

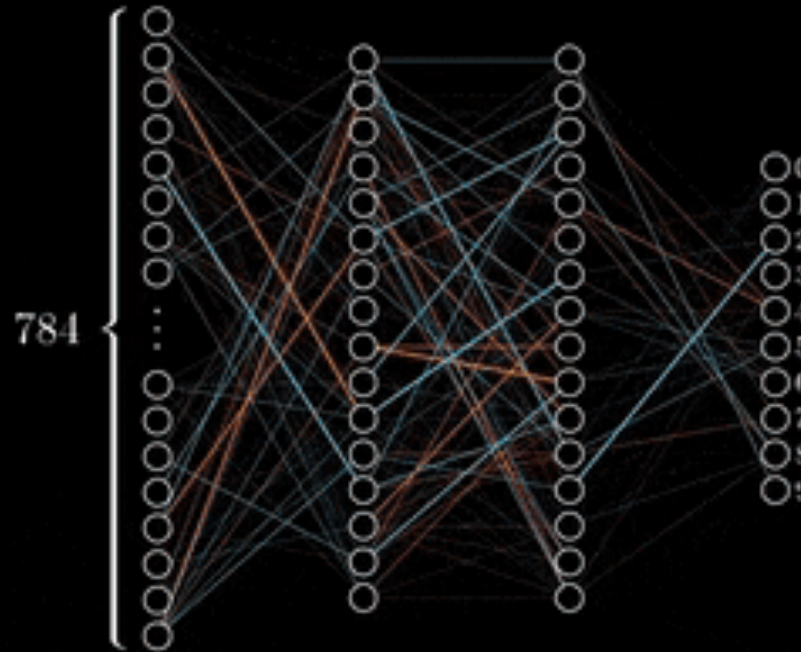
ENM 5310: Data-driven Modeling and Probabilistic Scientific Computing

Lecture #12: Training MLP networks



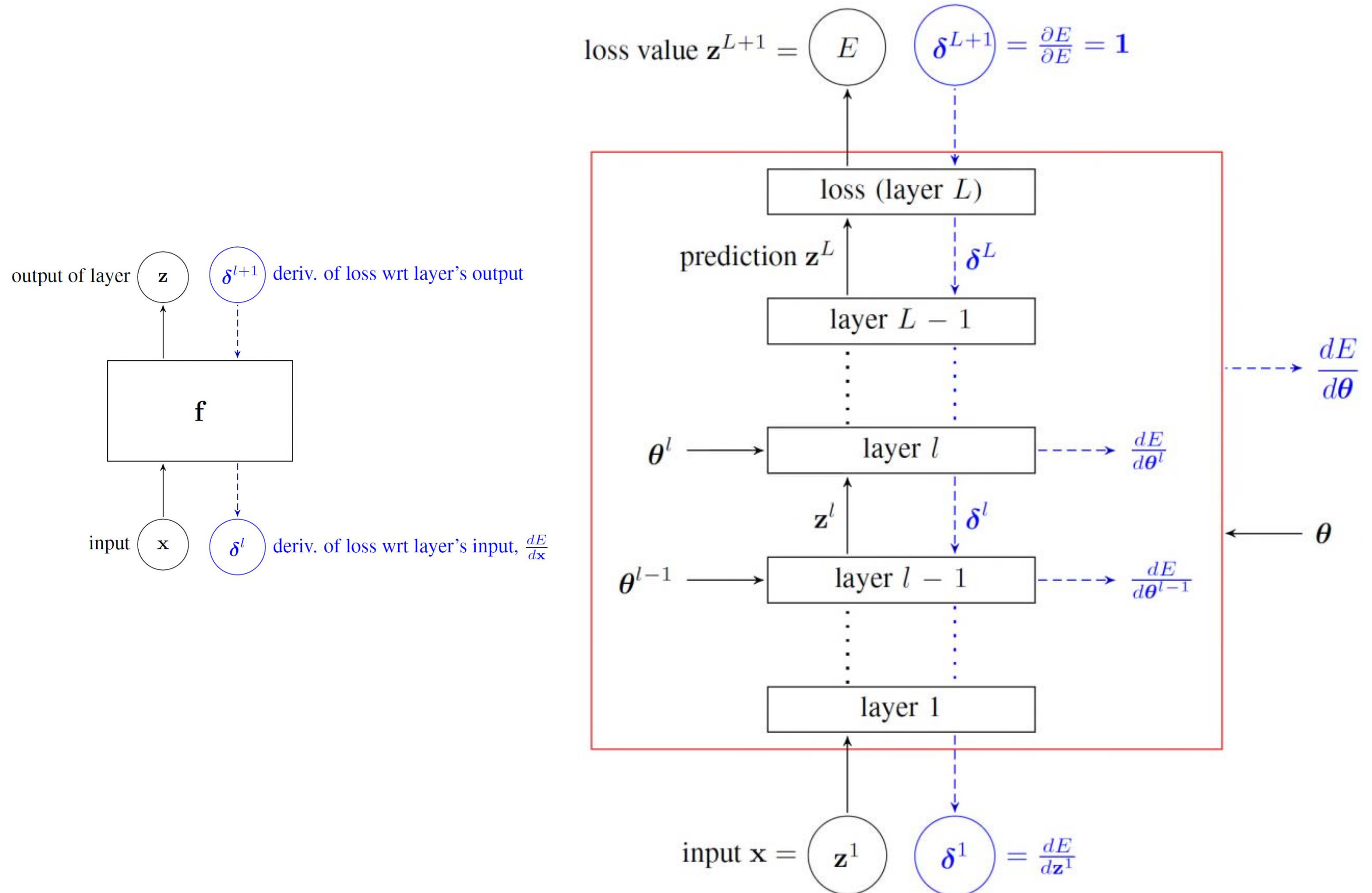
Backpropagation

Training in progress...



