DEPARTAMENTO DE
MATEMÁTICA

Teoria Efetiva de Redes Lineares Profundas na Inicialização

Luiz Fernando Bossa
Universidade Federal de Santa Catarina

16 de abril de 2025

Notações e Definições

Teoria Efetiva de Redes Lineares Profundas na Inicialização

Redes Lineares Profundas

Criticalidade

Uma rede neural com L camadas, cada camada tendo n_ℓ neurônios e dados de entrada x_α é dada por:

$$
\begin{aligned}
z^{(1)} &= W^{(1)} x_\alpha + b^{(1)} \\
z^{(\ell+1)} &= W^{(\ell+1)} \sigma\big(z^{(\ell)}\big) + b^{(\ell)}, \qquad \ell = 1, \ldots, L-1
\end{aligned}
\tag{2.5}
$$

- $z^{(\ell)}$ é um vetor de tamanho n_ℓ
- $W^{(\ell)}$ é uma matriz de tamanho $n_\ell \times n_{\ell-1}$

Distribuição inicial: médias zero e variâncias dadas por

$$\mathbb{E}\left(b_i^{(\ell)} b_j^{(\ell)}\right) = \delta_{ij} C_b^{(\ell)} \tag{2.19}$$

$$\mathbb{E}\left(W_{ij}^{(\ell)} W_{kl}^{(\ell)}\right) = \delta_{ik}\delta_{jl} \frac{C_W^{(\ell)}}{n_{\ell-1}} \tag{2.20}$$

Estamos trabalhando com distribuições unidimensionais.

Para duas variáveis aleatórias X e Y com médias zero, temos

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - 0)(Y - 0)) = \mathbb{E}(XY)$$

E em particular,

$$\mathrm{Cov}(X, X) = \mathbb{E}(X^2) = \mathrm{Var}(X)$$

Se A é uma matriz, utilizaremos a notação

- A_{ij} para o elemento da linha i e coluna j.
- A_{i*} para a linha i.
- A_{*j} para a coluna j.
- O produto interno dos vetores u e v será denotado por $u \cdot v$.

- Particularmente eu não gosto de salada de índice, não me cai bem.
- Fiz as seguintes transformações nos índices

Original	**Minha notação**	**Índices**
i_1, i_2	i, j	coordenada fixas
j_1, j_2, j	k, l, ν	coordenadas variáveis
α_1, α_2	α, β	dados de entrada

Assim, podemos escrever as equações (2.19) e (2.20) como

$$(2.19) = \begin{cases} \operatorname{Cov}\left(b_i^{(\ell)}, b_j^{(\ell)}\right) = 0, \quad i \neq j \\ \operatorname{Var}\left(b_i^{(\ell)}\right) = C_b^{(\ell)} \end{cases} \tag{2.19'}$$

$$(2.20) = \begin{cases} \operatorname{Cov}\left(W_{ij}^{(\ell)}, W_{kl}^{(\ell)}\right) = 0, \quad (i,j) \neq (k,l) \\ \operatorname{Var}\left(W_{ij}^{(\ell)}\right) = \frac{C_W^{(\ell)}}{n_{\ell-1}} \end{cases} \tag{2.20'}$$

Embora não valha para todas as distribuições[1], se X e Y são variáveis aleatórias gaussianas, então X e Y são independentes se e somente se $\operatorname{Cov}(X, Y) = 0$.

Segue que as $b_i^{(\ell)}$ e $W_{ij}^{(\ell)}$ são variáveis gaussianas independentes, com médias zero e variâncias dadas por $C_b^{(\ell)}$ e $\frac{C_W^{(\ell)}}{n_{\ell-1}}$.

