Regularization for Deep Learning

Lecture slides for Chapter 7 of *Deep Learning*

www.deeplearningbook.org

Ian Goodfellow

2016-09-27

Definition

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

(Goodfellow 2016)

Weight Decay as Constrained Optimization

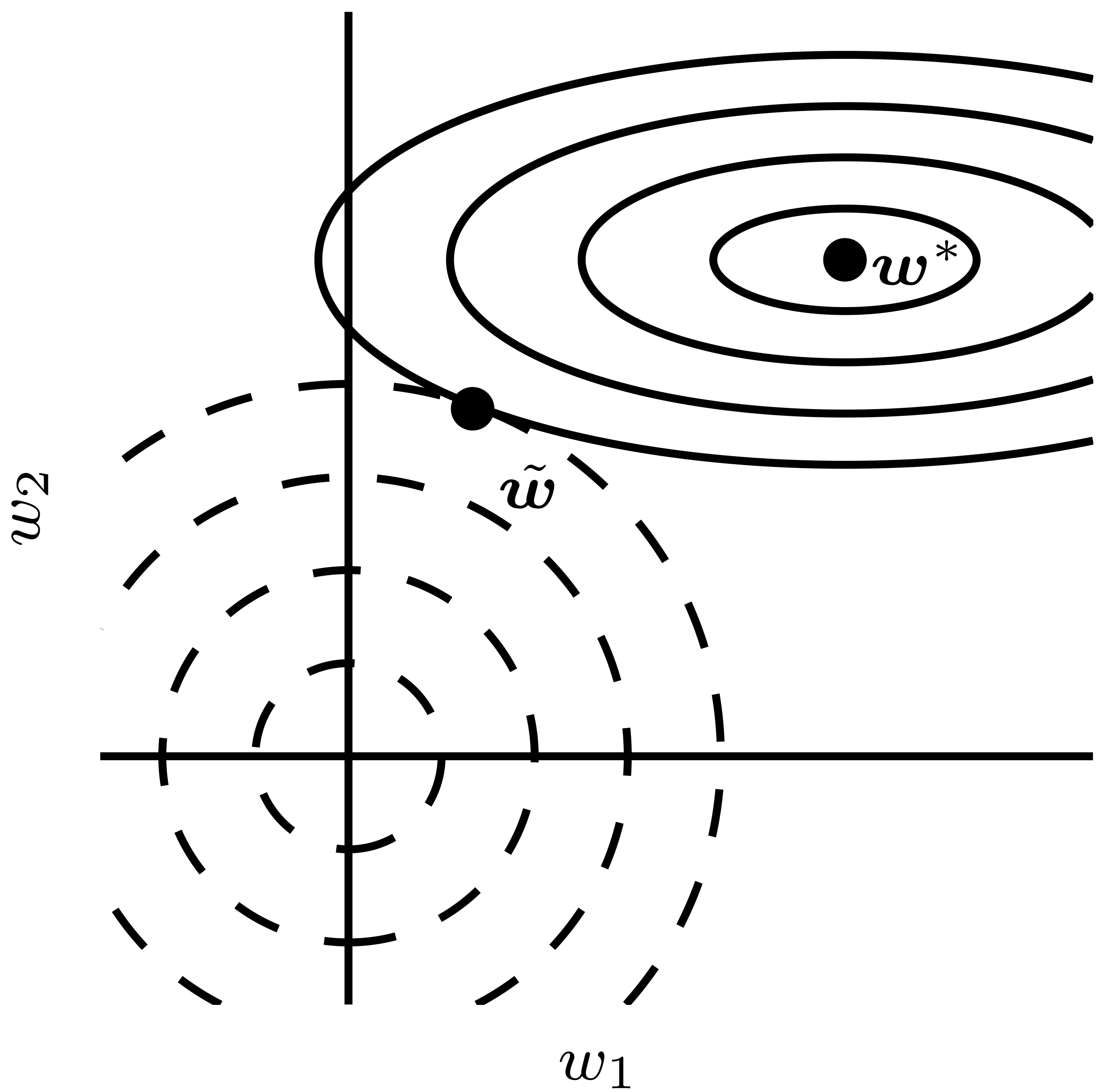


Figure 7.1

(Goodfellow 2016)

