Técnicas de Regularização em Aprendizado Profundo

Autor da Apresentação

Universidade Federal do Rio de Janeiro
UFRJ/COPPE/PEE

October 4, 2025

Agenda

Introdução

Definição

"Any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error."

— Goodfellow et al. Deep Learning

O Problema do Overfitting

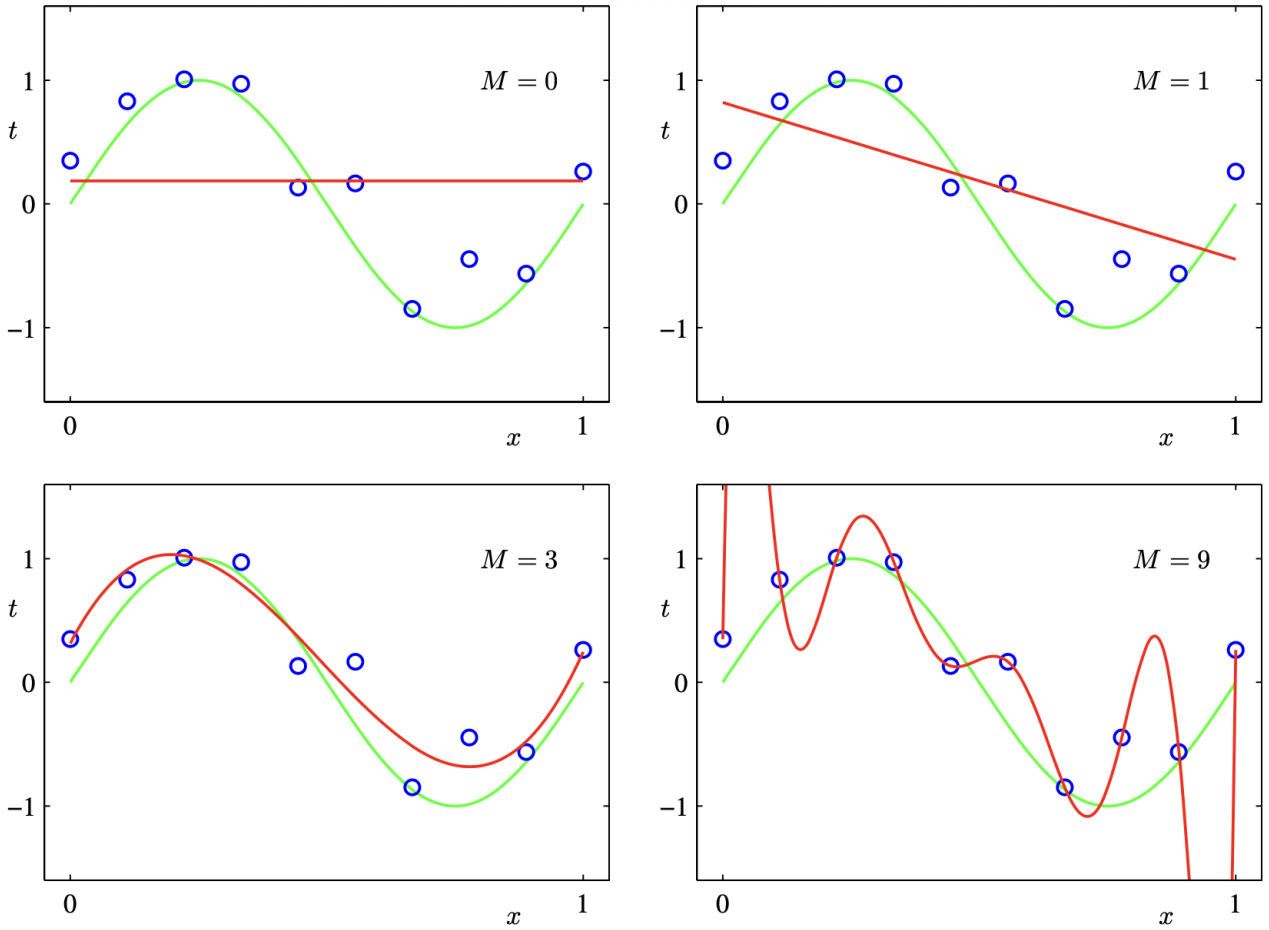
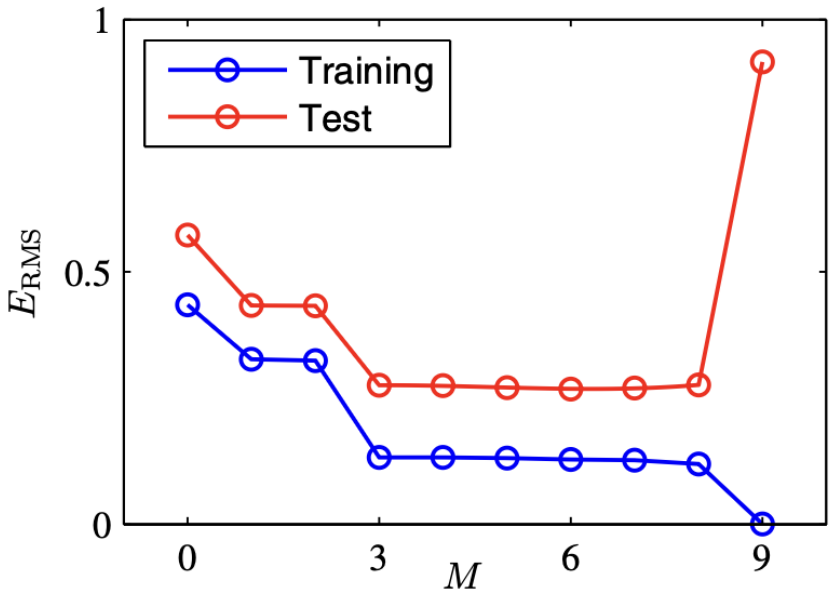


