DEPARTAMENTO DE
MATEMÁTICA

Teoria Efetiva de Redes Lineares Profundas na Inicialização

Luiz Fernando Bossa
Universidade Federal de Santa Catarina

3 de maio de 2025

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

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

- $z^{(\ell)}$ 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

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

E em particular,

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

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

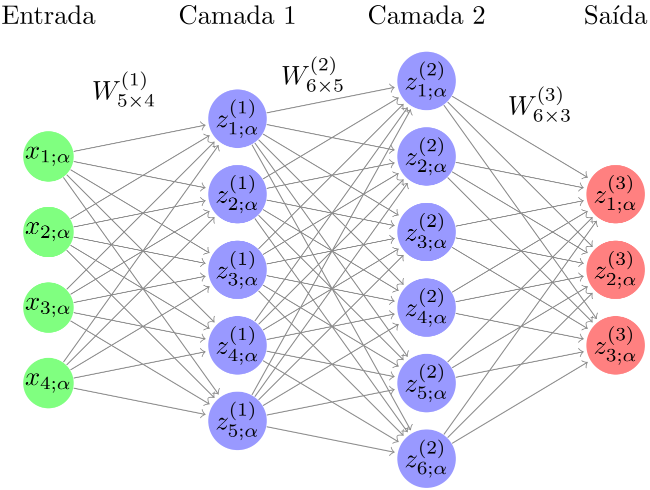
$$
\begin{aligned}
z^{(1)} &= W^{(1)} x_\alpha \\
z^{(\ell+1)} &= W^{(\ell+1)} \big(z^{(\ell)}\big), \qquad \ell = 1, \ldots, L-1
\end{aligned}
$$

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

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
$W^{(1)}_{5\times 4}$
$W^{(2)}_{6\times 5}$
$W^{(3)}_{6\times 3}$
$x_{1;\alpha}$
$x_{2;\alpha}$
$x_{3;\alpha}$
$x_{4;\alpha}$
$z^{(1)}_{1;\alpha}$
$z^{(1)}_{2;\alpha}$
$z^{(1)}_{3;\alpha}$
$z^{(1)}_{4;\alpha}$
$z^{(1)}_{5;\alpha}$
$z^{(2)}_{1;\alpha}$
$z^{(2)}_{2;\alpha}$
$z^{(2)}_{3;\alpha}$
$z^{(2)}_{4;\alpha}$
$z^{(2)}_{5;\alpha}$
$z^{(2)}_{6;\alpha}$
$z^{(3)}_{1;\alpha}$
$z^{(3)}_{2;\alpha}$
$z^{(3)}_{3;\alpha}$

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 $z_{\alpha}^{(\ell-1)}$.

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

- 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

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

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

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

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

- 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')

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

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

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

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

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

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

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

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

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

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

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

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 $G_{\alpha\beta}^{(\ell+1)}$.

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

Da equação (3.14) obtemos a recursão

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

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

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

Por outro lado, note que

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

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 $C_W = 1$.

Novamente, para evitar subscrito duplo, vamos fazer as seguintes mudanças de notação:

- i_1, i_2, i_3, i_4 para $\widehat{1}$, $\widehat{2}$, $\widehat{3}$, $\widehat{4}$
- j_1, j_2, j_3, j_4 para $\underline{1}, \underline{2}, \underline{3}, \underline{4}$

- O correlator de 4 pontos na camada ℓ é dado por

$$\mathbb{E}\left(z_{\widehat{1}}^{(\ell)} z_{\widehat{2}}^{(\ell)} z_{\widehat{3}}^{(\ell)} z_{\widehat{4}}^{(\ell)}\right)$$

- Vamos calcular o correlator de 4 pontos de maneira recursiva.
- Nessa sessão, vamos calcular as correlações em apenas uma entrada x_α, e portanto vamos abandonar o índice.

- Introduzimos a notação

$$G_2^{(\ell)} := G_{\alpha\alpha}^{(\ell)} = \frac{1}{n_\ell}\mathbb{E}\left(z_\alpha^{(\ell)} \cdot z_\alpha^{(\ell)}\right) \tag{3.17}$$

- Em particular, na camada 0,

$$G_2^{(0)} = \frac{1}{n_0}\mathbb{E}\left(x_\alpha \cdot x_\alpha\right) = \frac{1}{n_0} x \cdot x$$

Teorema de Wick

Para calcular momentos superiores de uma variável aleatória z, usamos a fórmula

$$\mathbb{E}\left(z_{\widehat{1}} z_{\widehat{2}} \dots z_{\widehat{2m}}\right) = \sum \mathbb{E}(z_{\widehat{k_1}} z_{\widehat{k_2}}) \mathbb{E}(z_{\widehat{k_3}} z_{\widehat{k_4}}) \dots \mathbb{E}(z_{\widehat{k_{2m-1}}} z_{\widehat{k_{2m}}})$$

em que a soma é feita sobre todos os pareamentos possíveis dos índices.