Norm Penalties

- L1: Encourages sparsity, equivalent to MAP Bayesian estimation with Laplace prior
- Squared L2: Encourages small weights, equivalent to MAP Bayesian estimation with Gaussian prior

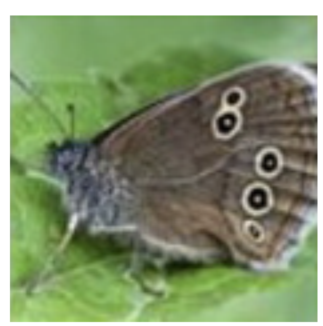
Dataset Augmentation

Affine Distortion

Noise

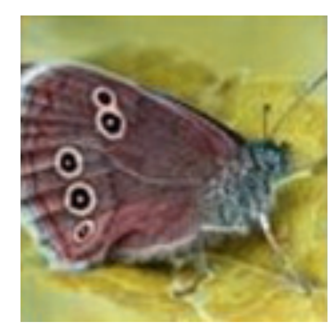
Elastic Deformation

Horizontal flip



Random Translation

Hue Shift



(Goodfellow 2016)

Multi-Task Learning

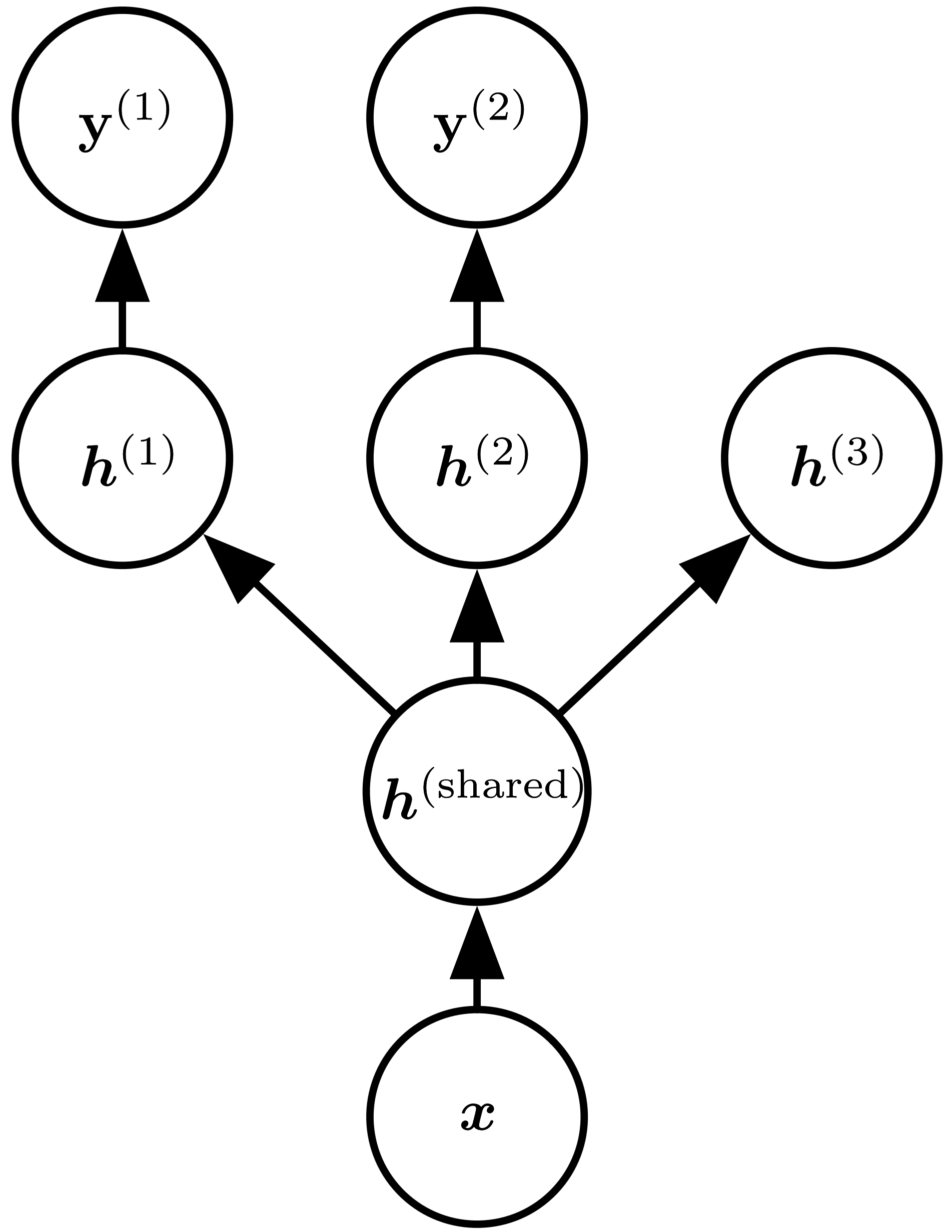


Figure 7.2

Learning Curves

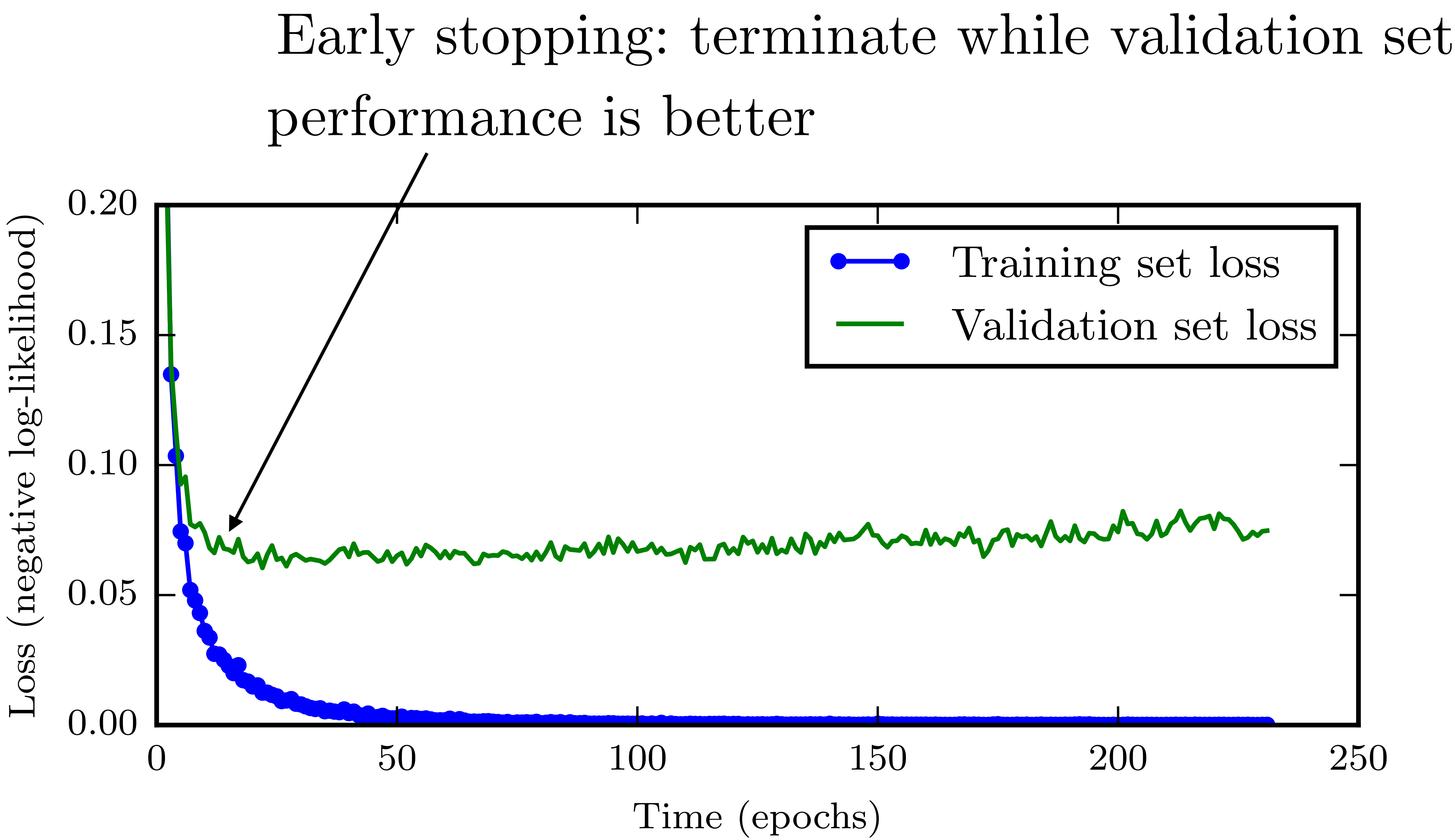


Figure 7.3

(Goodfellow 2016)

