Deep Learning - Theory and Practice

Basics of Machine Learning

23-01-2020

http://leap.ee.iisc.ac.in/sriram/teaching/DL20/

deeplearning.cce2020@gmail.com

Established 1911

INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Problems in Decision Theory

- Decision Theory
  - Inference problem
    - Finding the joint density $p(\mathbf{x}, \mathbf{t})$
  - Decision problem
    - Using the inference to make the classification or regression decision

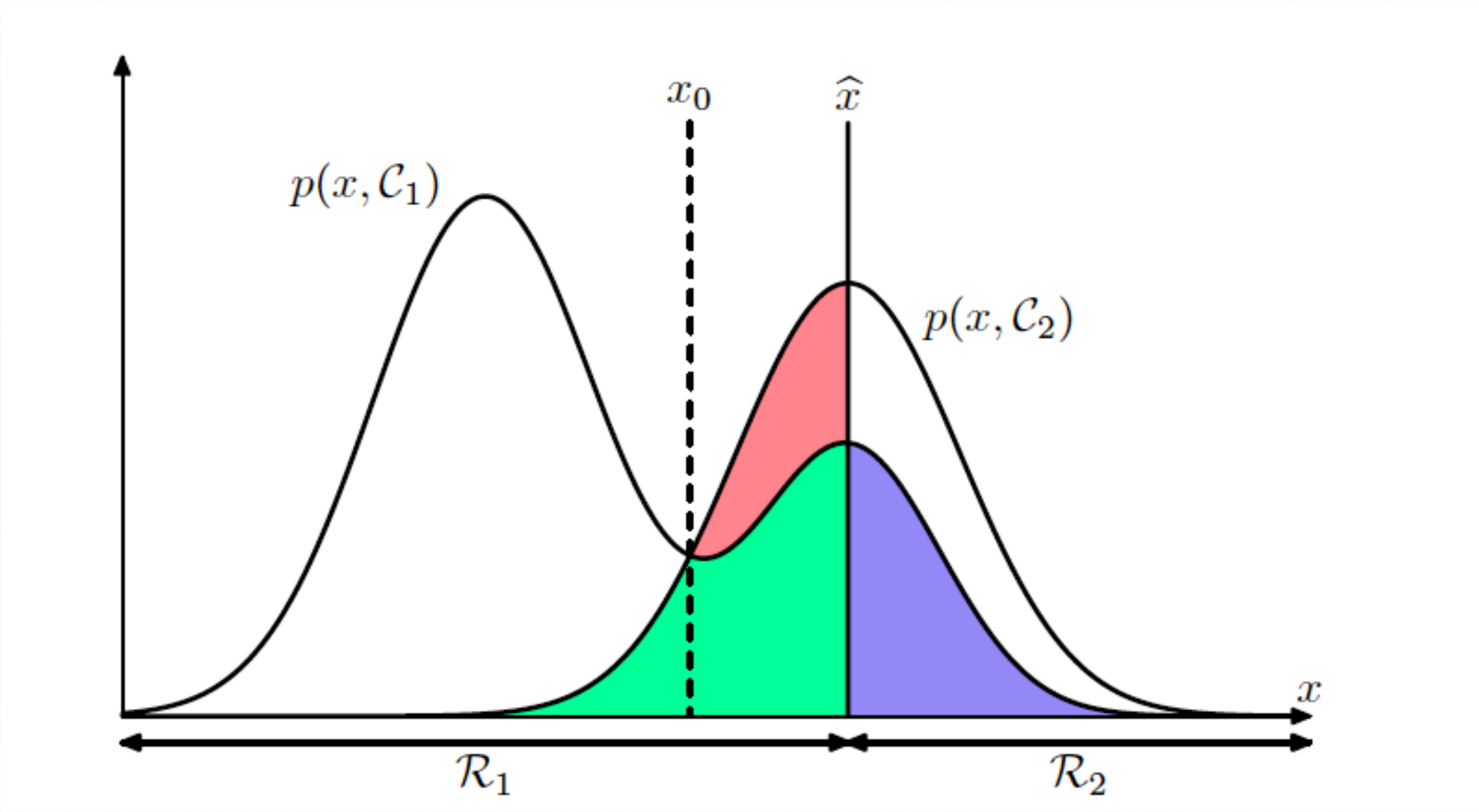
Decision Problem - Classification

- Minimizing the mis-classification error
- Decision based on maximum posteriors

$$argmax_j \; p(C_j|\mathbf{x})$$

- Loss matrix
  - Can be used for non uniform error weighting.

Decision Theory



Approaches for Inference and Decision

I. Finding the joint density from the data.

$$p(C_k|\mathbf{x}) \propto p(\mathbf{x}|C_k)p(C_k)$$

II. Finding the posteriors directly.

Neural Networks

III. Using discriminant functions for classification.

Advantages of Posteriors

- Minimizing the risk
- Reject Option
- Combining models
- Compensating for class priors