- O Teorema de Wick para 4 pontos nos diz que

$$\mathbb{E}\left(z_{\widehat{1}} z_{\widehat{2}} z_{\widehat{3}} z_{\widehat{4}}\right) = \\ \mathbb{E}\left(z_{\widehat{1}} z_{\widehat{2}}\right)\mathbb{E}\left(z_{\widehat{3}} z_{\widehat{4}}\right) + \mathbb{E}\left(z_{\widehat{1}} z_{\widehat{3}}\right)\mathbb{E}\left(z_{\widehat{2}} z_{\widehat{4}}\right) + \mathbb{E}\left(z_{\widehat{1}} z_{\widehat{4}}\right)\mathbb{E}\left(z_{\widehat{2}} z_{\widehat{3}}\right)$$

$$
\begin{aligned}
\mathbb{E}\left(z^{(1)}_{\hat{1}} z^{(1)}_{\hat{2}} z^{(1)}_{\hat{3}} z^{(1)}_{\hat{4}}\right) &= \mathbb{E}\left(W^{(1)}_{\hat{1}*} \cdot x \; W^{(1)}_{\hat{2}*} \cdot x \; W^{(1)}_{\hat{3}*} \cdot x \; W^{(1)}_{\hat{4}*} \cdot x\right) \\
&= \mathbb{E}\left(\sum_{\underline{1}=1}^{n_0} W^{(1)}_{\hat{1}\underline{1}} x_{\underline{1}} \sum_{\underline{2}=1}^{n_0} W^{(1)}_{\hat{2}\underline{2}} x_{\underline{2}} \sum_{\underline{3}=1}^{n_0} W^{(1)}_{\hat{3}\underline{3}} x_{\underline{3}} \sum_{\underline{4}=1}^{n_0} W^{(1)}_{\hat{4}\underline{4}} x_{\underline{4}}\right) \\
&= \sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_0} \mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}} W^{(1)}_{\hat{2}\underline{2}} W^{(1)}_{\hat{3}\underline{3}} W^{(1)}_{\hat{4}\underline{4}}\right) x_{\underline{1}} x_{\underline{2}} x_{\underline{3}} x_{\underline{4}} \qquad (3.18)
\end{aligned}
$$

Aplicamos o Teorema de Wick para o termo com esperança, lembrando que $\mathbb{E}\left(W^{(\ell)}_{ij} W^{(\ell)}_{kl}\right) = \delta_{ik}\delta_{jl}\frac{C_W}{n_{\ell-1}}$.

$$
\begin{aligned}
\mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}}W^{(1)}_{\hat{2}\underline{2}}W^{(1)}_{\hat{3}\underline{3}}W^{(1)}_{\hat{4}\underline{4}}\right) &= \mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}}W^{(1)}_{\hat{2}\underline{2}}\right)\mathbb{E}\left(W^{(1)}_{\hat{3}\underline{3}}W^{(1)}_{\hat{4}\underline{4}}\right) + \\
&\mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}}W^{(1)}_{\hat{3}\underline{3}}\right)\mathbb{E}\left(W^{(1)}_{\hat{2}\underline{2}}W^{(1)}_{\hat{4}\underline{4}}\right) + \mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}}W^{(1)}_{\hat{4}\underline{4}}\right)\mathbb{E}\left(W^{(1)}_{\hat{2}\underline{2}}W^{(1)}_{\hat{2}\underline{3}}\right) = \\
&\delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\frac{C_W}{n_0}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}}\frac{C_W}{n_0} + \delta_{\hat{1}\hat{3}}\delta_{\underline{13}}\frac{C_W}{n_0}\delta_{\hat{2}\hat{4}}\delta_{\underline{24}}\frac{C_W}{n_0} + \\
&\delta_{\hat{1}\hat{4}}\delta_{\underline{14}}\frac{C_W}{n_0}\delta_{\hat{2}\hat{3}}\delta_{\underline{23}}\frac{C_W}{n_0} =
\end{aligned}
$$

Agrupando os termos, obtemos

$$\mathbb{E}\left(W^{(1)}_{\hat{1}\underline{1}}W^{(1)}_{\hat{2}\underline{2}}W^{(1)}_{\hat{3}\underline{3}}W^{(1)}_{\hat{4}\underline{4}}\right) = \\ = \frac{C_W^2}{n_0^2}\left(\delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}} + \delta_{\hat{1}\hat{3}}\delta_{\underline{13}}\delta_{\hat{2}\hat{4}}\delta_{\underline{24}} + \delta_{\hat{1}\hat{4}}\delta_{\underline{14}}\delta_{\hat{2}\hat{3}}\delta_{\underline{23}}\right)$$

Voltando para (3.18), obtemos

$$\mathbb{E}\left(z^{(1)}_{\hat{1}} z^{(1)}_{\hat{2}} z^{(1)}_{\hat{3}} z^{(1)}_{\hat{4}}\right) =$$

$$\frac{C_W^2}{n_0^2} \sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_0} \begin{pmatrix} \delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}} \\ +\delta_{\hat{1}\hat{3}}\delta_{\underline{13}}\delta_{\hat{2}\hat{4}}\delta_{\underline{24}} \\ +\delta_{\hat{1}\hat{4}}\delta_{\underline{14}}\delta_{\hat{2}\hat{3}}\delta_{\underline{23}} \end{pmatrix} x_{\underline{1}} x_{\underline{2}} x_{\underline{3}} x_{\underline{4}}$$

Vamos nos atentar ao primeiro grupo de deltas, e os outros saem de maneira análoga.