Early Stopping and Weight Decay

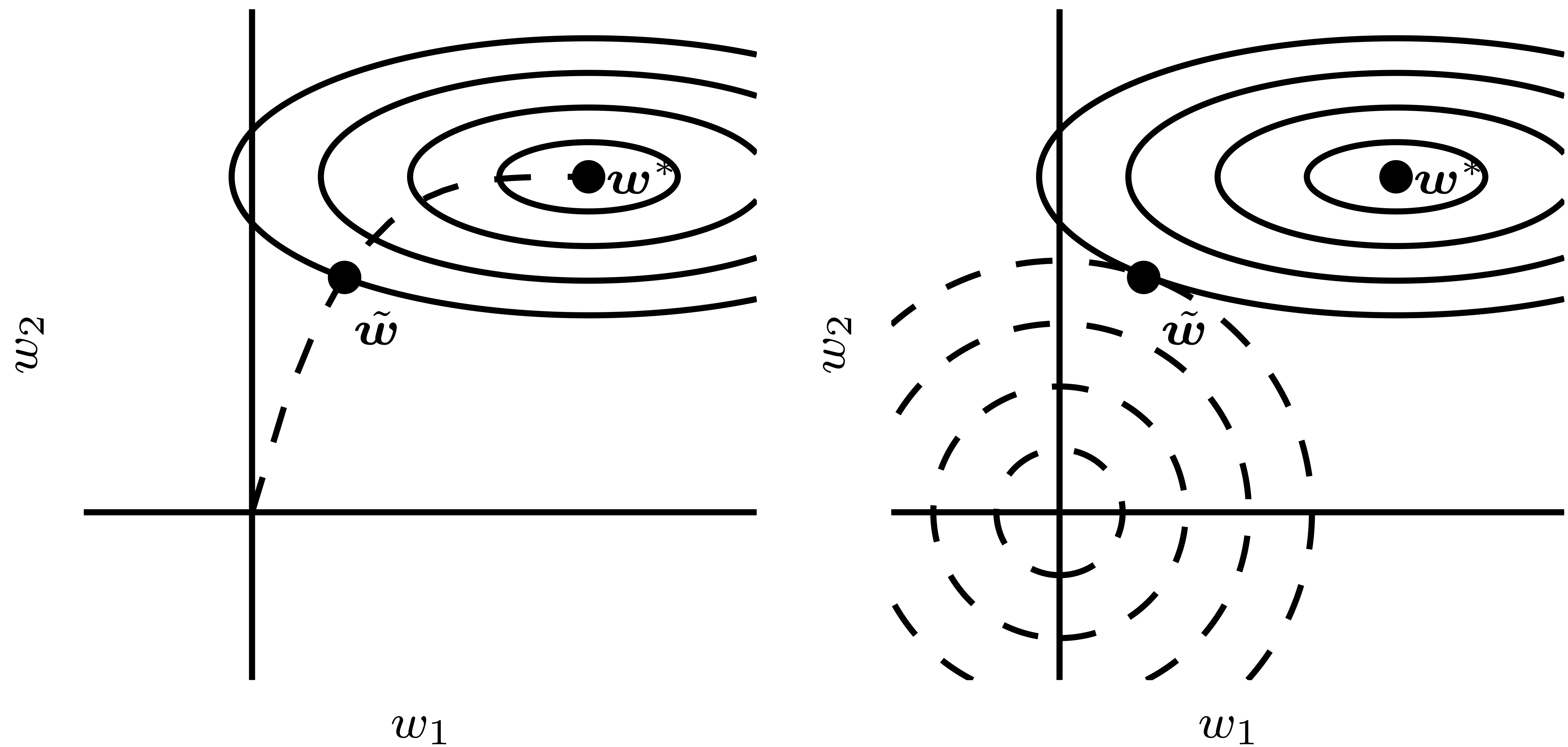


Figure 7.4

(Goodfellow 2016)

Sparse Representations

$$\underset{\boldsymbol{y} \in \mathbb{R}^m}{\begin{bmatrix} -14 \\ 1 \\ 19 \\ 2 \\ 23 \end{bmatrix}} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times n}}{\begin{bmatrix} 3 & -1 & 2 & -5 & 4 & 1 \\ 4 & 2 & -3 & -1 & 1 & 3 \\ -1 & 5 & 4 & 2 & -3 & -2 \\ 3 & 1 & 2 & -3 & 0 & -3 \\ -5 & 4 & -2 & 2 & -5 & -1 \end{bmatrix}} \underset{\boldsymbol{h} \in \mathbb{R}^n}{\begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \\ -3 \\ 0 \end{bmatrix}} \tag{7.47}$$