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.



Alternativa 1: aumentar a quantidade de dados no treinamento

"The best way to make a machine learning model generalize better is to train it on more data. Of course, in practice, the amount of data we have is limited." — Goodfellow et al., *Deep Learning*

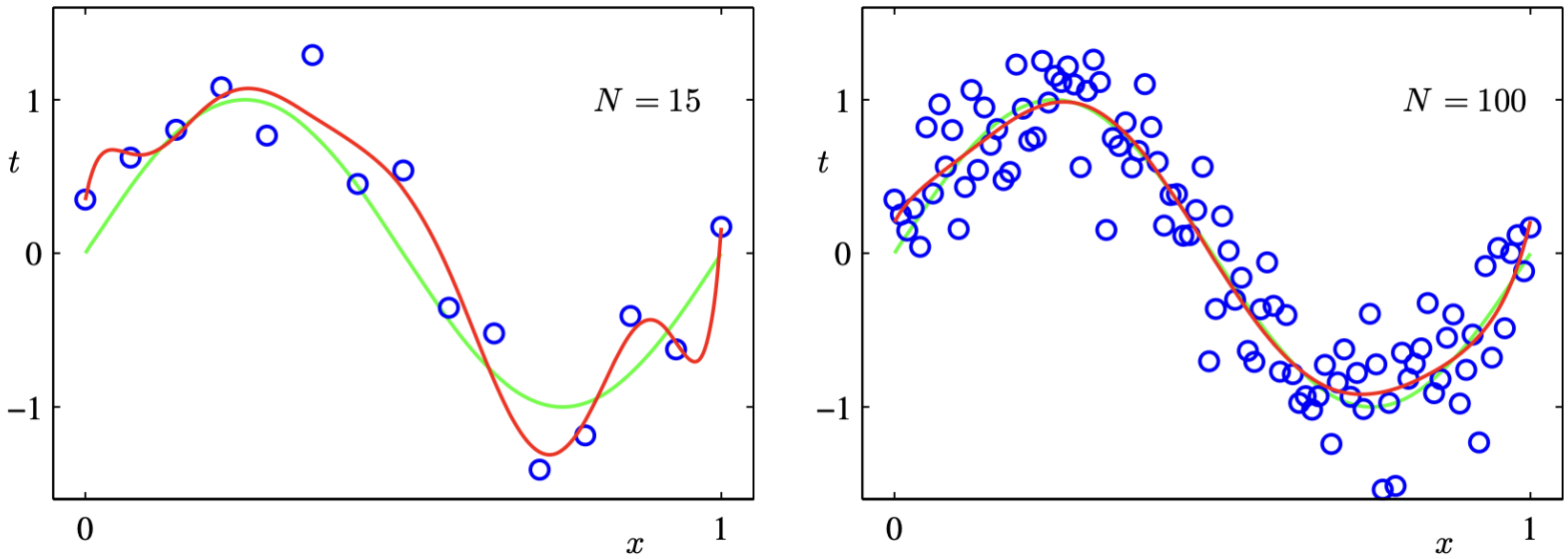


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

Nem sempre é possível. Ex: Experimento genético com 500 camundongos, 20000 a 25000 genomas