Os índices 'chapéu' são fixos, então podemos retirar da soma

$$\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_0} \delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}}x_{\underline{1}}x_{\underline{2}}x_{\underline{3}}x_{\underline{4}} = \delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_0} \delta_{\underline{12}}\delta_{\underline{34}}x_{\underline{1}}x_{\underline{2}}x_{\underline{3}}x_{\underline{4}}$$

O primeiro delta só é diferente de zero quando $\underline{1} = \underline{2}$, assim fazemos a mudança de variáveis $\nu = \underline{1} = \underline{2}$. De modo análogo, $\mu = \underline{3} = \underline{4}$. Assim, obtemos

$$\begin{aligned}\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\nu,\mu=1}^{n_0}\delta_{\nu\nu}\delta_{\mu\mu}x_\nu x_\nu x_\mu x_\mu &= \delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\nu=1}^{n_0}x_\nu x_\nu\sum_{\mu=1}^{n_0}x_\mu x_\mu \\ &= \delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}(x\cdot x)^2\end{aligned}$$

Voltando para (3.18), obtemos

$$
\begin{aligned}
&\mathbb{E}\big(z^{(1)}_{\widehat{1}} z^{(1)}_{\widehat{2}} z^{(1)}_{\widehat{3}} z^{(1)}_{\widehat{4}}\big) = \\
&= \frac{C_W^2}{n_0^2}\left(\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}}(x\cdot x)^2 + \delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}}(x\cdot x)^2 + \delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}(x\cdot x)^2\right) \\
&= \frac{C_W^2}{n_0^2}(x\cdot x)^2\left(\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}} + \delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}} + \delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}\right) \\
&= C_W^2\left(\frac{x\cdot x}{n_0}\right)^2\left(\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}} + \delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}} + \delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}\right) \\
&= C_W^2\left(G_2^{(0)}\right)^2\left(\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}} + \delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}} + \delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}\right) \quad (3.18)
\end{aligned}
$$

- O mesmo raciocínio pode ser aplicado para a camada ℓ, e obtemos

$$
\begin{aligned}
&\mathbb{E}\big(z^{(\ell+1)}_{\hat{1}} z^{(\ell+1)}_{\hat{2}} z^{(\ell+1)}_{\hat{3}} z^{(\ell+1)}_{\hat{4}}\big) = \\
&\qquad \mathbb{E}\left(W^{(\ell+1)}_{\hat{1}*} z^{(\ell)}\; W^{(\ell+1)}_{\hat{2}*} z^{(\ell)}\; W^{(\ell+1)}_{\hat{3}*} z^{(\ell)}\; W^{(\ell+1)}_{\hat{4}*} z^{(\ell)}\right) = \\
&\mathbb{E}\left(\sum_{\underline{1}=1}^{n_\ell} W^{(\ell+1)}_{\hat{1}\underline{1}} z^{(\ell)}_{\underline{1}} \sum_{\underline{2}=1}^{n_\ell} W^{(\ell+1)}_{\hat{2}\underline{2}} z^{(\ell)}_{\underline{2}} \sum_{\underline{3}=1}^{n_\ell} W^{(\ell+1)}_{\hat{3}\underline{3}} z^{(\ell)}_{\underline{3}} \sum_{\underline{4}=1}^{n_\ell} W^{(\ell+1)}_{\hat{4}\underline{4}} z^{(\ell)}_{\underline{4}}\right) = \\
&\qquad \sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}} W^{(\ell+1)}_{\hat{2}\underline{2}} W^{(\ell+1)}_{\hat{3}\underline{3}} W^{(\ell+1)}_{\hat{4}\underline{4}} z^{(\ell)}_{\underline{1}} z^{(\ell)}_{\underline{2}} z^{(\ell)}_{\underline{3}} z^{(\ell)}_{\underline{4}}\right) \qquad (3.20)
\end{aligned}
$$

O que acontece na camada $\ell + 1$ é independente do que acontece na camada ℓ, então a esperança do produto é o produto das esperanças.

$$\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}} W^{(\ell+1)}_{\hat{2}\underline{2}} W^{(\ell+1)}_{\hat{3}\underline{3}} W^{(\ell+1)}_{\hat{4}\underline{4}} z^{(\ell)}_{\underline{1}} z^{(\ell)}_{\underline{2}} z^{(\ell)}_{\underline{3}} z^{(\ell)}_{\underline{4}}\right) = \sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}} W^{(\ell+1)}_{\hat{2}\underline{2}} W^{(\ell+1)}_{\hat{3}\underline{3}} W^{(\ell+1)}_{\hat{4}\underline{4}}\right) \mathbb{E}\left(z^{(\ell)}_{\underline{1}} z^{(\ell)}_{\underline{2}} z^{(\ell)}_{\underline{3}} z^{(\ell)}_{\underline{4}}\right) \tag{3.20}$$

Usando o Teorema de Wick, e fazendo o mesmo cálculo que fizemos para a camada 1, obtemos

$$
\begin{aligned}
\mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}}W^{(\ell+1)}_{\hat{2}\underline{2}}W^{(\ell+1)}_{\hat{3}\underline{3}}W^{(\ell+1)}_{\hat{4}\underline{4}}\right) = \\
&\mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}}W^{(\ell+1)}_{\hat{2}\underline{2}}\right)\mathbb{E}\left(W^{(\ell+1)}_{\hat{3}\underline{3}}W^{(\ell+1)}_{\hat{4}\underline{4}}\right) + \\
&\mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}}W^{(\ell+1)}_{\hat{3}\underline{3}}\right)\mathbb{E}\left(W^{(\ell+1)}_{\hat{2}\underline{2}}W^{(\ell+1)}_{\hat{4}\underline{4}}\right) + \\
&\mathbb{E}\left(W^{(\ell+1)}_{\hat{1}\underline{1}}W^{(\ell+1)}_{\hat{4}\underline{4}}\right)\mathbb{E}\left(W^{(\ell+1)}_{\hat{2}\underline{2}}W^{(\ell+1)}_{\hat{3}\underline{3}}\right) = \\
&\frac{C_W^2}{n_\ell^2}\left(\delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}} + \delta_{\hat{1}\hat{3}}\delta_{\underline{13}}\delta_{\hat{2}\hat{4}}\delta_{\underline{24}} + \delta_{\hat{1}\hat{4}}\delta_{\underline{14}}\delta_{\hat{2}\hat{3}}\delta_{\underline{23}}\right)
\end{aligned}
$$