[1] Independence of Normals

§3.1 Redes Lineares Profundas

§3.2 Criticalidade: cálculo do correlator de 2 pontos

§3.3 ~~Flutuações: cálculo do correlator de 4 pontos~~

§3.4 ~~Caos: cálculo do correlator de 6 pontos~~

- São redes neurais com funções de ativação identidade $\sigma(x) = x$.
- Para simplificar a análise, zeramos os vieses $b^{(\ell)} \equiv \vec{0}$.
- A equação (2.5) se torna

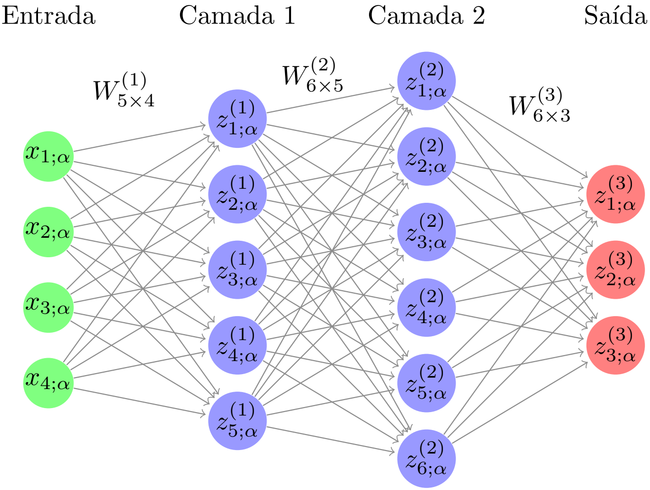
$$\begin{aligned} z^{(1)} &= W^{(1)} x_\alpha \\ z^{(\ell+1)} &= W^{(\ell+1)}\big(z^{(\ell)}\big), \qquad \ell = 1, \dots, L-1 \end{aligned}$$

$$z_{\alpha}^{(\ell)} = W^{(\ell)} W^{(\ell-1)} \cdots W^{(1)} x_{\alpha} \tag{3.2}$$

Introduzimos a notação

$$\mathcal{W}^{(\ell)} = W^{(\ell)} W^{(\ell-1)} \cdots W^{(1)} \tag{3.3}$$

Fazemos todas as variâncias constantes e independentes da camada $C_W^{(\ell)} \equiv C_W$.


Entrada
Camada 1
Camada 2
Saída
$W^{(1)}_{5\times4}$
$W^{(2)}_{6\times5}$
$W^{(3)}_{6\times3}$
$x_{1;\alpha}$
$x_{2;\alpha}$
$x_{3;\alpha}$
$x_{4;\alpha}$
$z^{(1)}_{1;\alpha}$
$z^{(1)}_{2;\alpha}$
$z^{(1)}_{3;\alpha}$
$z^{(1)}_{4;\alpha}$
$z^{(1)}_{5;\alpha}$
$z^{(2)}_{1;\alpha}$
$z^{(2)}_{2;\alpha}$
$z^{(2)}_{3;\alpha}$
$z^{(2)}_{4;\alpha}$
$z^{(2)}_{5;\alpha}$
$z^{(2)}_{6;\alpha}$
$z^{(3)}_{1;\alpha}$
$z^{(3)}_{2;\alpha}$
$z^{(3)}_{3;\alpha}$

Queremos calcular

$$p\left(z_{\alpha}^{(\ell)} \mid \mathcal{D}\right)$$

- Uma distribuição é completamente determinada pelos seus momentos, que são dados por seus correlatores de M pontos.

- Note que pela equação (3.2), temos que

$$z_\alpha^{(\ell)} = W^{(\ell)} z_\alpha^{(\ell-1)} \tag{3.2'}$$

- Podemos calcular a esperança de $z_\alpha^{(\ell)}$ componente a componente, lembrando que é o produto interno da i-ésima linha da matriz $W^{(\ell)}$ com o vetor $z_\alpha^{(\ell-1)}$.