Alternativa 2: aplicar regularização

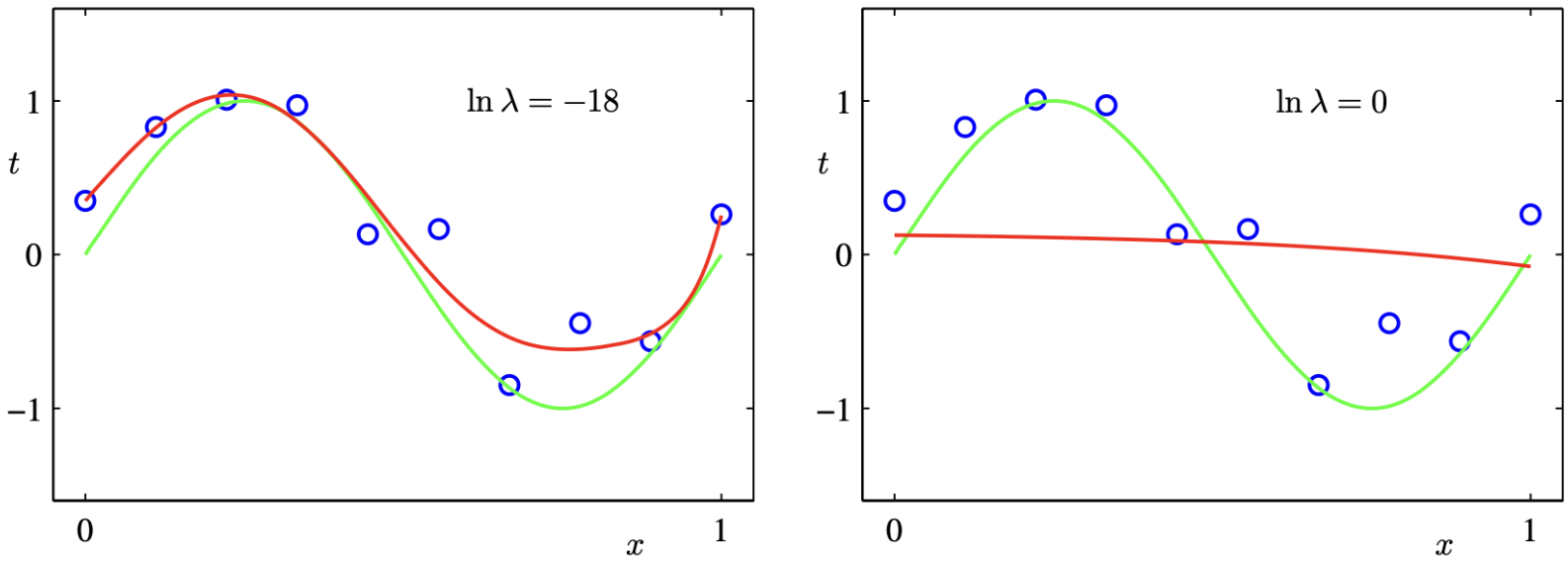


Figure 1.7 Plots of $M = 9$ polynomials fitted to the data set shown in Figure 1.2 using the regularized error function (1.4) for two values of the regularization parameter λ corresponding to $\ln \lambda = -18$ and $\ln \lambda = 0$. The case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln \lambda = -\infty$, is shown at the bottom right of Figure 1.4.

Table 1.2 Table of the coefficients $\mathbf{w}^\star$ for $M = 9$ polynomials with various values for the regularization parameter λ. Note that $\ln \lambda = -\infty$ corresponds to a model with no regularization, i.e., to the graph at the bottom right in Figure 1.4. We see that, as the value of λ increases, the typical magnitude of the coefficients gets smaller.