Bagging

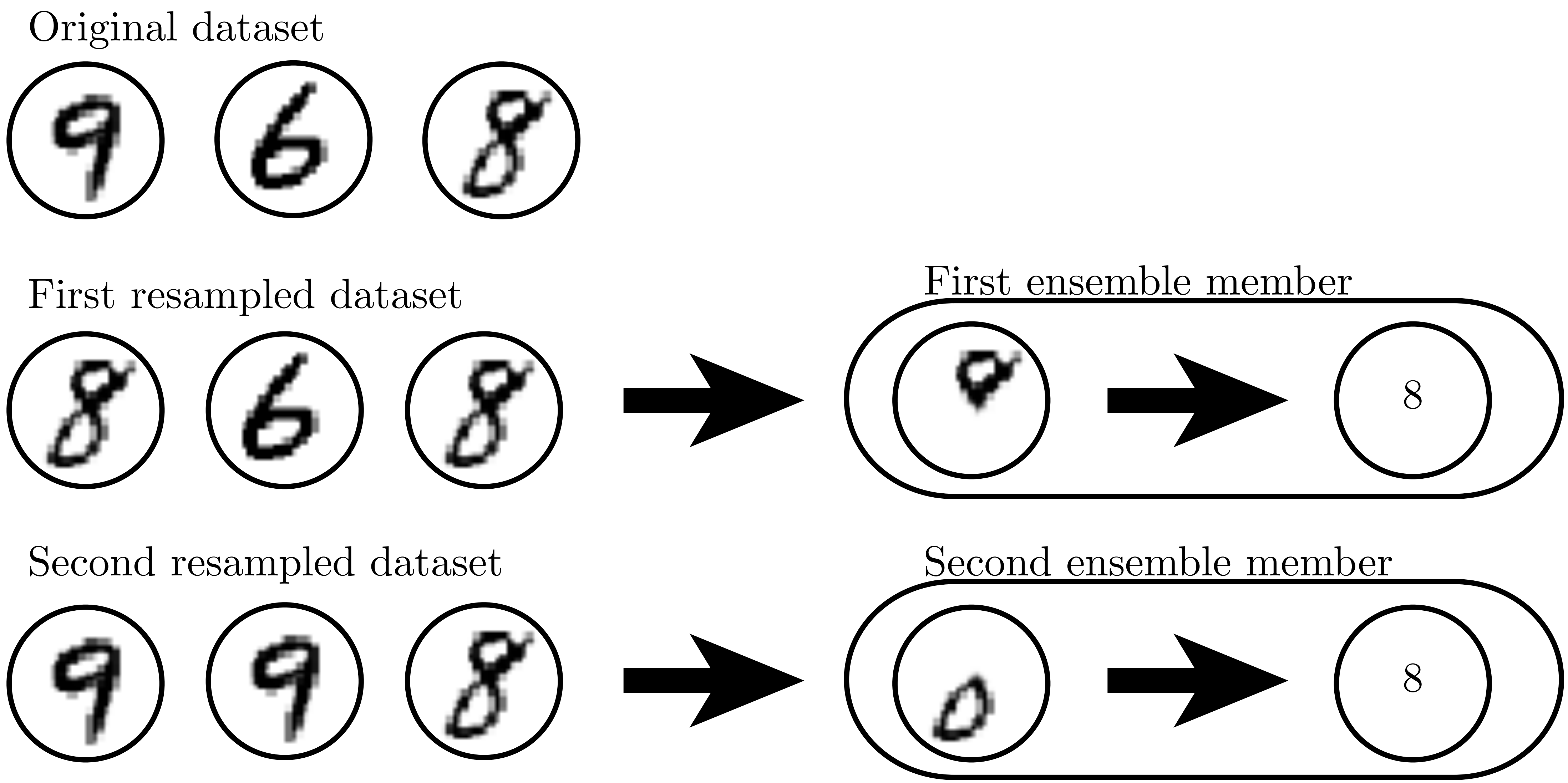