$$\begin{aligned}
 G_\ell &= F - Y \in \mathbb{R}^{n \times p_{\ell+1}}, & \frac{\partial \mathcal{L}}{\partial W_\ell} &= A_\ell^T G_\ell \in \mathbb{R}^{p_\ell \times p_{\ell+1}}, & \frac{\partial \mathcal{L}}{\partial b_\ell} &= \mathbf{1}^T G_\ell \in \mathbb{R}^{1 \times p_{\ell+1}}, \\
 G_{\ell-1} &= (1 - A_\ell \odot A_\ell) \odot (G_\ell W_\ell^T) \in \mathbb{R}^{n \times p_\ell}, & \frac{\partial \mathcal{L}}{\partial W_{\ell-1}} &= A_{\ell-1}^T G_{\ell-1} \in \mathbb{R}^{p_{\ell-1} \times p_\ell}, & \frac{\partial \mathcal{L}}{\partial b_{\ell-1}} &= \mathbf{1}^T G_{\ell-1} \in \mathbb{R}^{1 \times p_\ell}, \\
 G_{\ell-2} &= (1 - A_{\ell-1} \odot A_{\ell-1}) \odot (G_{\ell-1} W_{\ell-1}^T) \in \mathbb{R}^{n \times p_{\ell-1}}, & \frac{\partial \mathcal{L}}{\partial W_{\ell-2}} &= A_{\ell-2}^T G_{\ell-2} \in \mathbb{R}^{p_{\ell-2} \times p_{\ell-1}}, & \frac{\partial \mathcal{L}}{\partial b_{\ell-2}} &= \mathbf{1}^T G_{\ell-2} \in \mathbb{R}^{1 \times p_{\ell-1}}, \\
 &\vdots & &\vdots & &\vdots \\
 G_1 &= (1 - A_2 \odot A_2) \odot (G_2 W_2^T) \in \mathbb{R}^{n \times p_2}, & \frac{\partial \mathcal{L}}{\partial W_1} &= A_1^T G_1 \in \mathbb{R}^{p_1 \times p_2}, & \frac{\partial \mathcal{L}}{\partial b_1} &= \mathbf{1}^T G_1 \in \mathbb{R}^{1 \times p_2}, \\
 G_0 &= (1 - A_1 \odot A_1) \odot (G_1 W_1^T) \in \mathbb{R}^{n \times p_1}, & \frac{\partial \mathcal{L}}{\partial W_0} &= X^T G_0 \in \mathbb{R}^{p_0 \times p_1}, & \frac{\partial \mathcal{L}}{\partial b_0} &= \mathbf{1}^T G_0 \in \mathbb{R}^{1 \times p_1}.
 \end{aligned}$$