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Na prática, o valor de λ é obtido pela validação cruzada.

Otimização com Restrições

Penalização da Norma dos Parâmetros

Regularização L1

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \text{Custo}(y_i, f_\theta(x_i)) + \lambda||\theta||_1$$

- Pesos esparsos

- **exatamente zero**

Regularização L2

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \text{Custo}(y_i, f_\theta(x_i)) + \lambda||\theta||_2^2$$

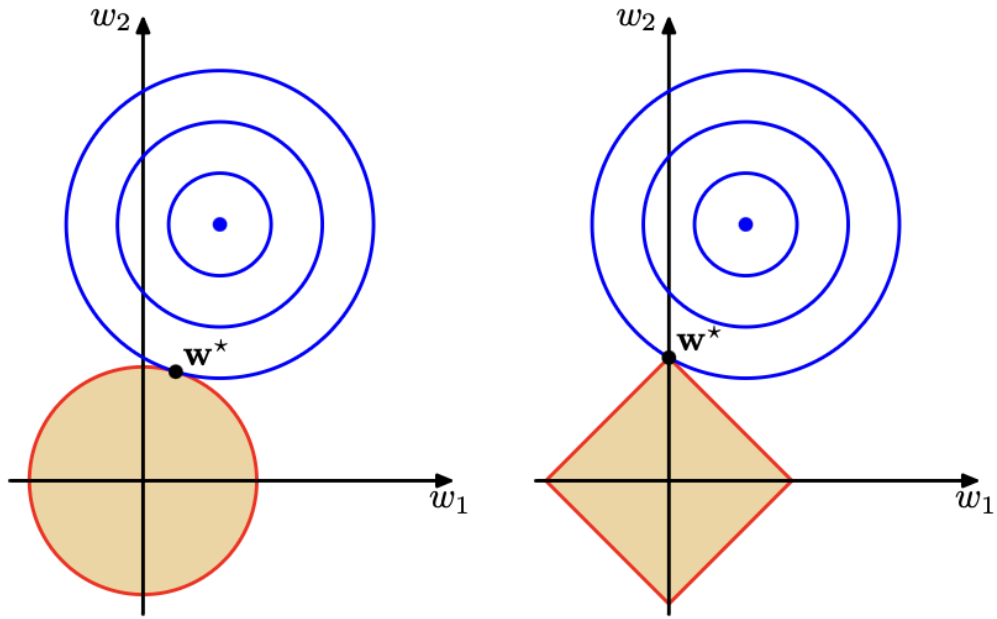
- Pesos pequenos

- **nunca exatamente zero**

"Regularization terms of the form (above) encourage the model weights to have a smaller magnitude and hence introduce a bias towards functions that vary more slowly with changes in the inputs." — Bishop et al., Deep Learning

Interpretação Geométrica

Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.