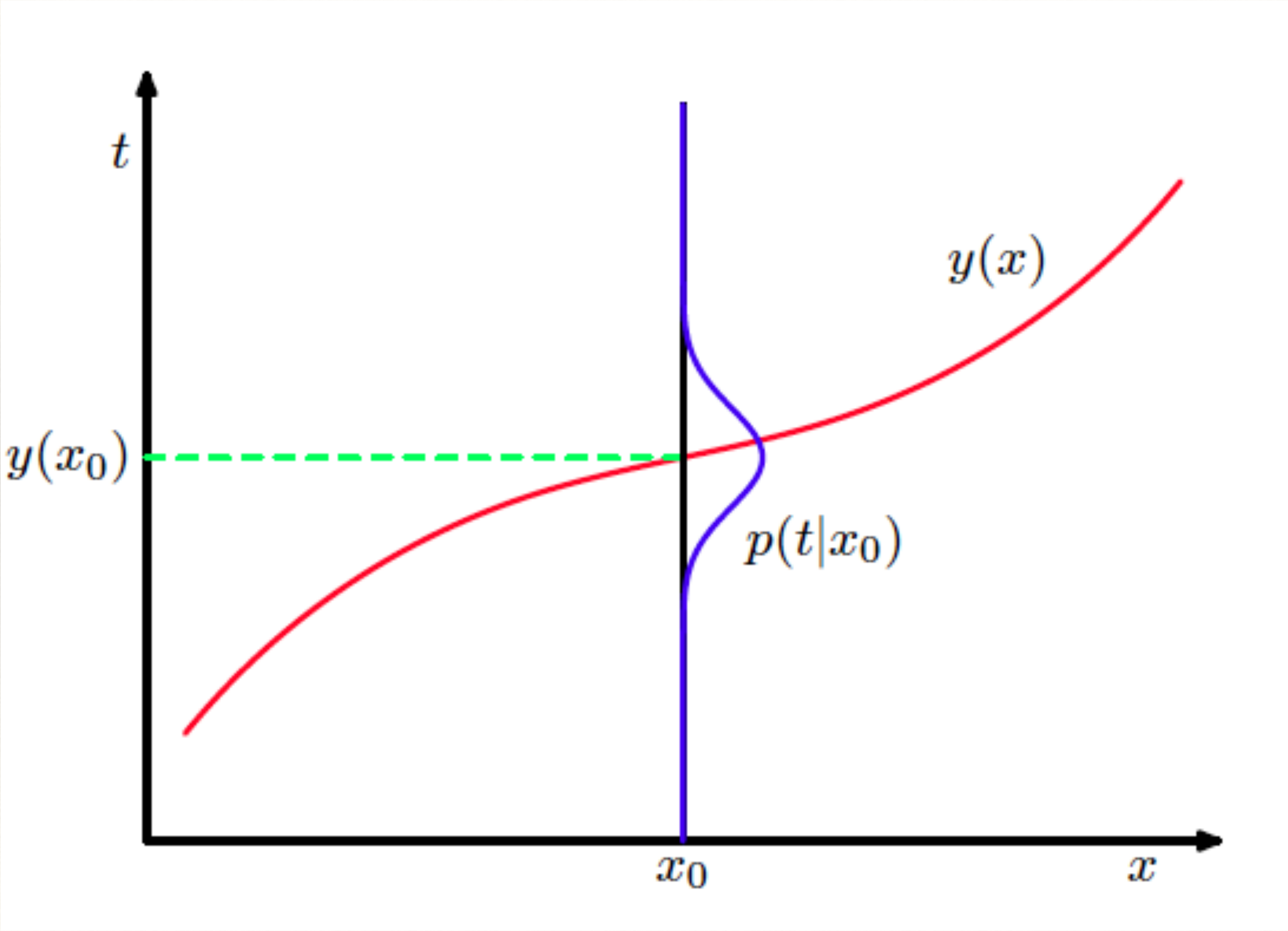
Loss Function for Regression

- With a mean square error loss

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t.$$

- The problem boils down to conditional expectation of the data given the