Modular implementation



Overfitting

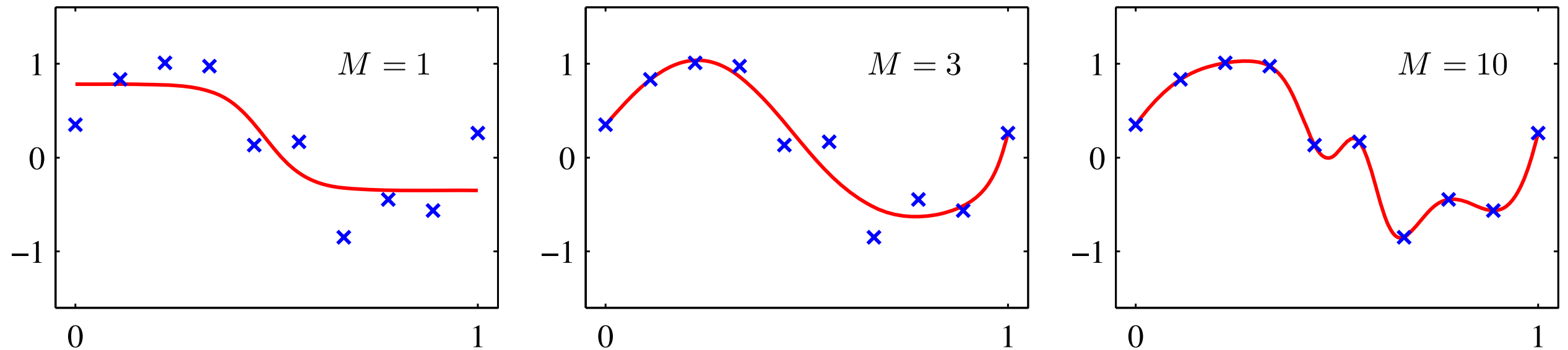
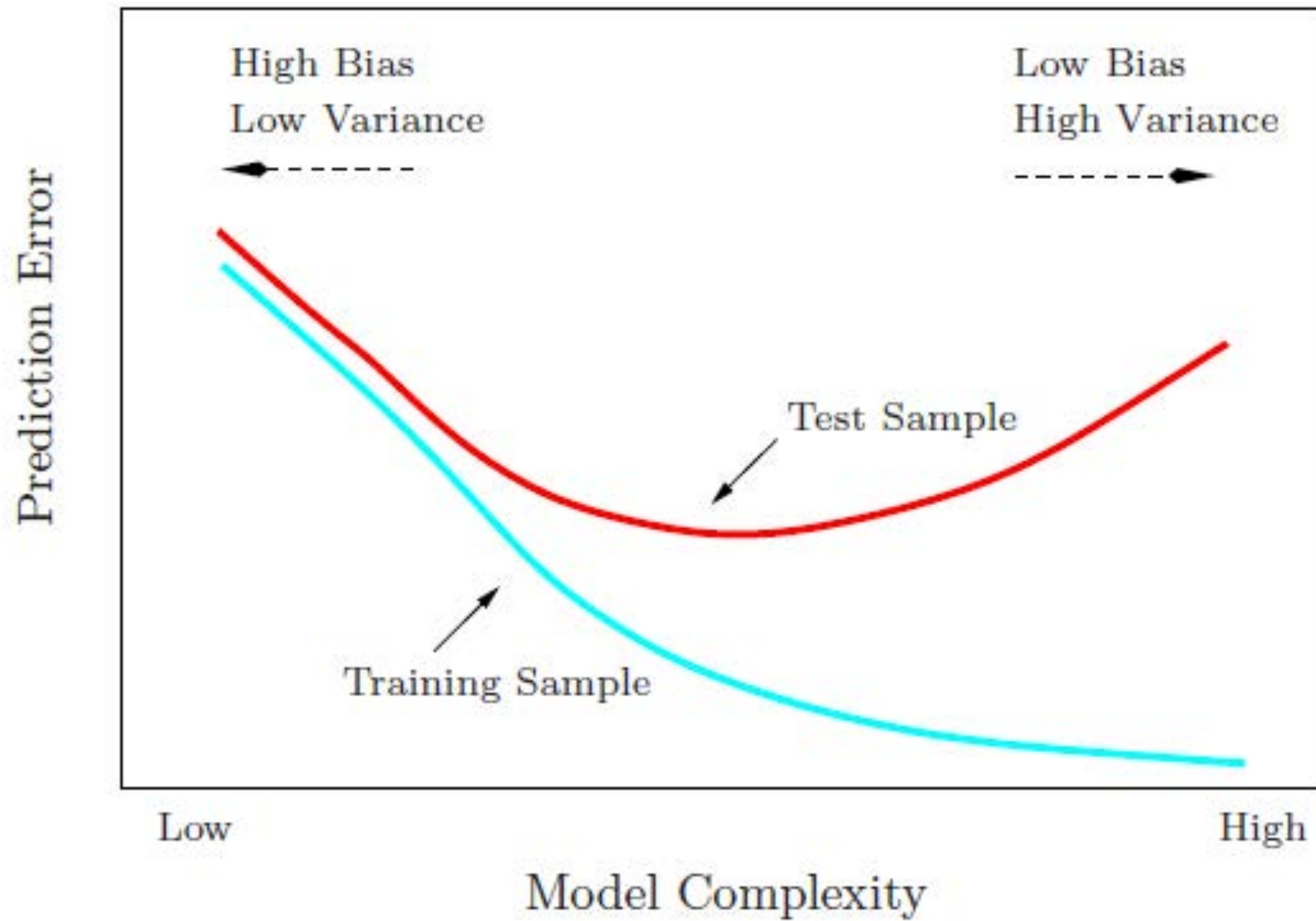
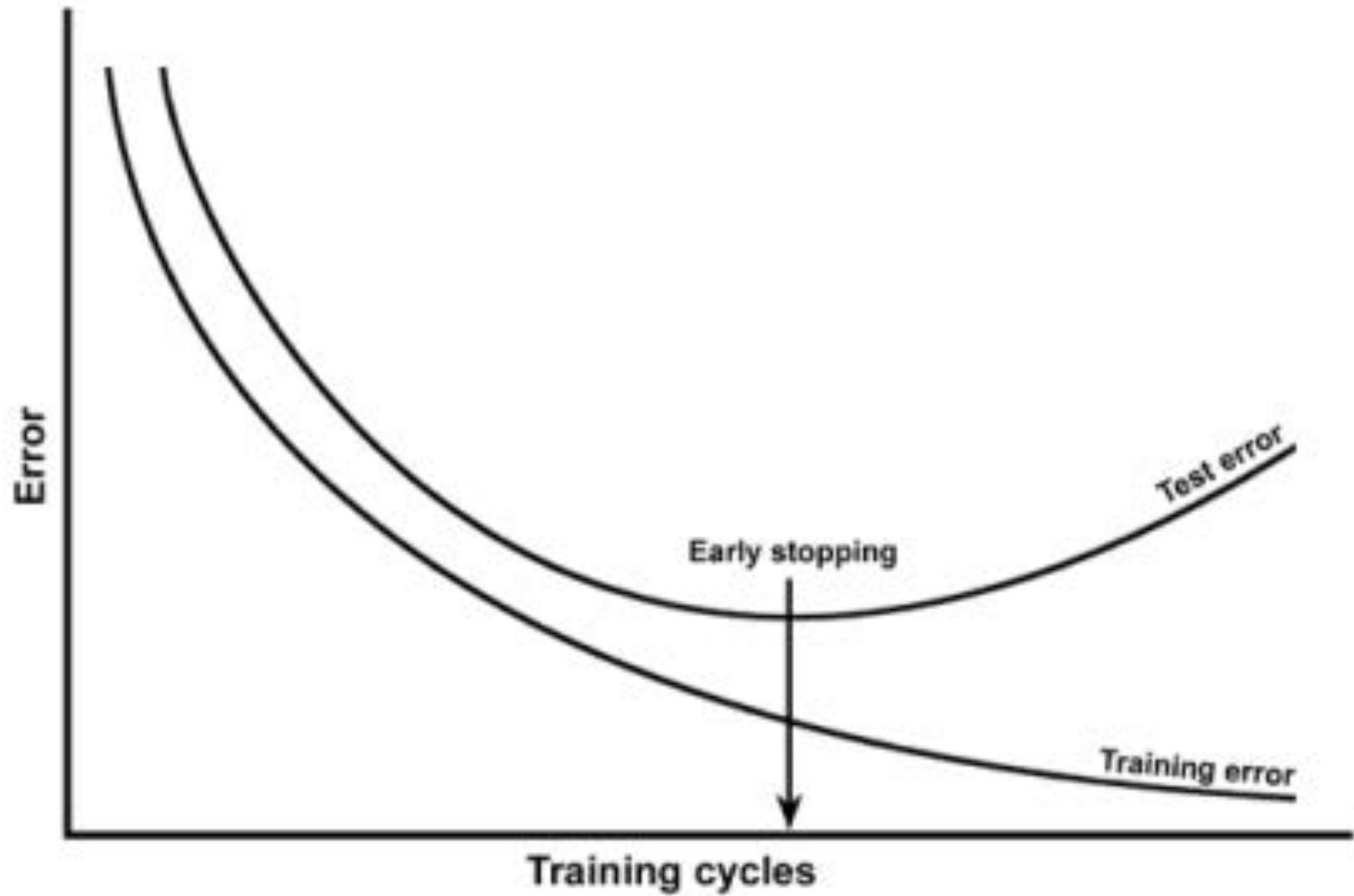