Otimização com Restrições Explícitas

- Quando há restrição, o ponto ótimo é aquele cujo gradiente da função de custo é **paralelo** ao da restrição
- Multiplicador de Lagrange encontra esse ponto
- A penalização da norma dos parâmetros é uma forma de **restrição suave** aplicada aos pesos:
  - Otimização não-convexa, presa em mínimos locais
  - Neurônios "mortos" em decorrência de pesos muito pequenos

Alternativas:

- **restrição explícita** com **reprojeção**
  - Restrição não "encoraja" pesos próximos da origem
  - Permite pesos maiores $\rightarrow$Gradiente maior $\rightarrow$
  - Mais stabilidade na otimização
- **Compartilhamento de parâmetros**:
  - Diminui os graus de liberdade
  - ex: Redes Convolucionais
- **Compartilhamento suave**: Termo de regularização que "encoraja" grupos de parâmetros a terem valores similares.

Representações Esparsas

Data Augmentation

Aumento de Dados

- Em determinadas tarefas, a predição deve ser **equivariante**. Ex: segmentação em objetos com translação
- Em outras, a predição deve ser **invariante** a uma ou mais transformações nos dados de entrada: translação, tamanho, rotação, **ruído**, etc.
- Formas de tornar o modelo invariante a transformações:
  - Pre-processamento: gerar *features* invariantes às transformações.
  - Regularização: penaliza alterações na saída do modelo para uma mesma entrada com as transformações aplicadas.
  - **Aumentar o conjunto de dados de treinamento**
  - Modificar a estrutura da rede