$$
\begin{aligned}
\mathbb{E}\left(z_{i;\alpha}^{(\ell)}\right) &= \mathbb{E}\left(W_{i*}^{(\ell)} \cdot z_{\alpha}^{(\ell-1)}\right) \\
&= \mathbb{E}\left(\sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} z_{j;\alpha}^{(\ell-1)}\right) \\
&= \sum_{j=1}^{n_{\ell-1}} \mathbb{E}\left(W_{ij}^{(\ell)} z_{j;\alpha}^{(\ell-1)}\right) \\
&= \sum_{j=1}^{n_{\ell-1}} \underbrace{\mathbb{E}\left(W_{ij}^{(\ell)}\right)}_{0} \mathbb{E}\left(z_{j;\alpha}^{(\ell-1)}\right) = 0 \qquad (3.6)
\end{aligned}
$$

- Os autores afirmam que, por um argumento similar, é possível mostrar que os momentos de ordem ímpar serão todos zerados.

- Vamos calcular o correlator de 2 pontos na primeira camada, coordenada a coordenada

$$
\begin{aligned}
\mathbb{E}\big(z_{i;\alpha}^{(1)} z_{j;\beta}^{(1)}\big) &= \mathbb{E}\left(W_{i*}^{(1)} \cdot x_\alpha W_{j*}^{(1)} \cdot x_\beta\right) \\
&= \mathbb{E}\left(\left(\sum_{k=1}^{n_0} W_{ik}^{(1)} x_{k;\alpha}\right)\left(\sum_{l=1}^{n_0} W_{il}^{(1)} x_{l;\alpha}\right)\right) \\
&= \mathbb{E}\left(\sum_{k=1}^{n_0}\sum_{l=1}^{n_0} W_{ik}^{(1)} x_{k;\alpha} W_{il}^{(1)} x_{l;\beta}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left(\sum_{k=1}^{n_0}\sum_{l=1}^{n_0} W_{ik}^{(1)} x_{k;\alpha} W_{il}^{(1)} x_{l;\beta}\right) = \sum_{k,l=1}^{n_0} \mathbb{E}\left(W_{ik}^{(1)} W_{jl}^{(1)}\right) x_{k;\alpha} x_{l;\beta} \\
&= \sum_{k,l=1}^{n_0} \delta_{ij}\delta_{kl}\frac{C_W}{n_0} x_{k;\alpha} x_{l;\beta} = \delta_{ij}\frac{C_W}{n_0}\sum_{k,l=1}^{n_0} \delta_{kl} x_{k;\alpha} x_{l;\beta} =^{\dagger} \\
&= \delta_{ij}\frac{C_W}{n_0}\sum_{\nu=1}^{n_0} x_{\nu;\alpha} x_{\nu;\beta} = \delta_{ij}\frac{C_W}{n_0} x_\alpha \cdot x_\beta \qquad (3.8)
\end{aligned}
$$

Na passagem †, note que as parcelas somem quando $k \neq l$, então fazemos uma mudança de variáveis $\nu = k = l$.

Criamos a notação

$$G^{(0)}_{\alpha\beta} = \frac{1}{n_0} x_\alpha \cdot x_\beta \tag{3.9}$$

Assim

$$\mathbb{E}\big(z^{(1)}_{i;\alpha} z^{(1)}_{j;\beta}\big) = \delta_{ij} C_W G^{(0)}_{\alpha\beta} \tag{3.10}$$

- Note que no lado direito da equação acima, o único termo que depende das coordenadas i, j é δ_{ij}.

- Vamos calcular o correlator de 2 pontos na camada $\ell + 1$ de maneira recursiva, utilizando a equação (3.2')