Figure 5.9 Examples of two-layer networks trained on 10 data points drawn from the sinusoidal data set. The graphs show the result of fitting networks having $M = 1$, 3 and 10 hidden units, respectively, by minimizing a sum-of-squares error function using a scaled conjugate-gradient algorithm.

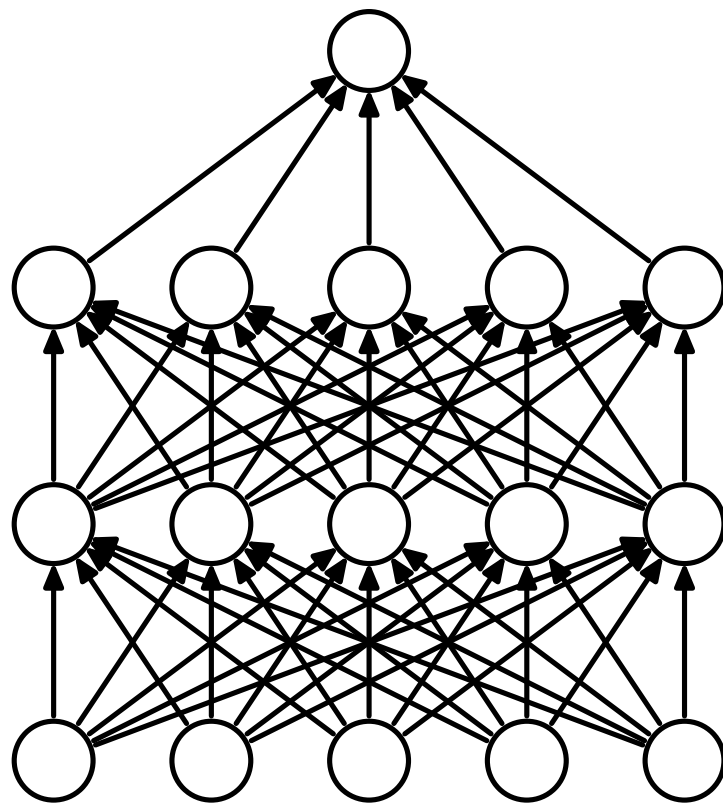
Overfitting



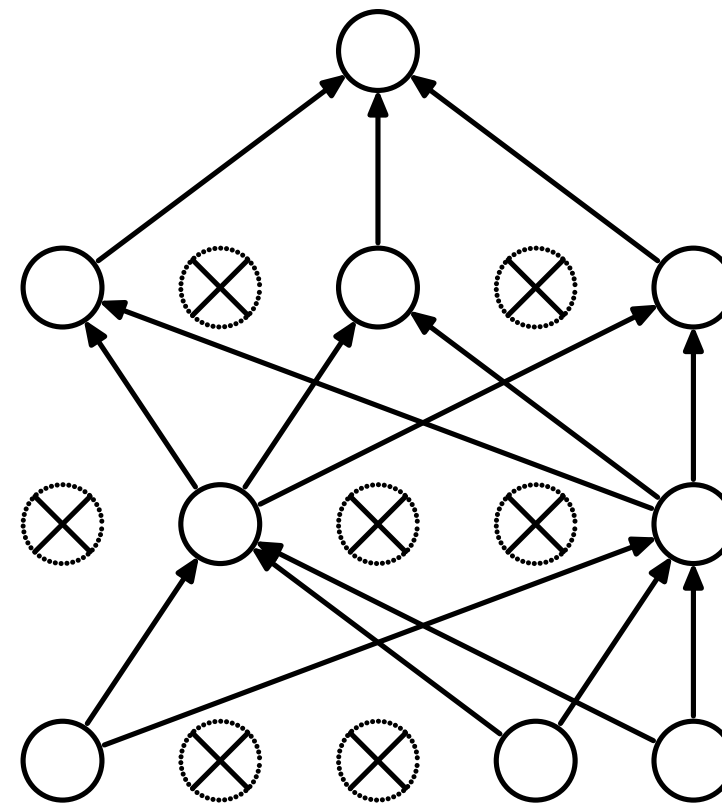
Early stopping



Dropout



(a) Standard Neural Net



(b) After applying dropout.