Voltando para (3.20), obtemos

$$
\begin{aligned}
&\mathbb{E}\left(z_{\hat{1}}^{(\ell+1)} z_{\hat{2}}^{(\ell+1)} z_{\hat{3}}^{(\ell+1)} z_{\hat{4}}^{(\ell+1)}\right) = \\
&= \sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \frac{C_W^2}{n_\ell^2} \begin{pmatrix} \delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}} \\ +\delta_{\hat{1}\hat{3}}\delta_{\underline{13}}\delta_{\hat{2}\hat{4}}\delta_{\underline{24}} \\ +\delta_{\hat{1}\hat{4}}\delta_{\underline{14}}\delta_{\hat{2}\hat{3}}\delta_{\underline{23}} \end{pmatrix} \mathbb{E}\left(z_{\underline{1}}^{(\ell)} z_{\underline{2}}^{(\ell)} z_{\underline{3}}^{(\ell)} z_{\underline{4}}^{(\ell)}\right)
\end{aligned} \tag{3.20}
$$

Vamos nos concentrar no primeiro grupo de deltas

$$\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \delta_{\hat{1}\hat{2}}\delta_{\underline{12}}\delta_{\hat{3}\hat{4}}\delta_{\underline{34}}\mathbb{E}\left(z_{\underline{1}}^{(\ell)}z_{\underline{2}}^{(\ell)}z_{\underline{3}}^{(\ell)}z_{\underline{4}}^{(\ell)}\right) =$$

$$= \delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \delta_{\underline{12}}\delta_{\underline{34}}\mathbb{E}\left(z_{\underline{1}}^{(\ell)}z_{\underline{2}}^{(\ell)}z_{\underline{3}}^{(\ell)}z_{\underline{4}}^{(\ell)}\right)$$

Novamente, fazendo $\nu = \underline{1} = \underline{2}$, $\mu = \underline{3} = \underline{4}$, temos

$$\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\nu,\mu=1}^{n_\ell} \delta_{\nu\nu}\delta_{\mu\mu}\mathbb{E}\left(z_{\nu}^{(\ell)}z_{\nu}^{(\ell)}z_{\mu}^{(\ell)}z_{\mu}^{(\ell)}\right) =$$

$$= \delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}}\sum_{\nu,\mu=1}^{n_\ell} \mathbb{E}\left(z_{\nu}^{(\ell)}z_{\nu}^{(\ell)}z_{\mu}^{(\ell)}z_{\mu}^{(\ell)}\right)$$

Aplicando a ideia acima para todos os os grupos de deltas, obtemos

$$\mathbb{E}\big(z^{(\ell+1)}_{\widehat{1}} z^{(\ell+1)}_{\widehat{2}} z^{(\ell+1)}_{\widehat{3}} z^{(\ell+1)}_{\widehat{4}}\big) =$$

$$\sum_{\underline{1},\underline{2},\underline{3},\underline{4}=1}^{n_\ell} \frac{C_W^2}{n_\ell^2} \begin{pmatrix} \delta_{\widehat{1}\widehat{2}}\delta_{\underline{12}}\delta_{\widehat{3}\widehat{4}}\delta_{\underline{34} } \\ +\delta_{\widehat{1}\widehat{3}}\delta_{\underline{13}}\delta_{\widehat{2}\widehat{4}}\delta_{\underline{24}} \\ +\delta_{\widehat{1}\widehat{4}}\delta_{\underline{14}}\delta_{\widehat{2}\widehat{3}}\delta_{\underline{23}} \end{pmatrix} \mathbb{E}\left(z^{(\ell)}_{\underline{1}} z^{(\ell)}_{\underline{2}} z^{(\ell)}_{\underline{3}} z^{(\ell)}_{\underline{4}} \right) =$$

$$= \frac{C_W^2}{n_\ell^2} \left(\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}} + \delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}} + \delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}\right) \sum_{\nu,\mu=1}^{n_\ell} \mathbb{E}\left(z^{(\ell)}_{\nu} z^{(\ell)}_{\nu} z^{(\ell)}_{\mu} z^{(\ell)}_{\mu} \right) \tag{3.20}$$

- Novamente, podemos argumentar que o correlator de 4 pontos é proporcional ao fator

$$\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}$$

  que chamaremos de *fator Wick 4*.
- Chamando a constante de proporcionalidade de $G_4^{(\ell)}$, escrevemos a relação

$$\mathbb{E}\left(z_{\hat{1}}^{(\ell)} z_{\hat{2}}^{(\ell)} z_{\hat{3}}^{(\ell)} z_{\hat{4}}^{(\ell)}\right) = \left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right) G_4^{(\ell)} \quad (3.21)$$