Regression Problem



Matrix Derivatives

$$\left(\frac{\partial \mathbf{a}}{\partial x}\right)_i = \frac{\partial a_i}{\partial x}$$

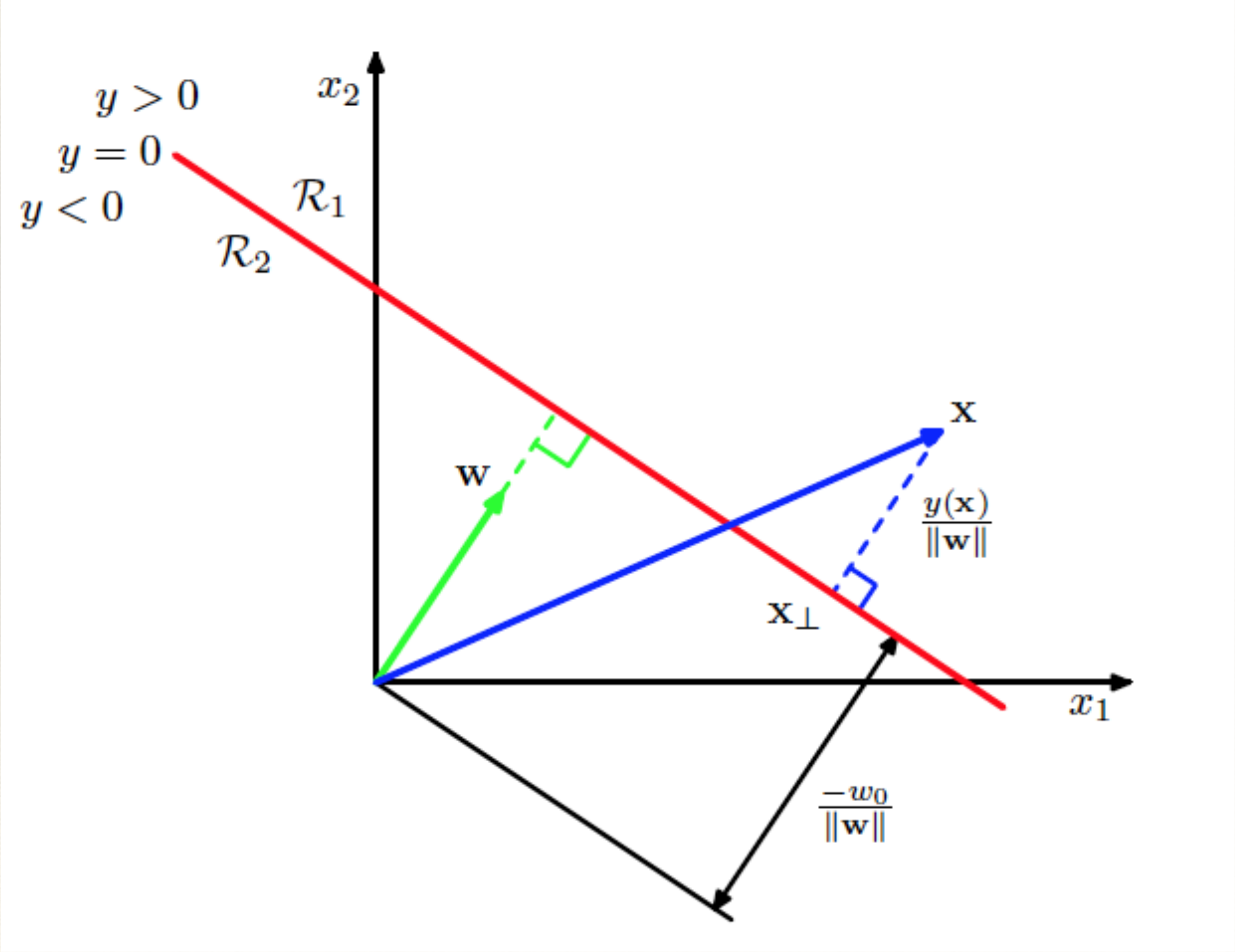
$$\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial a_i}$$