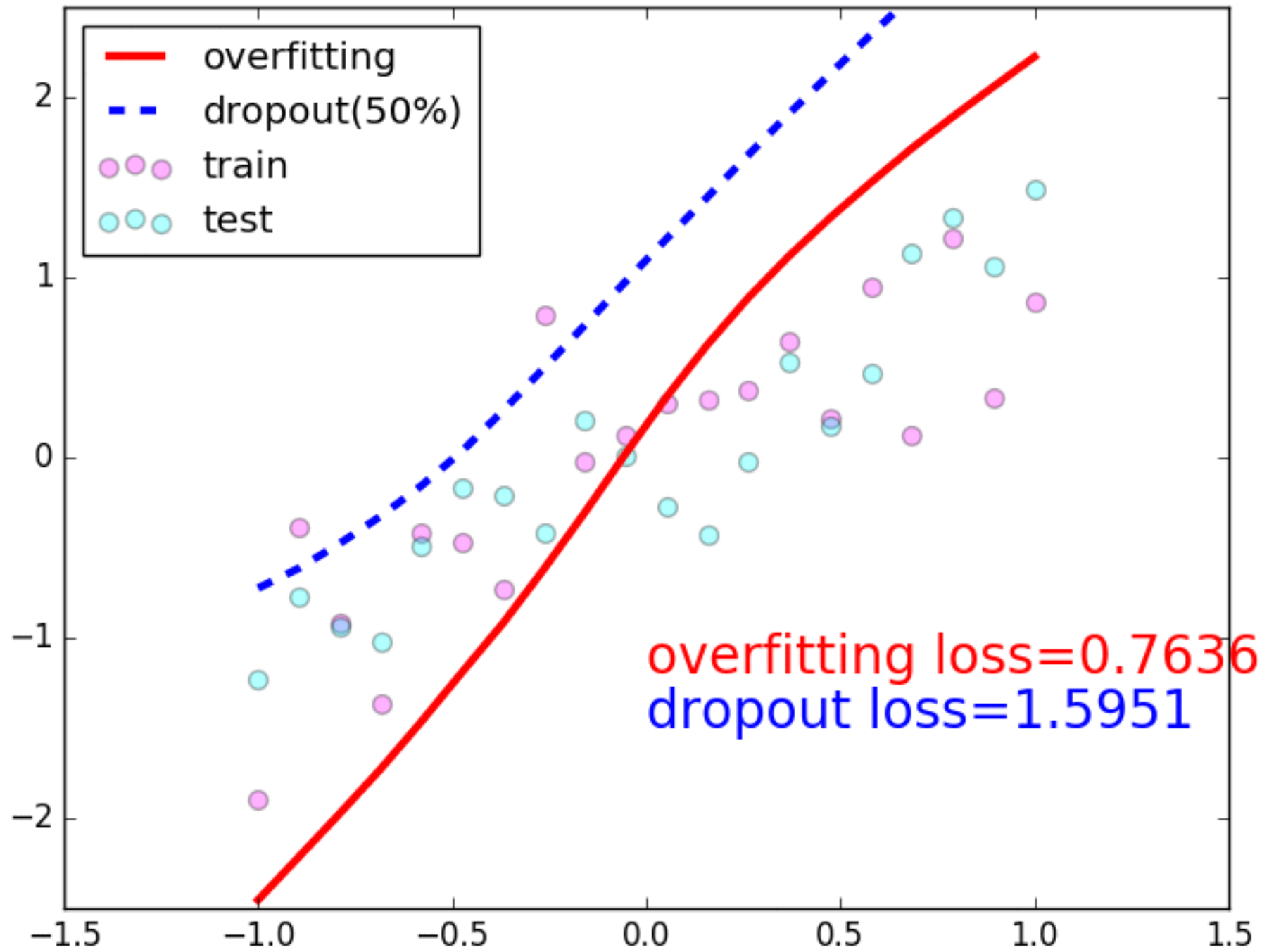
With probability `keep_prob`, outputs the input element scaled up by $1 / \text{keep_prob}$, otherwise outputs 0. The scaling is so that the expected sum is unchanged.

```
for W, b in params:
    outputs = np.dot(inputs, W) + b
    inputs = np.tanh(outputs)
    if dropout_train: inputs *= np.random.binomial([np.ones_like(inputs)], (1-
keep_prob)) [0] / (1-keep_prob)
```

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (pp. 1050-1059).

Dropout



Network initialization

Understanding the difficulty of training deep feedforward neural networks

Xavier Glorot

DIRO, Université de Montréal, Montréal, Québec, Canada

Yoshua Bengio

Whereas before 2006 it appears that deep multi-layer neural networks were not successfully trained, since then several algorithms have been shown to successfully train them, with experimental results showing the superiority of deeper vs less deep architectures. All these experimental results were obtained with new initialization or training mechanisms. Our objective here is to understand better why standard gradient descent from random initialization is doing so poorly with deep neural networks, to better understand these recent relative successes and help design better algorithms in the future. We first observe the influence of the non-linear activations functions. We find that the logistic sigmoid activation is unsuited for deep networks with random initialization because of its mean value, which can drive especially the top hidden layer into saturation. Surprisingly, we find that saturated units can move out of saturation by themselves, albeit slowly, and explaining the plateaus sometimes seen when training neural networks. We find that a new non-linearity that saturates less can often be beneficial. Finally, we study how activations and gradients vary across layers and during training, with the idea that training may be more difficult when the singular values of the Jacobian associated with each layer are far from 1. Based on these considerations, we propose a new initialization scheme that brings substantially faster convergence.