Comparando (3.21) com (3.18), obtemos a relação

$$G_4^{(1)} = C_W^2 \left(G_2^{(0)} \right)^2 \tag{3.22}$$

Aplicando (3.21) no somatório que aparece (3.20), obtemos

$$\sum_{\nu,\mu=1}^{n_\ell} \mathbb{E}\left(z_\nu^{(\ell)} z_\nu^{(\ell)} z_\mu^{(\ell)} z_\mu^{(\ell)}\right) = \sum_{\nu,\mu=1}^{n_\ell} \left(\delta_{\nu\nu}\delta_{\mu\mu} + \delta_{\nu\mu}\delta_{\nu\mu} + \delta_{\nu\mu}\delta_{\nu\mu}\right) G_4^{(\ell)} = \\ = \sum_{\nu,\mu=1}^{n_\ell} \left(\delta_{\nu\nu}\delta_{\mu\mu} + 2\delta_{\nu\mu}\delta_{\nu\mu}\right) G_4^{(\ell)} \quad (3.23)$$

- O primeiro par de deltas é sempre 1, então essa primeira parte da soma é n_ℓ^2.
- O segundo par de deltas só é diferente de zero quando $\nu = \mu$, então essa segunda parte da soma é n_ℓ.
- Portanto, a soma total é $\left(n_\ell^2 + 2n_\ell\right) G_4^{(\ell)}$.

Voltando para (3.20),

$$
\begin{aligned}
&\mathbb{E}\big(z_{\hat{1}}^{(\ell+1)} z_{\hat{2}}^{(\ell+1)} z_{\hat{3}}^{(\ell+1)} z_{\hat{4}}^{(\ell+1)}\big) = \\
&= \frac{C_W^2}{n_\ell^2}\left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right) \sum_{\nu,\mu=1}^{n_\ell} \mathbb{E}\left(z_\nu^{(\ell)} z_\nu^{(\ell)} z_\mu^{(\ell)} z_\mu^{(\ell)}\right) = \\
&= \frac{C_W^2}{n_\ell^2}\left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right)\left(n_\ell^2 + 2n_\ell\right) G_4^{(\ell)} = \\
&= \left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right)\left(1 + \frac{2}{n_\ell}\right) C_W^2 G_4^{(\ell)}
\end{aligned}
$$

Por outro lado, definimos que o correlator de 4 pontos na camada $\ell + 1$ é dado pelo fator Wick 4 multiplicado pela constante de proporcionalidade $G_4^{(\ell+1)}$. Assim, podemos escrever a recorrência

$$G_4^{(\ell+1)} = C_W^2 \left(1 + \frac{2}{n_\ell}\right) G_4^{(\ell)} \tag{3.24}$$

Se abrirmos essa recursão da camada ℓ até a camada 1, e aplicamos (3.22), obtemos

$$
\begin{aligned}
G_4^{(\ell)} = \left[\prod_{\hat{\ell}=1}^{\ell-1} C_W^2 \left(1 + \frac{2}{n_{\hat{\ell}}}\right)\right] G_4^{(1)} &= \left(C_W^2\right)^{\ell-1} G_4^{(1)} \prod_{\hat{\ell}=1}^{\ell-1} \left(1 + \frac{2}{n_{\hat{\ell}}}\right) \\
&= \left(C_W^2\right)^{\ell-1} C_W^2 \left(G_2^{(0)}\right)^2 \prod_{\hat{\ell}=1}^{\ell-1} \left(1 + \frac{2}{n_{\hat{\ell}}}\right) = \\
&= \left(C_W^{\ell} G_2^{(0)}\right)^2 \prod_{\hat{\ell}=1}^{\ell-1} \left(1 + \frac{2}{n_{\hat{\ell}}}\right)
\end{aligned}
$$

Aplicando (3.15), o fator $C_W^{\ell} G_2^{(0)} = G_2^{(\ell)}$, e obtemos

$$G_4^{(\ell)} = \left(G_2^{(\ell)}\right)^2 \prod_{\widehat{\ell}=1}^{\ell-1} \left(1 + \frac{2}{n_{\widehat{\ell}}}\right) \tag{3.25}$$

que relaciona o correlator de 4 pontos com o correlator de 2 pontos.

- Equalizando o número de neurônios em todas as camadas $n_i = n, i = 1, \ldots, L$, a equação (3.25) se torna

$$G_4^{(\ell)} = \left(G_2^{(\ell)}\right)^2 \left(1 + \frac{2}{n}\right)^{\ell-1} \tag{3.25'}$$

- Se fizermos $n \to \infty$, o correlator de 4 pontos converge para

$$G_4^{(\ell)} = \left(G_2^{(\ell)}\right)^2$$

o que tornaria a distribuição gaussiana.

Para medir o desvio da gaussianidade, usamos a aproximação de Taylor centrada em 0 para

$$(1+x)^{\ell-1} \approx 1 + (\ell-1)x + O(x^2)$$

e obtemos

$$\begin{aligned} G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 &= \left(G_2^{(\ell)}\right)^2 \left[\left(1+\frac{2}{n}\right)^{\ell-1} - 1\right] \\ &= \left(G_2^{(\ell)}\right)^2 \left[\frac{2}{n}(\ell-1) + O\left(\frac{1}{n^2}\right)\right] \\ &= \frac{2(\ell-1)}{n}\left(G_2^{(\ell)}\right)^2 + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{3.28}$$

- O desvio da gaussianidade é proporcional ao número de camadas ℓ e inversamente proporcional ao número de neurônios n.
- A magnitude do desvio é proporcional ao quociente $\frac{\ell}{n}$, chamado de *escala emergente*.