$$\left(\frac{\partial \mathbf{a}}{\partial \mathbf{b}}\right)_{ij} = \frac{\partial a_i}{\partial b_j}.$$

Linear Models for Classification

- Optimize a modified cost function

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$



Established 1911

Least Squares for Classification

- K-class classification problem

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

$$\mathbf{y}(\mathbf{x}) = \widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{x}}$$

- With 1-of-K hot encoding, and least squares regression

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^{\mathrm{T}}(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

Bishop - PRML book (Chap 3)

Principal Component Analysis

- First M eigenvectors of data covariance matrix

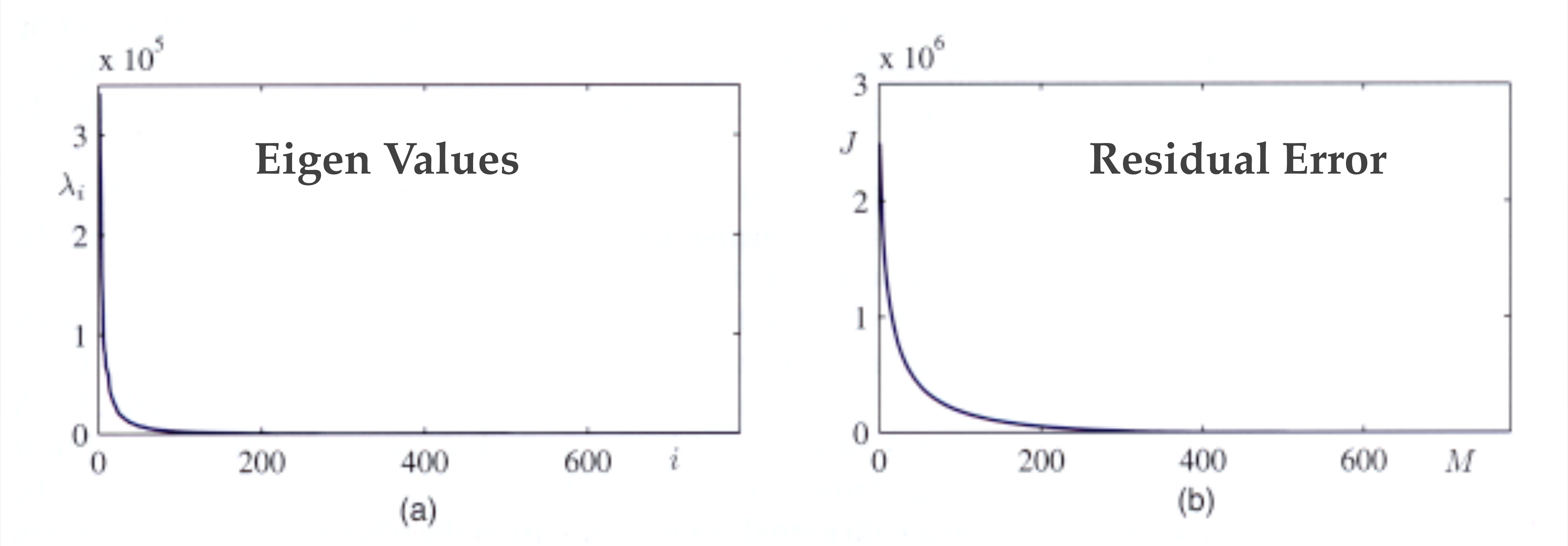
$$S = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