Network initialization

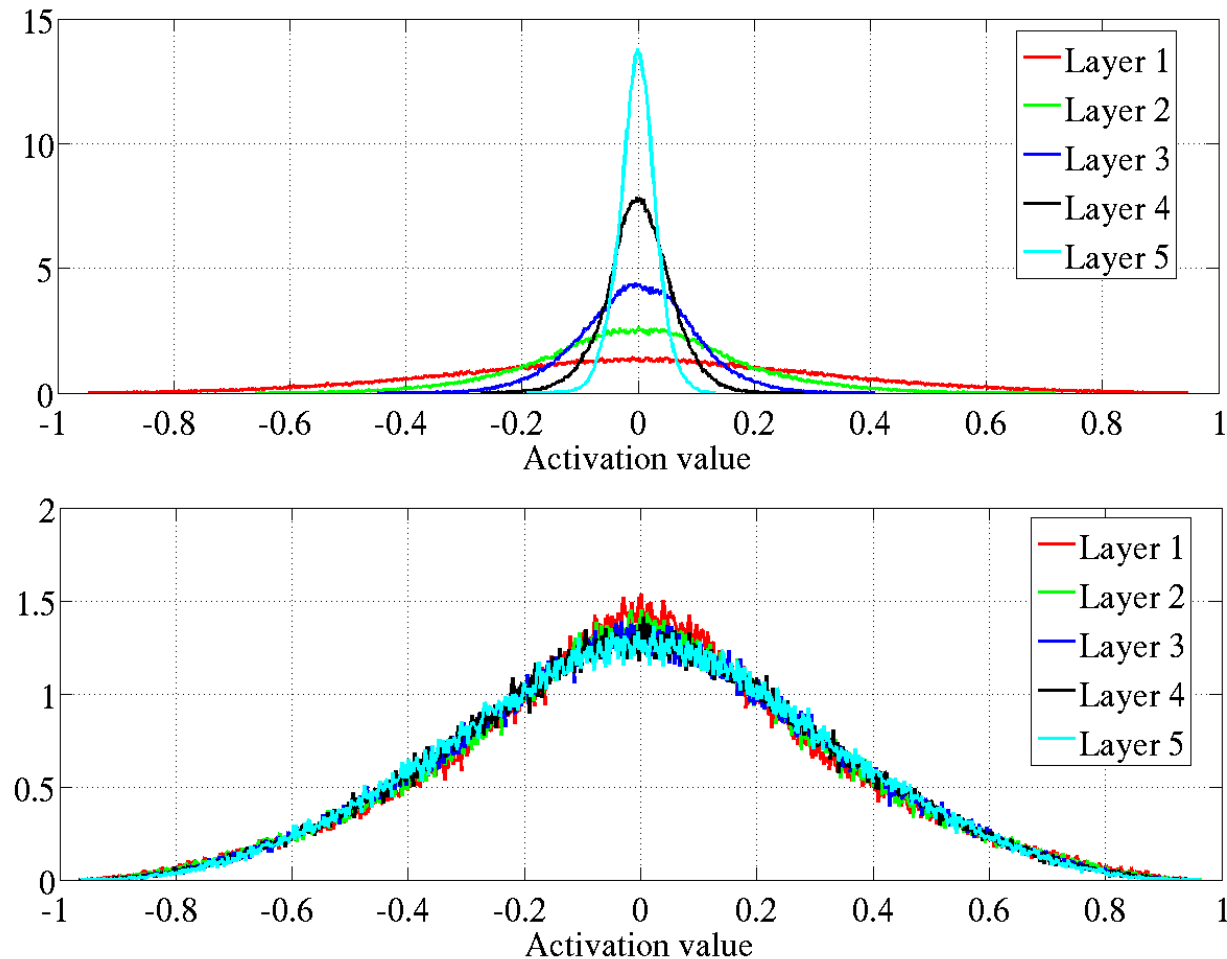


Figure 6: *Activation values normalized histograms with hyperbolic tangent activation, with standard (top) vs normalized initialization (bottom). Top: 0-peak increases for higher layers.*