$$z_{\alpha}^{(\ell+1)} = W^{(\ell+1)} z_{\alpha}^{(\ell)} \tag{3.2'}$$

$$
\begin{aligned}
\mathbb{E}\big(z_{i;\alpha}^{(\ell+1)} z_{j;\beta}^{(\ell+1)}\big) &= \mathbb{E}\left(W_{i*}^{(\ell+1)} \cdot z_{\alpha}^{(\ell)} W_{j*}^{(\ell+1)} \cdot z_{\beta}^{(\ell)}\right) \\
&= \mathbb{E}\left(\left(\sum_{k=1}^{n_\ell} W_{ik}^{(\ell+1)} z_{k;\alpha}^{(\ell)}\right)\left(\sum_{l=1}^{n_\ell} W_{il}^{(\ell+1)} z_{l;\alpha}^{(\ell)}\right)\right) \\
&= \sum_{k,l=1}^{n_\ell} \mathbb{E}\left(W_{ik}^{(\ell+1)} W_{jl}^{(\ell+1)} z_{k;\alpha}^{(\ell)} z_{l;\beta}^{(\ell)}\right) \\
&= \sum_{k,l=1}^{n_\ell} \mathbb{E}\left(W_{ik}^{(\ell+1)} W_{jl}^{(\ell+1)}\right) \mathbb{E}\left(z_{k;\alpha}^{(\ell)} z_{l;\beta}^{(\ell)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{k,l=1}^{n_\ell} \mathbb{E}\left(W_{ik}^{(\ell+1)} W_{jl}^{(\ell+1)}\right) \mathbb{E}\left(z_{k;\alpha}^{(\ell)} z_{l;\beta}^{(\ell)}\right) \\
&= \sum_{k,l=1}^{n_\ell} \delta_{ij}\delta_{kl} \frac{C_W}{n_\ell} \mathbb{E}\left(z_{k;\alpha}^{(\ell)} z_{l;\beta}^{(\ell)}\right) = \delta_{ij} \frac{C_W}{n_\ell} \sum_{k,l=1}^{n_\ell} \delta_{kl} \mathbb{E}\left(z_{k;\alpha}^{(\ell)} z_{l;\beta}^{(\ell)}\right) \\
&= \delta_{ij} \frac{C_W}{n_\ell} \sum_{\nu=1}^{n_\ell} \mathbb{E}\left(z_{\nu;\alpha}^{(\ell)} z_{\nu;\beta}^{(\ell)}\right) \\
&= \delta_{ij} \frac{C_W}{n_\ell} \mathbb{E}\left(\sum_{\nu=1}^{n_\ell} z_{\nu;\alpha}^{(\ell)} z_{\nu;\beta}^{(\ell)}\right) = \delta_{ij} \frac{C_W}{n_\ell} \mathbb{E}\big(z_{\alpha}^{(\ell)} \cdot z_{\beta}^{(\ell)}\big) \qquad (3.11)
\end{aligned}
$$

MTM

Correlator camada $\ell + 1$

- Em suma, a equação (3.11) vira

$$\mathbb{E}\left(z_{i;\alpha}^{(\ell+1)} z_{j;\beta}^{(\ell+1)}\right) = \delta_{ij} \frac{C_W}{n_\ell} \mathbb{E}\left(z_\alpha^{(\ell)} \cdot z_\beta^{(\ell)}\right) \tag{3.11}$$

- Em qualquer camada, o correlator das coordenadas i, j é sempre o delta de Kronecker vezes um número que não depende das coordenadas, permitindo assim introduzir a notação

$$\mathbb{E}\left(z_{i;\alpha}^{(\ell)} \cdot z_{j;\beta}^{(\ell)}\right) = \delta_{ij} G_{\alpha\beta}^{(\ell)} \tag{3.12}$$

- Para isolar $G_{\alpha\beta}^{(\ell)}$, vamos somar a equação (3.12) sobre todos os possíveis i e j.

$$\sum_{i,j=1}^{n_\ell} \mathbb{E}\big(z_{i;\alpha}^{(\ell)} z_{j;\beta}^{(\ell)}\big) = \sum_{i,j=1}^{n_\ell} \delta_{ij} G_{\alpha\beta}^{(\ell)}$$

$$\sum_{\nu=1}^{n_\ell} \mathbb{E}\big(z_{\nu;\alpha}^{(\ell)} z_{\nu;\beta}^{(\ell)}\big) = \sum_{\nu=1}^{n_\ell} \delta_{\nu\nu} G_{\alpha\beta}^{(\ell)}$$

