Deep Learning

Inicialização de Pesos: Explosão e Dissipação de Gradientes

Lucas Silveira Kupssinskü

Machine Learning Theory and Applications Laboratory
PUCRS

agosto de 2026

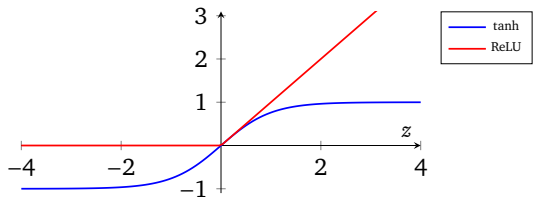
Visão Geral

1. O Problema da Inicialização
2. Análise da Variância
3. Inicializações He e Xavier

Visão Geral

Recap: nem toda ativação preserva o gradiente

- Sigmoid e tanh saturam para $|z|$ grande, com derivada tendendo a zero.
- ReLU e variantes preservam gradiente no lado positivo.
- A combinação **inicialização** $\times$ **ativação** controla a magnitude do sinal que se propaga em forward e em backward.



No início do treino, só os pesos definem o sinal

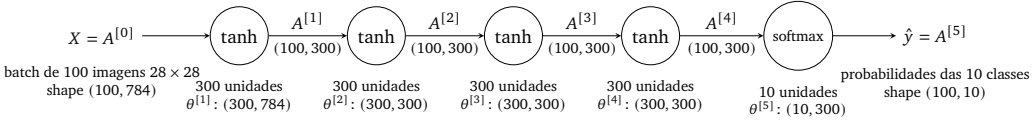
- Antes da primeira atualização, a rede não foi ajustada aos dados.
- A magnitude das ativações em cada camada é determinada inteiramente pelos pesos iniciais e pelas funções de ativação.
- Se a magnitude das ativações crescer ou encolher exponencialmente camada após camada, o gradiente faz o mesmo.

Vanishing e exploding gradients

- **Vanishing gradient**: gradientes encolhem a cada camada, camadas iniciais não recebem sinal de aprendizagem.
- **Exploding gradient**: gradientes crescem a cada camada, atualizações ficam instáveis numericamente.
- Ambos os casos quebram a descida de gradiente.
- A inicialização certa mantém a variância das pré-ativações aproximadamente constante ao longo das camadas.

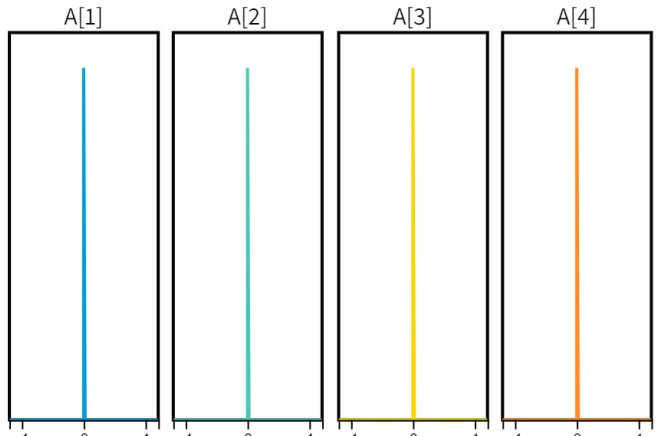
Setup do estudo de caso

- Dataset: MNIST, dígitos manuscritos.
- Arquitetura: MLP com 5 camadas, todas com 300 unidades tanh, saída softmax com 10 classes.
- Tamanho do *batch*: 100 imagens.
- Vamos olhar o histograma das ativações $A^{[1]}, \ldots, A^{[4]}$ em cada camada oculta.



Tentativa 1: $\theta = 0$

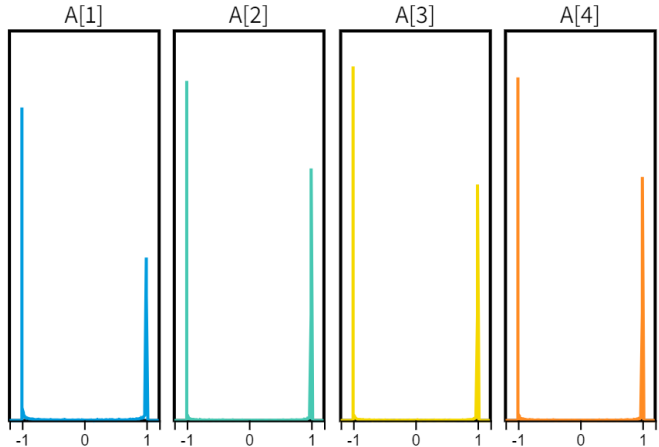
- Pesos iniciais zerados.
- Toda unidade computa a mesma pré-ativação.
- Todas as ativações colapsam em 0.
- Por simetria, qualquer atualização afeta todas as unidades igualmente: a rede nunca quebra a simetria.