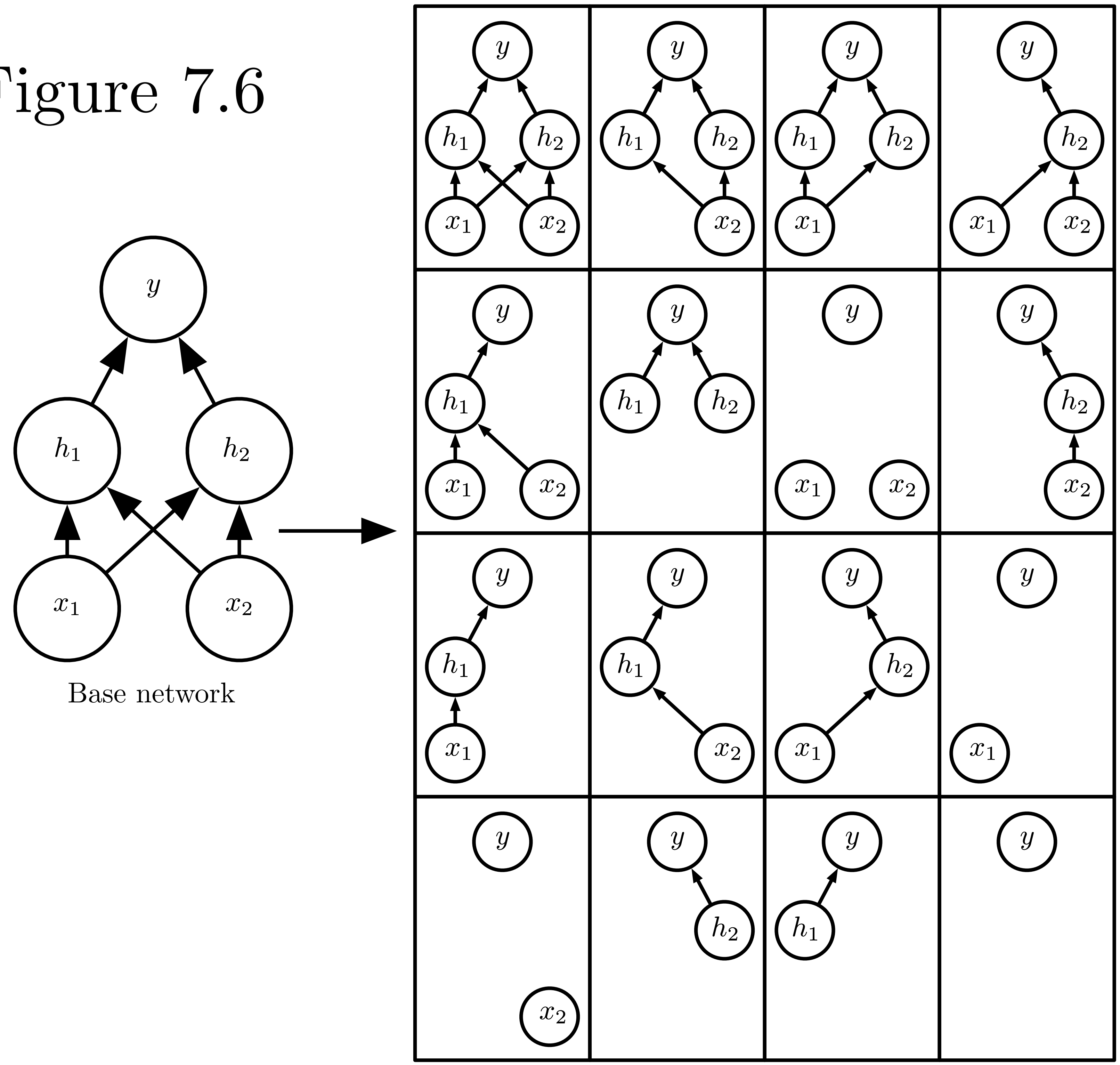