O correlator de 4 pontos conexo tem a fórmula:

$$\mathbb{E}(z_{\hat{1}} z_{\hat{2}} z_{\hat{3}} z_{\hat{4}})\Big|_C = \mathbb{E}(z_{\hat{1}} z_{\hat{2}} z_{\hat{3}} z_{\hat{4}}) - \mathbb{E}(z_{\hat{1}} z_{\hat{2}})\mathbb{E}(z_{\hat{3}} z_{\hat{4}}) - \mathbb{E}(z_{\hat{1}} z_{\hat{3}})\mathbb{E}(z_{\hat{2}} z_{\hat{4}}) - \mathbb{E}(z_{\hat{1}} z_{\hat{4}})\mathbb{E}(z_{\hat{2}} z_{\hat{3}}) \tag{1.54}$$

- O teorema de Wick garante que se a distribuição for gaussiana, o correlator de 4 pontos conexo é zero.
- Valores diferentes de zero indicam o desvio da gaussianidade.

Utilizando as equações (3.21) e (3.12), obtemos a fórmula para o correlator de 4 pontos conexo

$$
\begin{aligned}
\mathbb{E}(z_{\hat{1}} z_{\hat{2}} z_{\hat{3}} z_{\hat{4}})\Big|_C &= \left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right) G_4^{(\ell)} \\
&\quad - \delta_{\hat{1}\hat{2}} G_2^{(\ell)} \delta_{\hat{3}\hat{4}} G_2^{(\ell)} - \delta_{\hat{1}\hat{3}} G_2^{(\ell)} \delta_{\hat{2}\hat{4}} G_2^{(\ell)} \\
&\quad - \delta_{\hat{1}\hat{4}} G_2^{(\ell)} \delta_{\hat{2}\hat{3}} G_2^{(\ell)} \\
&= \left(\delta_{\hat{1}\hat{2}}\delta_{\hat{3}\hat{4}} + \delta_{\hat{1}\hat{3}}\delta_{\hat{2}\hat{4}} + \delta_{\hat{1}\hat{4}}\delta_{\hat{2}\hat{3}}\right) \left(G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2\right)
\end{aligned}
\tag{3.29}
$$

Outra maneira de interpretar a não-gaussianidade é através das interações: quebras da independência estatística entre os neurônios. Para $\widehat{1} = \widehat{2} = j \neq \widehat{3} = \widehat{4} = k$

$$
\begin{aligned}
&\mathbb{E}\left(\left(z_j^{(\ell)} z_j^{(\ell)} - G_2^{(\ell)}\right)\left(z_k^{(\ell)} z_k^{(\ell)} - G_2^{(\ell)}\right)\right) = \\
&= \mathbb{E}\left(z_j^{(\ell)} z_j^{(\ell)} z_k^{(\ell)} z_k^{(\ell)}\right) - G_2^{(\ell)} \mathbb{E}\left(z_j^{(\ell)} z_j^{(\ell)}\right) - G_2^{(\ell)} \mathbb{E}\left(z_k^{(\ell)} z_k^{(\ell)}\right) + G_2^{(\ell)^2} = \\
&= (1 + 0 + 0) G_4^{(\ell)} - G_2^{(\ell)} \delta_{jj} G_2^{(\ell)} - G_2^{(\ell)} \delta_{kk} G_2^{(\ell)} + G_2^{(\ell)^2} \\
&= G_4^{(\ell)} - G_2^{(\ell)^2}
\end{aligned}
\tag{3.30}
$$

- O quanto $z_j z_j$ desvia de sua média $G_2^{(\ell)}$ está correlacionado com o quanto $z_k z_k$ desvia de sua média $G_2^{(\ell)}$.

Observável

$$\mathcal{O}^{(\ell)} := \frac{1}{n} z^{(\ell)} \cdot z^{(\ell)}$$

mede a magnitude média do vetor de ativação na camada ℓ.
Seu valor médio é

$$\mathbb{E}\left(\mathcal{O}^{(\ell)}\right) = \frac{1}{n}\mathbb{E}\left(z^{(\ell)} \cdot z^{(\ell)}\right) = G_2^{(\ell)}$$

Quanto esse observável desvia de sua média?

$$
\begin{aligned}
\mathbb{E}\left(\left(\mathcal{O}^{(\ell)} - G_2^{(\ell)}\right)^2\right) &= \mathbb{E}\left(\mathcal{O}^{(\ell)^2}\right) - 2G_2^{(\ell)}\mathbb{E}\left(\mathcal{O}^{(\ell)}\right) + \left(G_2^{(\ell)}\right)^2 \\
&= \mathbb{E}\left(\left(\frac{1}{n} z^{(\ell)} \cdot z(\ell)\right)^2\right) - \left(G_2^{(\ell)}\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{\mu=1}^{n} z_\mu^{(\ell)} z_\mu^{(\ell)} \sum_{\nu=1}^{n} z_\nu^{(\ell)} z_\nu^{(\ell)}\right) - \left(G_2^{(\ell)}\right)^2 \\
&= \frac{1}{n^2}\sum_{\mu,\nu=1}^{n} \mathbb{E}\left(z_\mu^{(\ell)} z_\mu^{(\ell)} z_\nu^{(\ell)} z_\nu^{(\ell)}\right) - \left(G_2^{(\ell)}\right)^2 \\
&= \frac{1}{n^2}\sum_{\mu,\nu=1}^{n} \left(\delta_{\nu\nu}\delta_{\mu\mu} + 2\delta_{\nu\mu}\delta_{\nu\mu}\right) G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 \\
&= \frac{1}{n^2}\left(n^2 + 2n\right) G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 = \left(1 + \frac{2}{n}\right) G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2
\end{aligned}
$$