Histogramas de $A^{[1]}, \ldots, A^{[4]}$ com $\theta = 0$.

Tentativa 2: $\theta \sim \mathcal{N}(0, 1)$

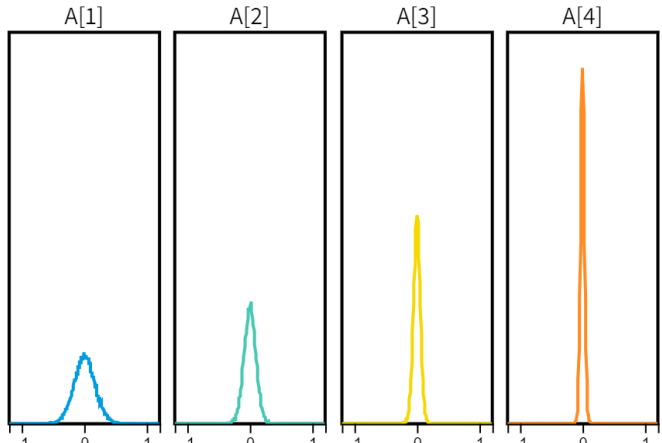
- Pesos com variância 1.
- As pré-ativações de cada camada explodem em magnitude.
- A tanh satura em ±1.
- Em camadas internas, quase todas as unidades estão na região de derivada nula.



Massa concentrada em ±1: zona morta da tanh.

Tentativa 3: $\theta \sim \mathcal{U}(-1, 1)$

- Variância $\sigma_\theta^2 = 1/3$, menor que o caso anterior.
- Mesmo assim, as ativações encolhem a cada camada.
- Em $A^{[4]}$ a distribuição é uma agulha em zero.
- Sintoma de *vanishing* em forward.



Distribuição encolhe a cada camada.

Precisamos calibrar σ_θ^2 por camada

- Inicializações ad-hoc falham de formas opostas dependendo da escala.
- A escala certa depende do número de unidades de entrada de cada camada e da função de ativação utilizada.
- Visualização interativa que motiva a derivação:

Demo interativo: Inicialização de Pesos

Visão Geral

Definição da pré-ativação por camada

Cada exemplo é um vetor linha e o peso multiplica a entrada pela direita:

$$z^{(l)} = \varphi\left(z^{(l-1)}\right)\theta^{(l)\top} + b^{(l)} \qquad \Longrightarrow \qquad z_i^{(l)} = b_i^{(l)} + \sum_{j=1}^{n^{(l-1)}} \varphi\left(z_j^{(l-1)}\right)\theta_{ij}^{(l)}$$

- $z^{(l)}$: pré-ativação da camada l, vetor linha com $n^{(l)}$ componentes.
- $b^{(l)}$: bias, inicializado em 0.
- $\theta^{(l)}$: pesos, matriz de shape $(n^{(l)}, n^{(l-1)})$, com $\theta_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_\theta^2)$.
- φ: ativação (tomamos ReLU como referência).

Hipóteses

$\theta_{ij}^{(l)}$ independentes entre si e independentes de $\varphi(z_j^{(l-1)})$. A distribuição de $z_j^{(l-1)}$ é simétrica em zero.

Esperança: separar bias do somatório

Aplicando a linearidade da esperança:

$$E[z_i^{(l)}] = E[b_i^{(l)}] + E\left[\sum_{j=1}^{n^{(l-1)}} \varphi(z_j^{(l-1)})\, \theta_{ij}^{(l)}\right].$$

- Como inicializamos $b_i^{(l)} = 0$, $E[b_i^{(l)}] = 0$.
- Sobra a esperança do somatório, que reescrevemos como soma das esperanças.

$$E[z_i^{(l)}] = \sum_{j=1}^{n^{(l-1)}} E\left[\varphi(z_j^{(l-1)})\, \theta_{ij}^{(l)}\right].$$