Invariância

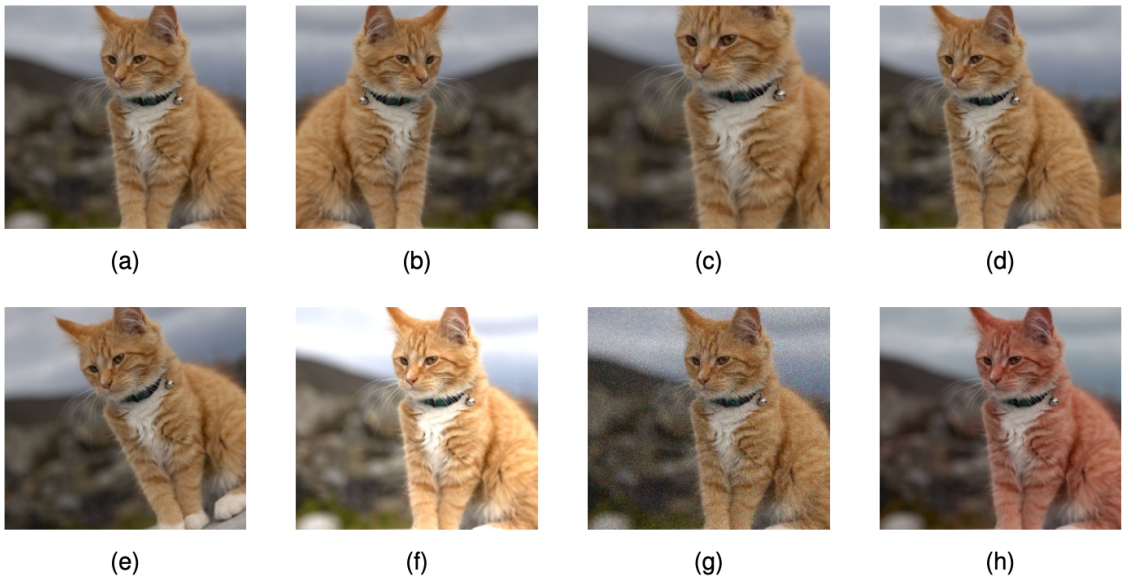
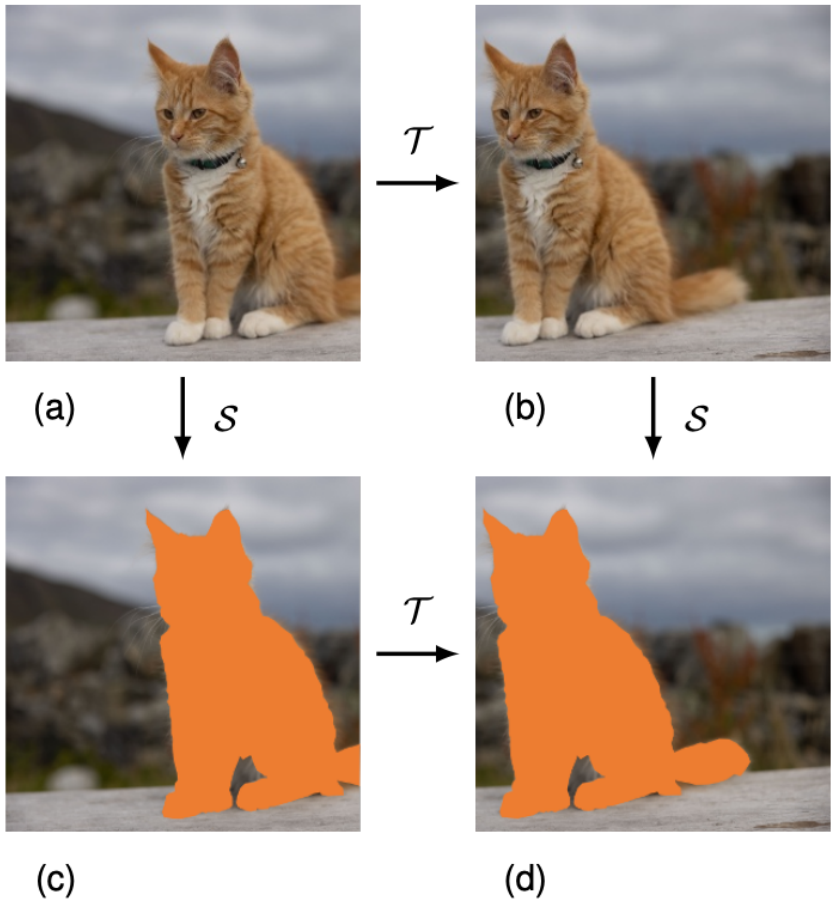


Figure 9.1 Illustration of data set augmentation, showing (a) the original image, (b) horizontal inversion, (c) scaling, (d) translation, (e) rotation, (f) brightness and contrast change, (g) additive noise, and (h) colour shift.

- Reconhecimento de imagem, reconhecimento de fala: funciona
- OCR: cuidado (b e d, 6 e 9)

Equivariância

Figure 9.2 Illustration of equivariance, corresponding to (9.2). If an image (a) is first translated to give (b) and then segmented to give (d), the result is the same as if the image is first segmented to give (c) and then translated to give (d).



