right)^2 \tag{2}$$

因为 W_{ij}^l 是独立同分布且 $\boldsymbol{W}^l$`.shape=(n_l,n_l+1)`，所以有：

$$\left\|\nabla_{W_{ij}^l} J\right\|^2 \approx N_l N_{l+1}\mathbb{E}\left[\left(\frac{\partial J}{\partial W_{ij}^l}\right)^2\right] = N_l N_{l+1}\mathbb{E}\left[\left(\delta_i^l \cdot \phi(z_j^{l-1})\right)^2\right] \tag{3}$$

由于 δ_i^l 和 $\phi(z_j^{l-1})$ 独立同分布，有：

$$\mathbb{E}\left[\left(\frac{\partial J}{\partial W_{ij}^l}\right)^2\right] = \mathbb{E}[(\delta_i^l)^2]\cdot\mathbb{E}[\phi^2(z_j^{l-1})] \tag{4}$$

得证。

Theorem 1.2

在平均场理论中，误差项（即 δ_i^l）的方差满足以下递推关系：

$$\tilde{q}^l = \tilde{q}^{l+1}\left(\mathbb{E}[(\phi'(z_i^l))^2]\sigma_w^2 N_{l+1}\right) \tag{5}$$

Proof：我们直接计算梯度的方差，并应用平均场近似，得到：

$$\tilde{q}^l = \mathbb{E}[(\delta_i^l)^2] = \mathbb{E}[(\phi'(z_i^l))^2] \sum_j \mathbb{E}[(\delta_j^{l+1})^2] \cdot \mathbb{E}[(W_{ji}^{l+1})^2] \tag{6}$$

考虑权重方差为 σ_w^2，有：

$$\tilde{q}^l = \mathbb{E}[(\phi'(z_i^l))^2] \left(\sigma_w^{l+1}\right)^2 \sum_j \mathbb{E}[(\delta_j^{l+1})^2] \tag{7}$$

代入下一层误差项的平均方差 $\tilde{q}^{l+1}$，得到：

$$\tilde{q}^l = \mathbb{E}[(\phi'(z_i^l))^2] \left(\sigma_w^{l+1}\right)^2 N_{l+1} \tilde{q}^{l+1} \tag{8}$$

2 He 初始化

有了上述两个定理，很容易可以得到使用 ReLU 作为 activation function 的层的权重矩阵初始化时的方差。

为了让梯度的波动尺度保持稳定，我们需要让其正比的 $\mathbb{E}\left[(\delta_i^l)^2\right]$ 保持稳定（$\mathbb{E}[\phi^2(z_j^{l-1})]$ 项是激活函数的期望，其本身即保持稳定），也就是要让 $\tilde{q}^l$ 保持稳定，即：

$$\tilde{q}^l \sim \tilde{q}^{l+1} \tag{9}$$

这即要求：

$$\mathbb{E}[(\phi'(z_i^l))^2] \left(\sigma_w^{l+1}\right)^2 N_{l+1} \tilde{q}^{l+1} = \tilde{q}^{l+1} \tag{10}$$

也即：

$$\left(\sigma_w^{l+1}\right)^2 = \frac{1}{\mathbb{E}[(\phi'(z_i^l))^2] N_{l+1}} \tag{11}$$

对于 ReLU，显然有：

$$\mathbb{E}[(\phi'(z_i^l))^2] = \mathbb{E}[\phi'(z_i^l)] = \frac{1}{2} \tag{12}$$

所以：

$$\left(\sigma_w^{l+1}\right)^2 = \frac{2}{N_{l+1}} \tag{13}$$

是能够让梯度的波动尺度保持稳定的方差。

这即是 He 初始化。