Esperança: independência $\varphi \perp \theta$

Como $\varphi(z_j^{(l-1)})$ é independente de $\theta_{ij}^{(l)}$,

$$E\left[\varphi(z_j^{(l-1)})\,\theta_{ij}^{(l)}\right] = E[\varphi(z_j^{(l-1)})]\,E[\theta_{ij}^{(l)}].$$

- Por construção, $\theta_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_\theta^2)$, logo $E[\theta_{ij}^{(l)}] = 0$.
- O produto inteiro é zero, independentemente de φ.

$$E[z_i^{(l)}] = \sum_{j=1}^{n^{(l-1)}} E[\varphi(z_j^{(l-1)})] \cdot 0 = 0.$$

Resultado intermediário

Média das pré-ativações é zero

$$E[z_i^{(l)}] = 0 \quad \text{para toda camada } l.$$

- Vamos usar este resultado para calcular a variância em seguida.
- Como $E[z_i^{(l)}] = 0$, temos $\text{Var}[z_i^{(l)}] = E[(z_i^{(l)})^2]$.

Variância: ponto de partida

$$\sigma^2_{z^{(l)}} = E\left[(z_i^{(l)})^2\right] - \cancel{\left(E[z_i^{(l)}]\right)^2}^{\,0} = E\left[\left(b_i^{(l)} + \sum_j \varphi(z_j^{(l-1)})\,\theta_{ij}^{(l)}\right)^2\right].$$

- Como $b_i^{(l)} = 0$, expandimos sem o termo do bias.
- Cross-terms entre pesos diferentes desaparecem por independência e $E[\theta] = 0$.

$$\sigma^2_{z^{(l)}} = \sum_{j=1}^{n^{(l-1)}} E\left[\varphi(z_j^{(l-1)})^2\right] E\left[(\theta_{ij}^{(l)})^2\right].$$

Identificando σ_θ^2

Como $\theta_{ij}^{(l)} \sim \mathcal{N}(0, \sigma_\theta^2)$ e $E[\theta] = 0$:

$$E\left[\left(\theta_{ij}^{(l)}\right)^2\right] = \mathrm{Var}[\theta_{ij}^{(l)}] = \sigma_\theta^2.$$

Substituindo:

$$\sigma_{z^{(l)}}^2 = \sum_{j=1}^{n^{(l-1)}} E\left[\varphi\left(z_j^{(l-1)}\right)^2\right] \sigma_\theta^2 = \sigma_\theta^2 \sum_{j=1}^{n^{(l-1)}} E\left[\varphi\left(z_j^{(l-1)}\right)^2\right].$$

- Falta calcular $E[\varphi(z)^2]$ para $\varphi = \mathrm{ReLU}$.

Para ReLU sobre z simétrica em zero

Vamos provar (próximo slide) que, se z é simétrica em zero com variância σ_z^2,

$$E[\mathrm{ReLU}(z)^2] = \frac{\sigma_z^2}{2}.$$

- ReLU corta os valores negativos, mantendo metade da massa.
- A simetria garante que a metade positiva tem o mesmo segundo momento da distribuição inteira, dividido por 2.

Variância da pré-ativação em função de σ_θ^2 e $\sigma^2_{z^{(l-1)}}$

Aplicando o resultado anterior, $E[\varphi(z_j^{(l-1)})^2] = \sigma^2_{z^{(l-1)}}/2$ para toda unidade j:

$$\sigma^2_{z^{(l)}} = \sigma_\theta^2 \sum_{j=1}^{n^{(l-1)}} \frac{\sigma^2_{z^{(l-1)}}}{2} = \frac{n^{(l-1)}}{2}\, \sigma_\theta^2\, \sigma^2_{z^{(l-1)}}.$$

Recorrência para a variância

$$\boxed{\sigma^2_{z^{(l)}} = \frac{n^{(l-1)}}{2}\, \sigma_\theta^2\, \sigma^2_{z^{(l-1)}}}$$

- Se a constante $\frac{n^{(l-1)}}{2}\, \sigma_\theta^2$ for maior que 1, exploding.
- Se for menor que 1, vanishing. Queremos exatamente 1.

Exercício: segundo momento de ReLU

Enunciado

Seja a uma variável aleatória contínua, simétrica em torno de $E[a] = 0$, com variância $\mathrm{Var}[a] = \sigma^2$. Mostre que

$$E\left[\mathrm{ReLU}(a)^2\right] = \frac{\sigma^2}{2}.$$

- Lembre que $\mathrm{ReLU}(a) = \max(0, a)$.
- Use $E[a^2] = \mathrm{Var}[a] + (E[a])^2 = \sigma^2$.
- Use a simetria de a em torno de zero.