Curvas de Aprendizado

Convenção "clássica": Encerramento Antecipado

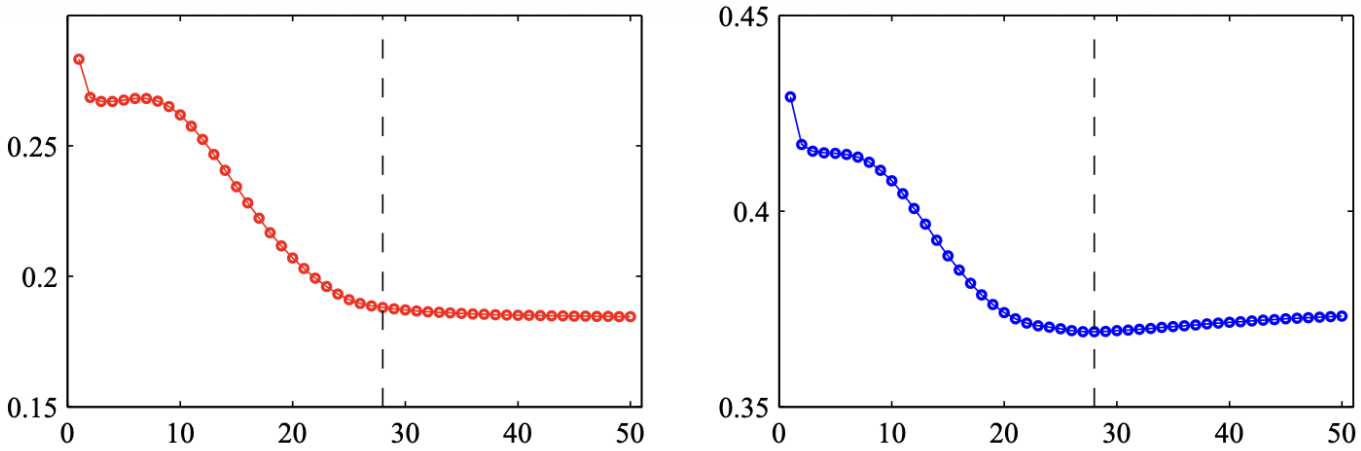


Figure 9.7 An illustration of the behaviour of training set error (left) and validation set error (right) during a typical training session, as a function of the iteration step, for the sinusoidal data set. To achieve the best generalization performance , the training should be stopped at the point shown by the vertical dashed lines, corresponding to the minimum of the validation set error.

- **Trade-off Viés-Variância:**
  - Poucos parâmetros: erro de teste alto (viés alto)
  - Mais parâmetros: erro de teste diminui
  - Parâmetros demais: erro de teste aumenta novamente (alta variância)

Crença clássica: número de parâmetros deve ser limitado ao tamanho do conjunto de dados

Convenção "moderna": **Double Descent**

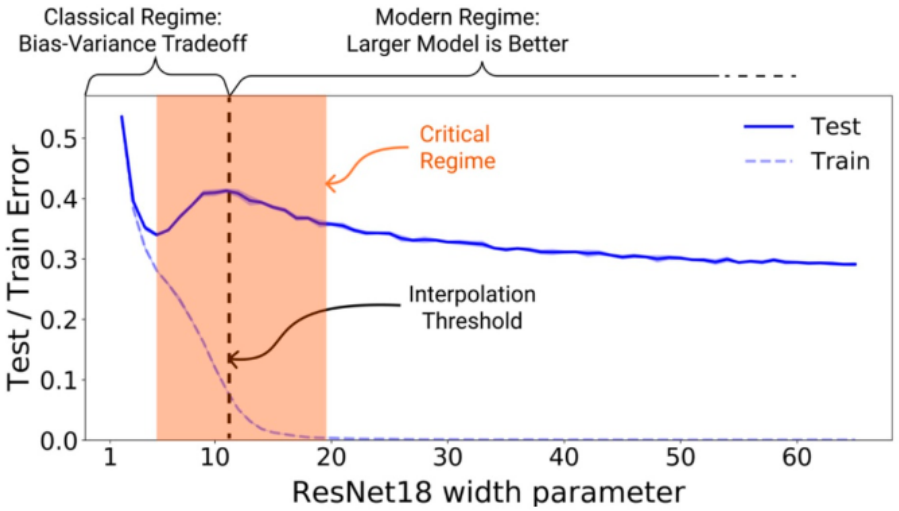


Figure 9.9 Plot of training set and test set errors for a large neural network model called ResNet18 trained on an image classification problem versus the complexity of a model. The horizontal axis represents a hyperparameter governing the number of hidden units and hence the overall number of weights and biases in the network. The vertical dashed line, labelled 'interpolation threshold' indicates the level of model complexity at which the model is capable, in principle, of achieving zero error on the training set. [From Nakkiran *et al.* (2019) with permission.]