- Residual error from PCA

$$J = \sum_{i=M+1}^{D} \lambda_i$$

PRML - C. Bishop (Sec. 12.1)

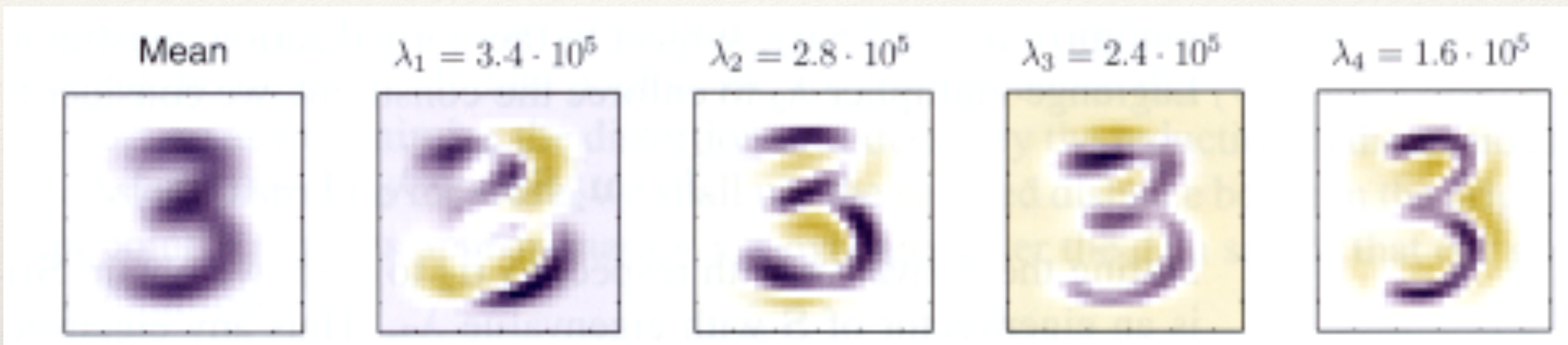
PCA



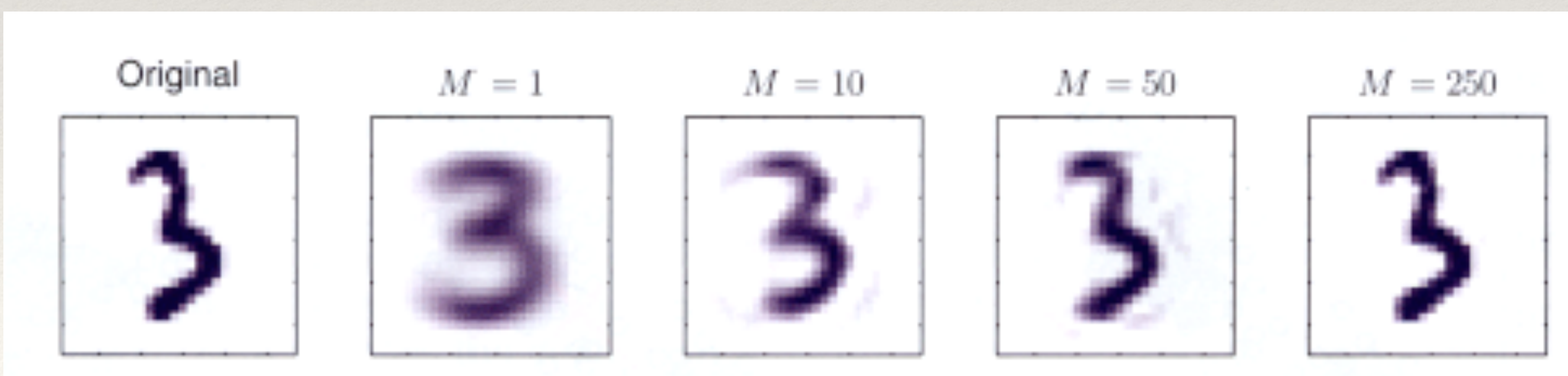
PRML - C. Bishop (Sec. 12.1)