Solução

Pela definição de ReLU:

$$E[\text{ReLU}(a)^2] = \int_{-\infty}^{\infty} \max(0, a)^2 p(a)\, da = \int_{0}^{\infty} a^2 p(a)\, da.$$

Pela simetria de p em torno de zero, $\int_{-\infty}^{0} a^2 p(a)\, da = \int_{0}^{\infty} a^2 p(a)\, da$. Logo,

$$E[a^2] = 2\int_{0}^{\infty} a^2 p(a)\, da \quad \Longrightarrow \quad \int_{0}^{\infty} a^2 p(a)\, da = \frac{E[a^2]}{2} = \frac{\sigma^2}{2}.$$

$$\boxed{E[\text{ReLU}(a)^2] = \frac{\sigma^2}{2}.}$$

Visão Geral

Condição: variância igual a cada camada

Queremos $\sigma^2_{z^{(l)}} = \sigma^2_{z^{(l-1)}}$ para todo l. Da recorrência,

$$\sigma^2_{z^{(l-1)}} = \frac{n^{(l-1)}}{2}\,\sigma^2_\theta\,\sigma^2_{z^{(l-1)}} \implies \frac{n^{(l-1)}}{2}\,\sigma^2_\theta = 1.$$

He / Kaiming

$$\boxed{\sigma^2_\theta = \frac{2}{n^{(l-1)}}} \qquad \theta^{(l)}_{ij} \sim \mathcal{N}\left(0,\ \frac{2}{n^{(l-1)}}\right).$$

- $n^{(l-1)}$ é o número de *fan-in* da camada (entradas).
- Padrão de fato para ReLU e GELU.

Origem da inicialização He[1]

- Apresentada em 2015 por Kaiming He et al. para redes muito profundas com ReLU.
- Permitiu treinar redes com mais de 30 camadas convolucionais sem *batch norm*.
- Variante uniforme: $\theta \sim \mathcal{U}\left(-\sqrt{6/n},\ \sqrt{6/n}\right)$, que tem a mesma variância $2/n$.

[1] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". Em: ICCV. 2015.

Inicialização Xavier (Glorot)[2]

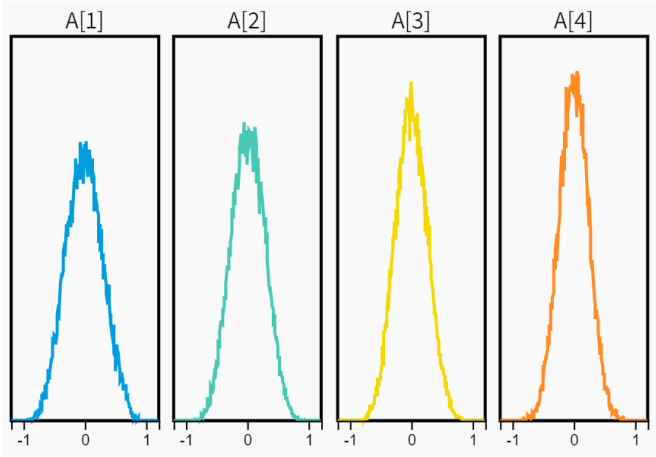
- Antes de He, Glorot e Bengio (2010) fizeram o mesmo argumento para **tanh** ao redor de zero.
- Para tanh, $\varphi(z) \approx z$ perto da origem, então $E[\varphi(z)^2] \approx \sigma_z^2$ sem o fator $1/2$.
- Isso leva a $\sigma_\theta^2 = 1/n^{(l-1)}$.

$$\boxed{\sigma_\theta^2 = \frac{1}{n^{(l-1)}}} \qquad \theta_{ij}^{(l)} \sim \mathcal{N}\left(0,\ \tfrac{1}{n^{(l-1)}}\right).$$

- Padrão para sigmoid, tanh e softmax.

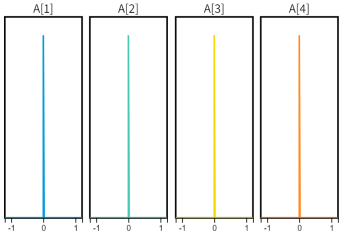
[2] Xavier Glorot e Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". Em: AISTATS. 2010.