- **Redes profundas:** bom desempenho mesmo quando a quantidade de parâmetros excede em muito o necessário para ajustar aos dados de treinamento

Convenção "moderna": **Double Descent**

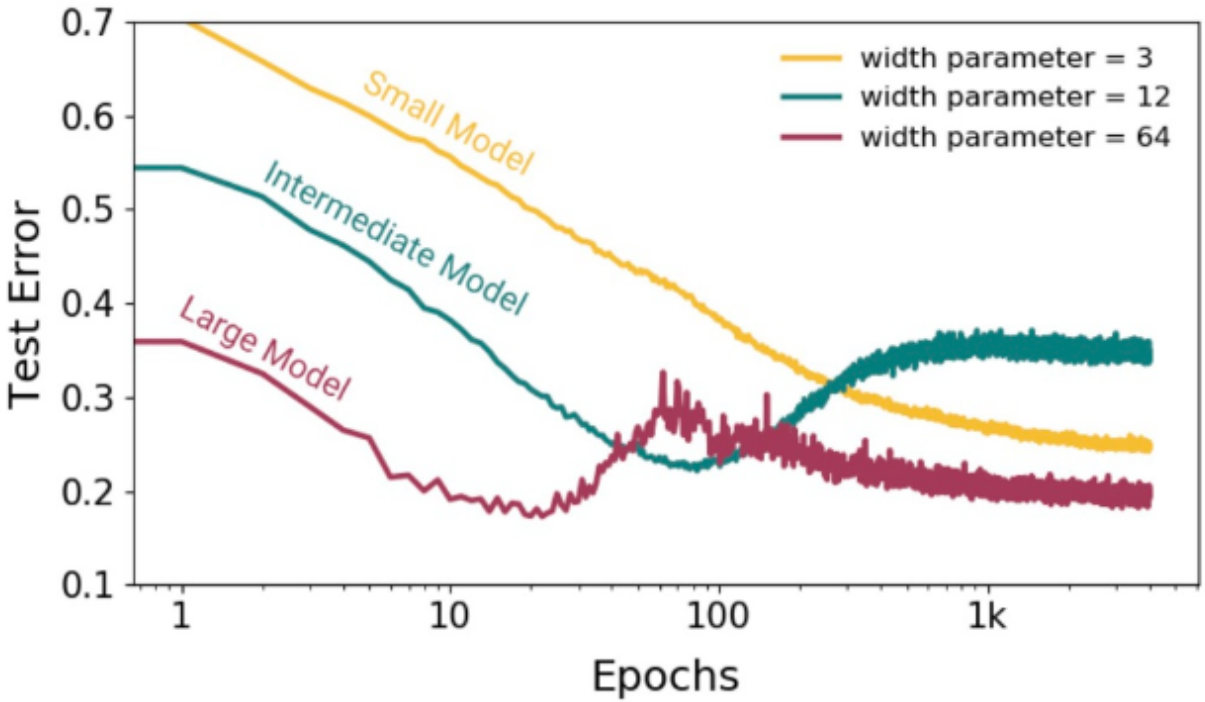


Figure 9.10 Plot of test set error versus number of epochs of gradient descent training for ResNet18 models of various sizes. The effective model complexity increases with the number of training epochs, and the double descent phenomenon is observed for a sufficiently large model. [From Nakkiran *et al.* (2019) with permission.]

Complexidade efetiva do modelo: quantidade máxima de dados de treinamento tal que o modelo atinja erro zero (limiar de interpolação). Dali em diante, a descendência dupla ocorre quando a complexidade do modelo excede esse limiar.

Outras Formas de Regularização

Batch Normalization

Bagging e Ensemble de Modelos

Aprendizado Residual

Dropout

Semi-Supervised Learning

Multi-Task Learning

Adversarial Training

Alguns Casos Práticos

Alguns Casos Práticos

Conclusão

Conclusão

Referências

Bishop Goodfellow