Aqui, vamos utilizar (3.25'), que nos diz que

$$G_4^{(\ell)} = \left(G_2^{(\ell)}\right)^2 \left(1 + \frac{2}{n}\right)^{\ell-1}$$

Logo, multiplicando tudo pelo fator $(1 + 2/n)$, obtemos

$$\left(1 + \frac{2}{n}\right) G_4^{(\ell)} = \left(G_2^{(\ell)}\right)^2 \left(1 + \frac{2}{n}\right)^{\ell}$$

e assim

$$\begin{aligned}\left(1 + \frac{2}{n}\right) G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 &= \left(G_2^{(\ell)}\right)^2 \left[\left(1 + \frac{2}{n}\right)^{\ell} - 1\right] \\ &= \left(G_2^{(\ell)}\right)^2 \left[\frac{2}{n}\ell + O\left(\frac{1}{n^2}\right)\right]\end{aligned}$$

Com isso, concluímos que

$$\mathbb{E}\left(\left(\mathcal{O}^{(\ell)} - G_2^{(\ell)}\right)^2\right) = \left(1 + \frac{2}{n}\right) G_4^{(\ell)} - \left(G_2^{(\ell)}\right)^2 \qquad (3.33)$$
$$= \frac{2\ell}{n}\left(G_2^{(\ell)}\right)^2 + O\left(\frac{1}{n^2}\right)$$

Assim, a escala emergente $\frac{\ell}{n}$ mede a magnitude do desvio do observável $\mathcal{O}^{(\ell)}$ de sua média $G_2^{(\ell)}$.

- Pra calcular o correlator de 6 pontos, precisamos do *fator Wick 6*.
- Como vou construir esse monstro de 15 termos?
- Usando a ordem dentro de cada par.
- Fixo o par (1,2) e faço os 3 agrupamentos dos indices 3, 4, 5 6.
- Passo pro par (1,3) e faço os 3 agrupamentos dos índices 2, 4, 5, 6.
- E assim por diante.

$$\mathrm{Wick}_6 = \begin{array}{lll} +\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{4}}\delta_{\widehat{5}\widehat{6}} & +\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{5}}\delta_{\widehat{4}\widehat{6}} & +\delta_{\widehat{1}\widehat{2}}\delta_{\widehat{3}\widehat{6}}\delta_{\widehat{4}\widehat{5}} \\ +\delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{4}}\delta_{\widehat{5}\widehat{6}} & +\delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{5}}\delta_{\widehat{4}\widehat{6}} & +\delta_{\widehat{1}\widehat{3}}\delta_{\widehat{2}\widehat{6}}\delta_{\widehat{4}\widehat{5}} \\ +\delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{3}}\delta_{\widehat{5}\widehat{6}} & +\delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{5}}\delta_{\widehat{3}\widehat{6}} & +\delta_{\widehat{1}\widehat{4}}\delta_{\widehat{2}\widehat{6}}\delta_{\widehat{3}\widehat{5}} \\ +\delta_{\widehat{1}\widehat{5}}\delta_{\widehat{2}\widehat{3}}\delta_{\widehat{4}\widehat{6}} & +\delta_{\widehat{1}\widehat{5}}\delta_{\widehat{2}\widehat{4}}\delta_{\widehat{3}\widehat{6}} & +\delta_{\widehat{1}\widehat{5}}\delta_{\widehat{2}\widehat{6}}\delta_{\widehat{3}\widehat{4}} \\ +\delta_{\widehat{1}\widehat{6}}\delta_{\widehat{2}\widehat{3}}\delta_{\widehat{4}\widehat{5}} & +\delta_{\widehat{1}\widehat{6}}\delta_{\widehat{2}\widehat{4}}\delta_{\widehat{3}\widehat{5}} & +\delta_{\widehat{1}\widehat{6}}\delta_{\widehat{2}\widehat{5}}\delta_{\widehat{3}\widehat{4}} \end{array}$$

Assim, o correlator de 6 pontos é dado por

$$\mathbb{E}\left(z^{(\ell+1)}_{\widehat{1}}z^{(\ell+1)}_{\widehat{2}}z^{(\ell+1)}_{\widehat{3}}z^{(\ell+1)}_{\widehat{4}}z^{(\ell+1)}_{\widehat{5}}z^{(\ell+1)}_{\widehat{6}}\right) =$$

$$= \sum_{\substack{\underline{k}=1 \\ k=1\ldots 6}}^{n_\ell} \mathbb{E}\left(W^{(\ell+1)}_{\widehat{1}\underline{1}}W^{(\ell+1)}_{\widehat{2}\underline{2}}W^{(\ell+1)}_{\widehat{3}\underline{3}}W^{(\ell+1)}_{\widehat{4}\underline{4}}W^{(\ell+1)}_{\widehat{5}\underline{5}}W^{(\ell+1)}_{\widehat{6}\underline{6}}\right)$$