Figure 7.5

(Goodfellow 2016)

Dropout

Figure 7.6

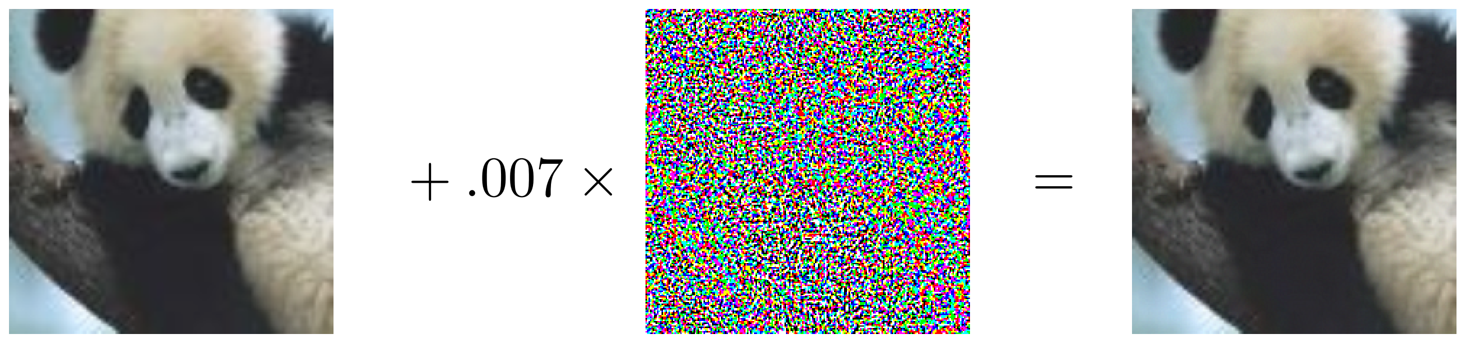


Base network

Ensemble of subnetworks

(Goodfellow 2016)

Adversarial Examples



$\boldsymbol{x}$

$y =$"panda"
w/ 57.7%
confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
w/ 8.2%
confidence

$\boldsymbol{x} + \epsilon \, \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
w/ 99.3 %
confidence

Figure 7.8

Training on adversarial examples is mostly intended to improve security, but can sometimes provide generic regularization.