$$\mathbb{E}\left(\sum_{\nu=1}^{n_\ell} z_{\nu;\alpha}^{(\ell)} z_{\nu;\beta}^{(\ell)}\right) = \sum_{\nu=1}^{n_\ell} G_{\alpha\beta}^{(\ell)}$$

$$\mathbb{E}\big(z_{\alpha}^{(\ell)} \cdot z_{\beta}^{(\ell)}\big) = n_\ell G_{\alpha\beta}^{(\ell)}$$

$$G_{\alpha\beta}^{(\ell)} = \frac{1}{n_\ell}\mathbb{E}(z_\alpha^{(\ell)} \cdot z_\beta^{(\ell)}) \tag{3.13}$$

Assim (3.11) se torna

$$\mathbb{E}\Big(z_{i;\alpha}^{(\ell+1)} z_{j;\beta}^{(\ell+1)}\Big) = \delta_{ij} C_W G_{\alpha\beta}^{(\ell)} \tag{3.11'}$$

Usando (3.11'), podemos encontrar a recursão para $G_{\alpha\beta}^{(\ell+1)}$.

$$
\begin{aligned}
G_{\alpha\beta}^{(\ell+1)} &= \frac{1}{n_{\ell+1}} \mathbb{E}\big(z_{\alpha}^{(\ell+1)} \cdot z_{\beta}^{(\ell+1)}\big) \\
&= \frac{1}{n_{\ell+1}} \mathbb{E}\left(\sum_{\nu=1}^{n_{\ell+1}} z_{\nu;\alpha}^{(\ell+1)} z_{\nu;\beta}^{(\ell+1)}\right) \\
&= \frac{1}{n_{\ell+1}} \sum_{\nu=1}^{n_{\ell+1}} \mathbb{E}\left(z_{\nu;\alpha}^{(\ell+1)} z_{\nu;\beta}^{(\ell+1)}\right) \\
&= \frac{1}{n_{\ell+1}} \sum_{\nu=1}^{n_{\ell+1}} \delta_{\nu\nu} C_W G_{\alpha\beta}^{(\ell)} \\
&= \frac{C_W}{n_{\ell+1}} \sum_{\nu=1}^{n_{\ell+1}} G_{\alpha\beta}^{(\ell)} = \frac{C_W}{n_{\ell+1}} n_{\ell+1} G_{\alpha\beta}^{(\ell)} = C_W G_{\alpha\beta}^{(\ell)}
\end{aligned}
\tag{3.14}
$$

Da equação (3.14) obtemos a recursão

$$G^{(\ell)}_{\alpha\beta} = (C_W)^{\ell} G^{(0)}_{\alpha\beta} \tag{3.15}$$

O observável $G^{(L)}_{\alpha\alpha}$ mede o tamanho médio do output da rede neural.

$$G^{(L)}_{\alpha\alpha} = \frac{1}{n_L}\mathbb{E}\left(z^{(L)}_{\alpha} \cdot z^{(L)}_{\alpha}\right) = \frac{1}{n_L}\mathbb{E}\left(\left\|z^{(L)}_{\alpha}\right\|^2\right) \tag{3.16}$$

Por outro lado, note que

$$G^{(L)}_{\alpha\alpha} = (C_W)^L G^{(0)}_{\alpha\alpha}$$

Assim, dependendo do valor da variância C_W, podemos ter três cenários:

$$\lim_{L\to\infty} G_{\alpha\alpha}^{(L)} = \lim_{L\to\infty} (C_W)^L G_{\alpha\alpha}^{(0)} = \begin{cases} 0 & \text{se } C_W < 1 \\ G_{\alpha\alpha}^{(0)} & \text{se } C_W = 1 \\ \infty & \text{se } C_W > 1 \end{cases}$$

- Se $C_W < 1$, a rede neural não consegue aprender, pois o output tende a zero.
- Se $C_W > 1$, o valor do output diverge, o que significa instabilidade numérica.
- O único caso no qual a rede neural consegue aprender é quando $C_W = 1$.