$$\mathbb{E}\left(z^{(\ell)}_{\underline{1}}z^{(\ell)}_{\underline{2}}z^{(\ell)}_{\underline{3}}z^{(\ell)}_{\underline{4}}z^{(\ell)}_{\underline{5}}z^{(\ell)}_{\underline{6}}\right) =$$

$$= \frac{C_W^3}{n_\ell^3}(\mathrm{Wick}_6)\sum_{\mu,\nu,\kappa=1}^{n_\ell}\mathbb{E}\left(z^{(\ell)}_{\mu}z^{(\ell)}_{\mu}z^{(\ell)}_{\nu}z^{(\ell)}_{\nu}z^{(\ell)}_{\kappa}z^{(\ell)}_{\kappa}\right) \quad (3.36)$$

Novamente, assumimos que o correlator de 6 pontos tem a forma

$$\mathbb{E}\left(z_{\widehat{1}}^{(\ell)} z_{\widehat{2}}^{(\ell)} z_{\widehat{3}}^{(\ell)} z_{\widehat{4}}^{(\ell)} z_{\widehat{5}}^{(\ell)} z_{\widehat{6}}^{(\ell)}\right) = (\mathrm{Wick}_6)\, G_6^{(\ell)} \tag{3.37}$$

Para calcular $G_6^{(\ell)}$, vamos usar a mesma ideia que usamos para o correlator de 4 pontos, e precisamos calcular

$$\sum_{\mu,\nu,\kappa=1}^{n_\ell} \mathbb{E}\left(z_\mu^{(\ell)} z_\mu^{(\ell)} z_\nu^{(\ell)} z_\nu^{(\ell)} z_\kappa^{(\ell)} z_\kappa^{(\ell)}\right)$$

Fazendo $\mu = \widehat{1} = \widehat{2}$, $\nu = \widehat{3} = \widehat{4}$, $\kappa = \widehat{5} = \widehat{6}$, na fórmula para Wick_6, obtemos

$$\text{Wick}_6^3 = \begin{array}{lll} +\delta_{\widehat{\mu}\widehat{\mu}}\delta_{\widehat{\nu}\widehat{\nu}}\delta_{\widehat{\kappa}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\mu}}\delta_{\widehat{\nu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\mu}}\delta_{\widehat{\nu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} \\ +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\kappa}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} \\ +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\kappa}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\kappa}} \\ +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\nu}} \\ +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\nu}}\delta_{\widehat{\nu}\widehat{\kappa}} & +\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\mu}\widehat{\kappa}}\delta_{\widehat{\nu}\widehat{\nu}} \end{array}$$

Temos

- 1 termo $\delta_{ii}\delta_{jj}\delta_{kk}$ em verde, cujo valor é sempre 1 em todas as n_ℓ^3 ocorrências;
- 6 termos do tipo $\delta_{ii}\delta_{jk}\delta_{kj}$ em azul, cujo valor é 1 quando $j = k$, ou seja, em n_ℓ^2 ocorrências;
- 8 termos do tipo $\delta_{ij}\delta_{jk}\delta_{ki}$ em preto, cujo valor é 1 quando $i = j = k$, ou seja, em n_ℓ ocorrências.

Assim, (3.36) nos dá a recorrência

$$G_6^{(\ell+1)} = \frac{C_W^3}{n_\ell^3} \sum_{\mu,\nu,\kappa=1}^{n_\ell} \left(\mathrm{Wick}_6^3\right) G_6^{(\ell)} \tag{3.42}$$
$$= \frac{C_W^3}{n_\ell^3} \left(n_\ell^3 + 6n_\ell^2 + 8n_\ell\right) G_6^{(\ell)}$$
$$= C_W^3 \left(1 + \frac{6}{n_\ell} + \frac{8}{n_\ell^2}\right) G_6^{(\ell)}$$

Descendo até a camada 0, obtemos a relação

$$\begin{aligned} G_6^{(\ell)} &= \left(C_W^3\right)^\ell G_6^{(0)} \prod_{\widehat{\ell}=1}^{\ell-1} \left(1 + \frac{6}{n_{\widehat{\ell}}} + \frac{8}{n_{\widehat{\ell}}^2}\right) \\ &= \left(C_W^3\right)^\ell \left(G_2^{(0)}\right)^3 \prod_{\widehat{\ell}=1}^{\ell-1} \left(1 + \frac{6}{n_{\widehat{\ell}}} + \frac{8}{n_{\widehat{\ell}}^2}\right) \\ &= \left(G_2^{(\ell)}\right)^3 \prod_{\widehat{\ell}=1}^{\ell-1} \left(1 + \frac{6}{n_{\widehat{\ell}}} + \frac{8}{n_{\widehat{\ell}}^2}\right) \end{aligned} \tag{3.43}$$

Novamente fazendo todos os n_ℓ iguais a n, obtemos

$$G_6^{(\ell)} = \left(G_2^{(\ell)}\right)^3 \left(1 + \frac{6}{n} + \frac{8}{n^2}\right)^{\ell-1} \tag{3.43'}$$

- Tomando $n \to \infty$, temos $(1 + 6/n + 8/n^2) \to 1$, e o correlator de 6 pontos converge para a distribuição gaussiana.
- Fixando n e fazendo $\ell \to \infty$, o correlator de 6 pontos explode para o infinito, mesmo com a variância $C_W = 1$.