PCA - Reconstruction

Eigenvectors



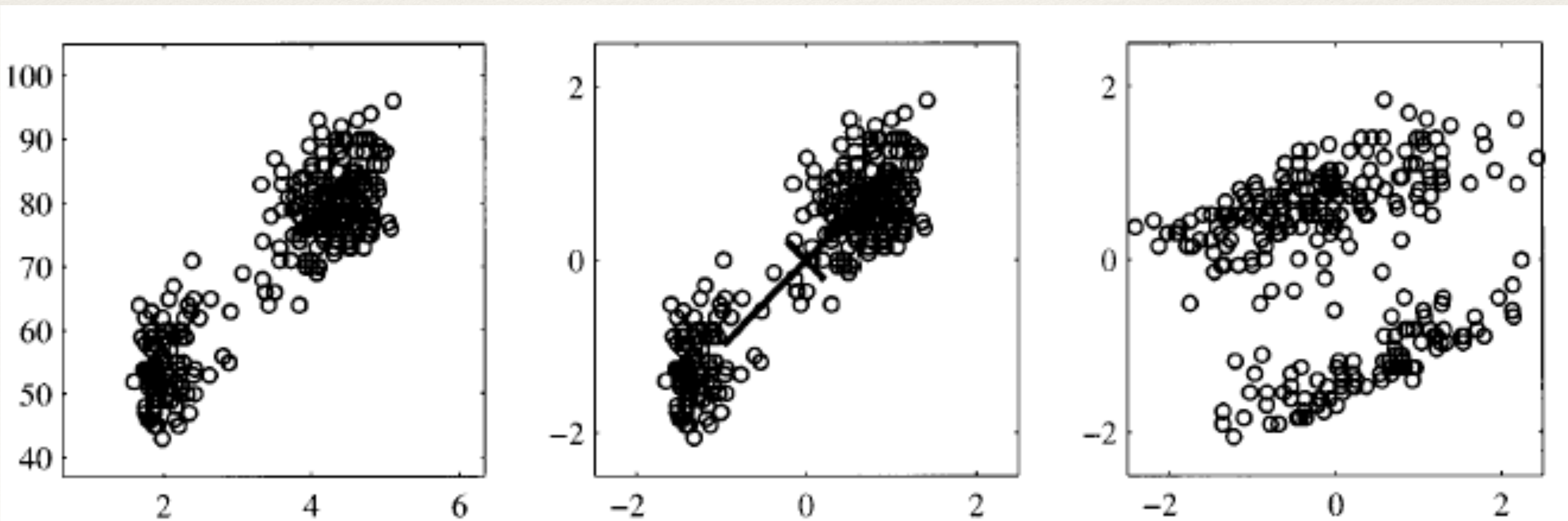
PCA - Reconstruction



PRML - C. Bishop (Sec. 12.1)

Whitening the Data

Original Data **Mean Removed data** **Whitened data**



PRML - C. Bishop (Sec. 12.1)