Mesmo MLP, agora com inicialização He

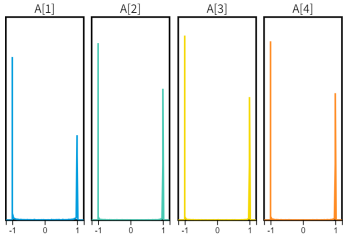


- Histogramas das ativações $A^{[1]}, \ldots, A^{[4]}$ ficam estáveis ao longo das camadas.
- A magnitude não cresce nem encolhe, e a tanh não satura.
- A rede consegue treinar.

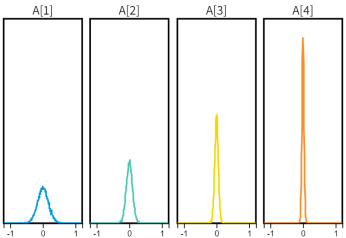
Comparação visual das quatro inicializações



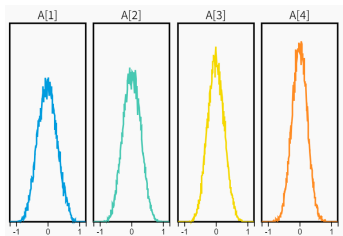
$\theta = 0$



$\theta \sim \mathcal{N}(0, 1)$



$\theta \sim \mathcal{U}(-1, 1)$



He: $\theta \sim \mathcal{N}(0, 2/n)$

O argumento simétrico para o gradiente

- O mesmo raciocínio aplicado à variância dos gradientes em *backward* leva a uma condição análoga.
- Em vez de $n^{(l-1)}$ (*fan-in*), aparece $n^{(l)}$ (*fan-out*): $\sigma_\theta^2 = 2/n^{(l)}$.
- Os dois lados nem sempre são compatíveis (camadas com formas diferentes).

Compromisso fan_avg

Glorot propôs balancear forward e backward usando a média de $n^{(l-1)}$ e $n^{(l)}$:

$$\sigma_\theta^2 = \frac{2}{n^{(l-1)} + n^{(l)}} \quad \text{(Xavier)}$$

$$\sigma_\theta^2 = \frac{4}{n^{(l-1)} + n^{(l)}} \quad \text{(He, fan\_avg)}$$

- Conhecido como *fan_avg* em frameworks.
- Em camadas onde $n^{(l-1)} = n^{(l)}$, recupera as fórmulas originais.

A versão He original (He et al. 2015) usa apenas *fan_in* ($2/n^{(l-1)}$). A média *fan_avg* vem de Glorot/Xavier e é o default do `Dense` no Keras (`glorot_uniform`); o PyTorch usa *fan_in* por padrão (ver próximo slide).

Qual inicialização usar?

Ativação principal	Inicialização recomendada	Variância σ_θ^2
ReLU, Leaky ReLU, ELU	He / Kaiming	$2/n^{(l-1)}$
GELU	He / Kaiming	$2/n^{(l-1)}$
Sigmoid, tanh	Xavier / Glorot	$1/n^{(l-1)}$
Softmax (saída)	Xavier / Glorot	$1/n^{(l-1)}$

- Bias sempre inicializado em 0 (ou em valor pequeno positivo para ReLU, evita unidades mortas no início).

Em PyTorch

- He (Kaiming): `torch.nn.init.kaiming_normal_(tensor, mode='fan_in')`
- Xavier (Glorot): `torch.nn.init.xavier_normal_(tensor)`
- Padrão dos módulos `nn.Linear` e `nn.Conv2d`: Kaiming uniform com `fan_in`, $a = \sqrt{5}$. Bom o suficiente na maioria dos casos.
- Para arquiteturas grandes (transformers, ResNet), inicializações específicas por bloco (DeepNorm, MUP) podem dar resultados ainda melhores.

Leituras Recomendadas

- Capítulo 7 de Simon J. D. Prince. Understanding Deep Learning. MIT Press, 2023
- Xavier (Glorot e Bengio, 2010): (Glorot e Bengio, 2010)
- Kaiming (He et al., 2015): (He et al., 2015)
- Visualização interativa: *lucaskup.github.io/deep-learning-course/weight-init*

Referências I

[1] Xavier Glorot e Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. Em: AISTATS. 2010.

[2] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. Em: ICCV. 2015.

[3] Simon J. D. Prince. Understanding Deep Learning. MIT Press, 2023.