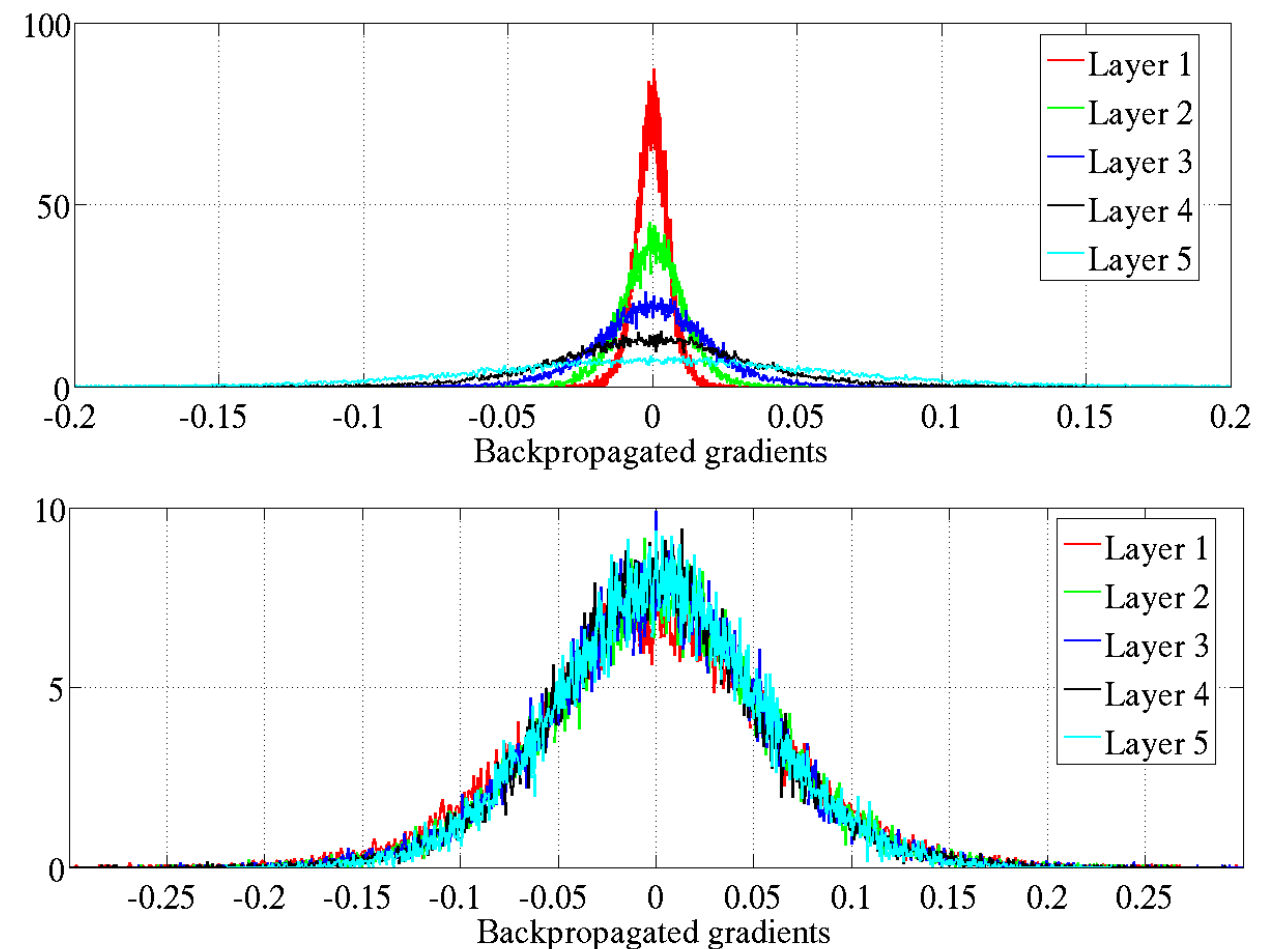


Figure 7: *Back-propagated gradients normalized histograms with hyperbolic tangent activation, with standard (top) vs normalized (bottom) initialization. Top: 0-peak decreases for higher layers.*

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249-256).

Tricks of the trade

Efficient BackProp

Yann LeCun¹, Leon Bottou¹, Genevieve B. Orr², and Klaus-Robert Müller³

¹ Image Processing Research Department AT&T Labs - Research, 100 Schulz Drive,
Red Bank, NJ 07701-7033, USA

² Willamette University, 900 State Street, Salem, OR 97301, USA

³ GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
{yann,leonb}@research.att.com, gorr@willamette.edu, klaus@first.gmd.de

originally published in

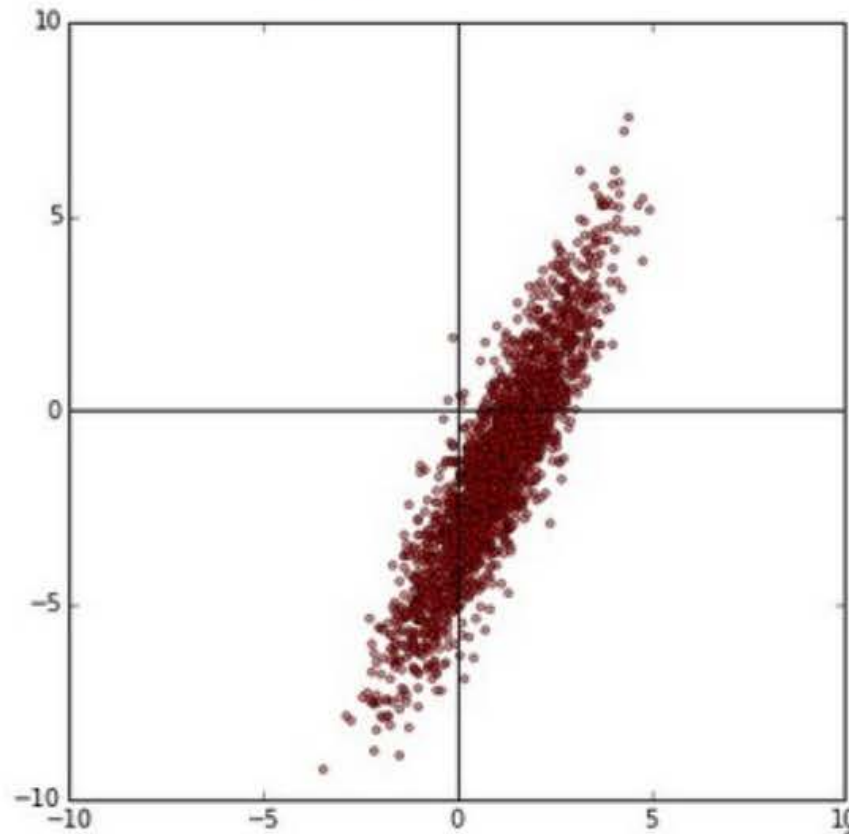
Orr, G. and Müller, K. “Neural Networks: tricks of the trade”,
Springer, 1998.

Abstract. The convergence of back-propagation learning is analyzed so as to explain common phenomenon observed by practitioners. Many undesirable behaviors of backprop can be avoided with tricks that are rarely exposed in serious technical publications. This paper gives some of those tricks, and offers explanations of why they work.