Tangent Propagation

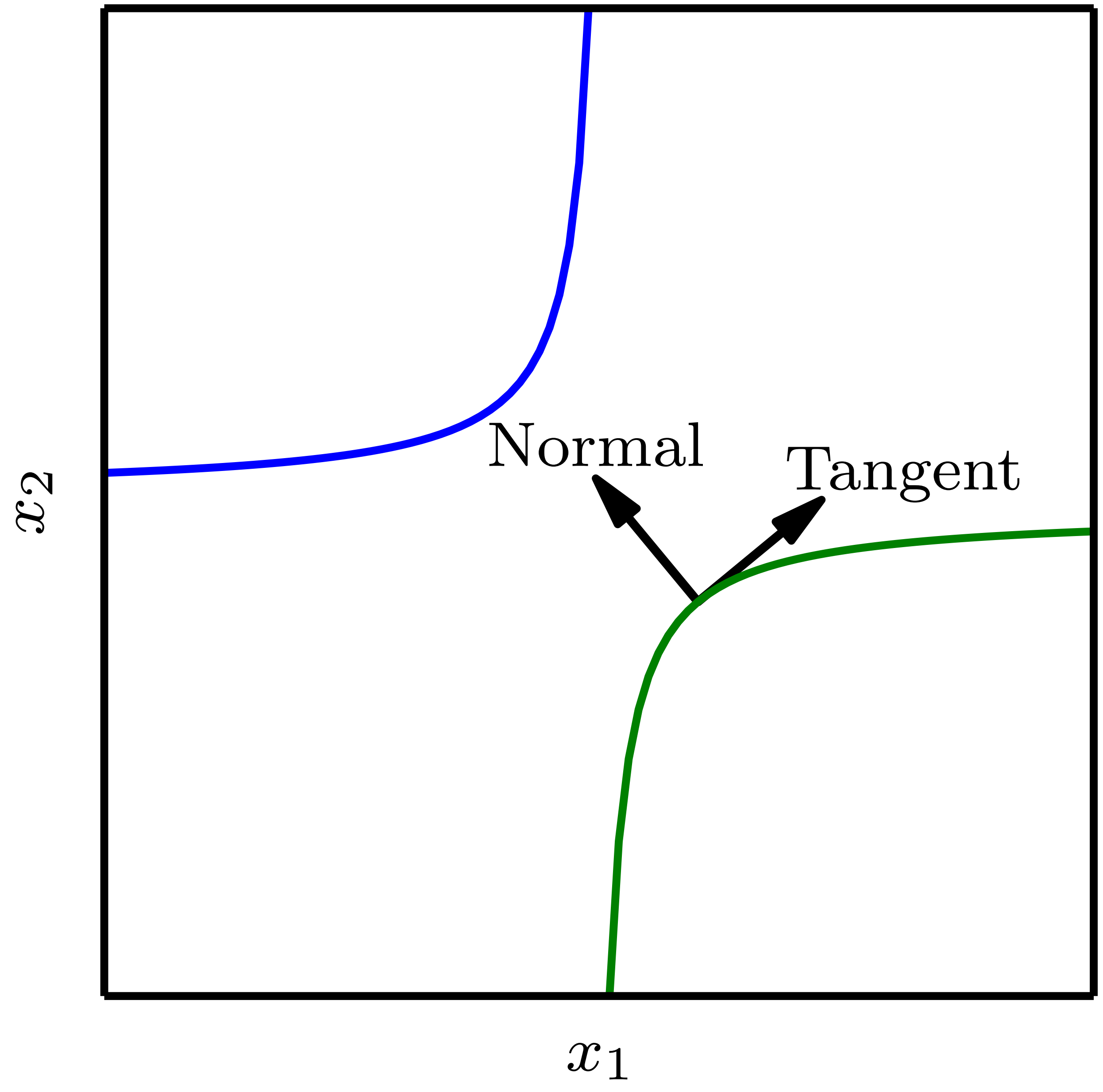


Figure 7.9

(Goodfellow 2016)