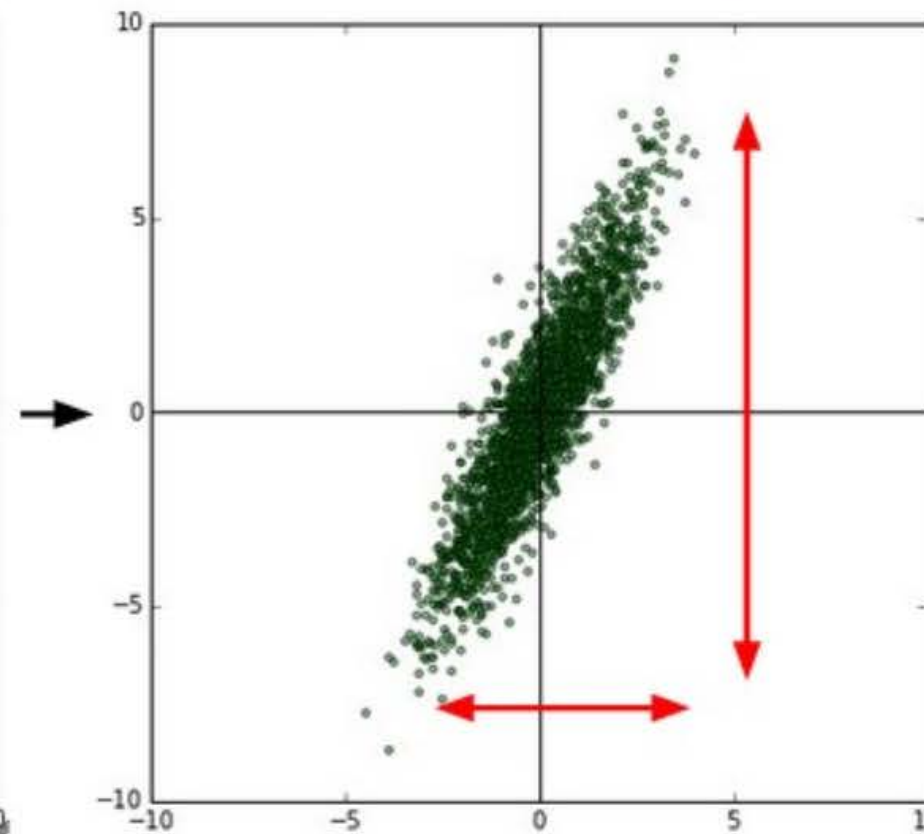
Many authors have suggested that second-order optimization methods are advantageous for neural net training. It is shown that most “classical” second-order methods are impractical for large neural networks. A few methods are proposed that do not have these limitations.

Normalizing the inputs

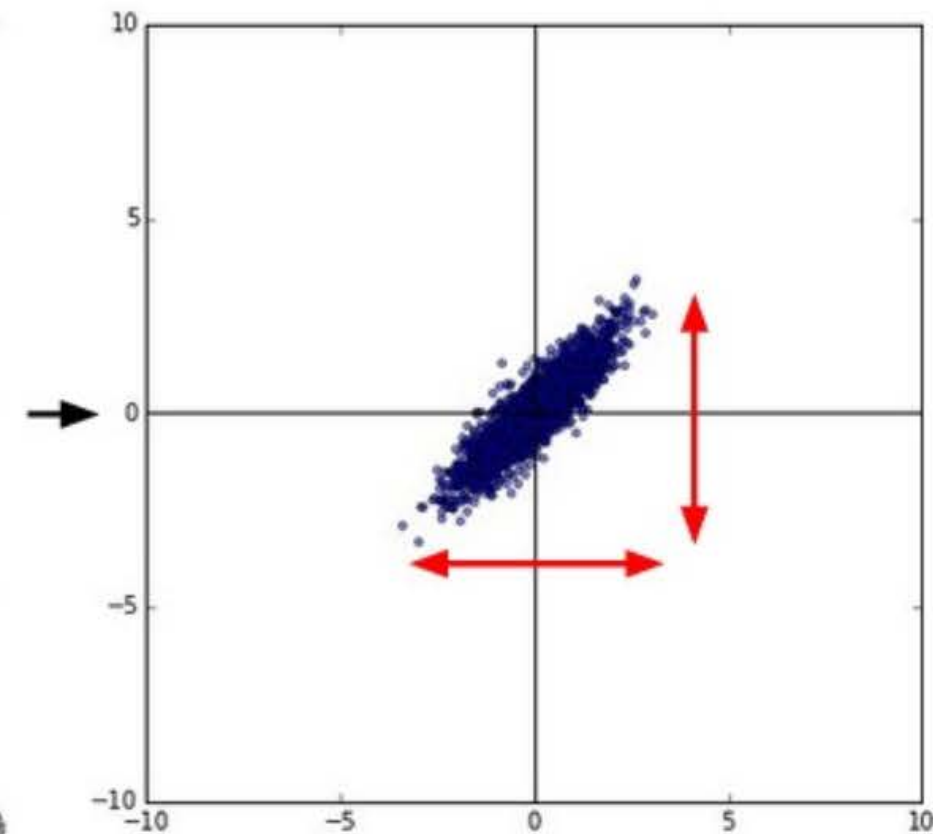
original data



zero-centered data



normalized data



Common data preprocessing pipeline. **Left:** Original toy, 2-dimensional input data. **Middle:** The data is zero-centered by subtracting the mean in each dimension. The data cloud is now centered around the origin. **Right:** Each dimension is additionally scaled by its standard deviation. The red lines indicate the extent of the data - they are of unequal length in the middle, but of equal length on the right.