Chapter 10. Information-Theoretic Learning Models

Neural Networks and Learning Machines (Haykin)

Lecture Notes on

Self-learning Neural Algorithms

V.2017.09.18

Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University

Contents

10.1 Introduction

- Information-theoretic models that lead to self-organization in a principled manner
- Maximum-mutual information principle (Linsker, 1988):

  The synaptic connections of a multilayered neural network develop in such a way as to maximize the amount of information that is preserved when signals are transformed at each processing stage of the network, subject to certain constraints

- Information-theoretic function of perceptual systems (Attneave, 1954):

  A major function of the perceptual machines is to strip away some of the redundancy of stimulation, to describe or encode information in a form more economical than that in which it impinges on the receptors.

10.2 Entropy (1/2)

Discrete random variable

$$X = \{x_k \mid k = 0, \pm 1, ..., \pm K\}$$

Probability of the event $X = x_k$

$$p_k = P(X = x_k)$$

$$\left(0 \le p_k \le 1 \quad \text{and} \quad \sum_{k=-K}^{K} p_k = 1 \right)$$

Amount of information gained after observing the event $X = x_k$ with probability p_k

$$I(x_k) = \log\left(\frac{1}{p_k}\right) = -\log p_k$$

If the event occurs with probability $p_k = 1$, there is no "surprise", and therefore no "information" is conveyed by the occurreence of the event $X = x_k$, since we know what the message must be.

Properties of information $I(x_k)$

1. $I(x_k) = 0 \quad \text{for } p_k = 1$
2. $I(x_k) \ge 0 \quad \text{for } 0 \le p_k \le 1$
3. $I(x_k) > I(x_i) \quad \text{for } p_k < p_i$

10.2 Entropy (2/2)

Entropy

$$H(X) = \mathbf{E}[I(x_k)] = \sum_{k=-K}^{K} p_k I(x_k) = -\sum_{k=-K}^{K} p_k \log p(x_k)$$

i.e. average amount of information conveyed per message

Entropy is bounded by

$$0 \le H(X) \le \log\ (2K+1)$$

1. $H(X) = 0$: no uncertainty
2. $H(X) = 1$: maximum uncertainty

Differential entropy of continous random variables

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log p_X(x) dx = -\mathbf{E}[\log p_X(x)]$$

$$h(\mathbf{X}) = -\int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) \log p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = -\mathbf{E}[\log p_{\mathbf{X}}(\mathbf{x})]$$

10.3 Maximum Entropy Principle (1/2)

- Maximum entropy principle is a constrained optimization problem
  1. A set of known states
  2. Unknown probabilities of the states
  3. Constraints on the probability distribution of the states

  When an inference is made on the basis of incomplete information, it should be drawn from the probability distribution that maximizes the entropy, subject to constraints on the distribution

$$h(X) = -\int_{-\infty}^{\infty} p_X(x) \log p_X(x) dx$$

1. $p_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} p_X(x) dx = 1$
3. $\int_{-\infty}^{\infty} p_X(x) g_i(x) = \alpha_i \quad \text{for } i = 1,2,\ldots,m$

10.3 Maximum Entropy Principle (2/2)

Method of Lagrange multiplyers for solving

the constrained optimization problem

$$J(p)=\int_{-\infty}^{\infty}\left[-p_X(x)\log p_X(x)+\lambda_0 p_X(x)+\sum_{i=1}^{m}\lambda_i g_i(x)\lambda_0 p_X(x)\right]dx$$

Setting $\frac{\partial J(p)}{\partial p_X(x)}=0$, we get

$$-1-\log p_X(x)+\lambda_0+\sum_{i=1}^{m}\lambda_i g_i(x)=0$$

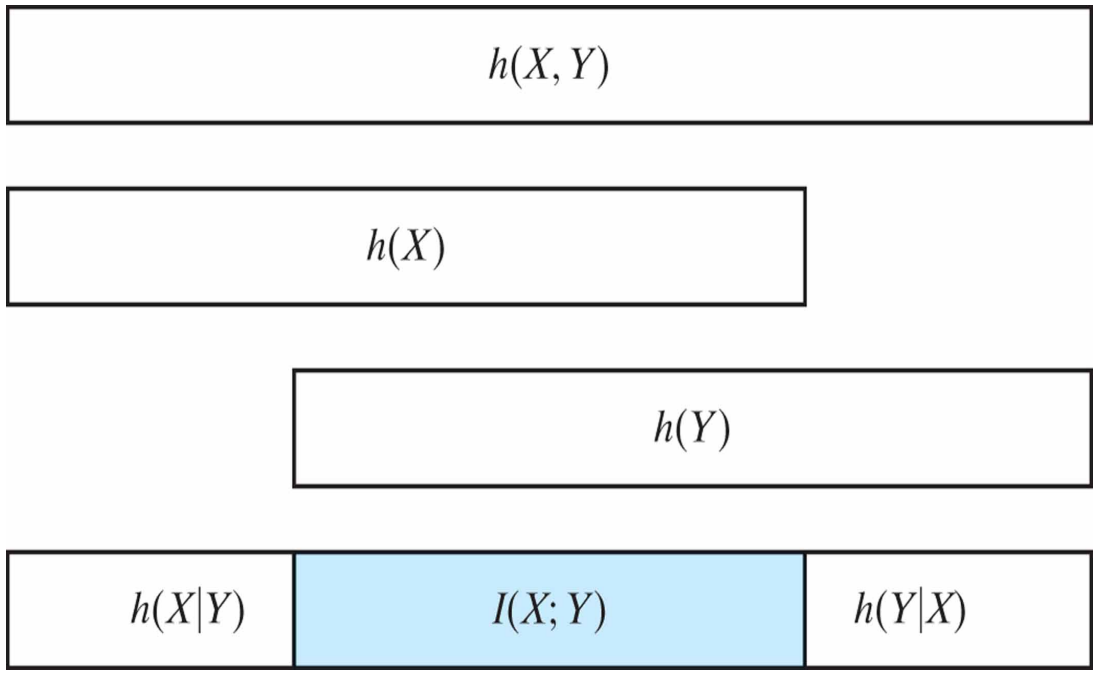
$$p_X(x)=\exp\left(-1+\lambda_0\sum_{i=1}^{m}\lambda_i g_i(x)\right)$$

$$h(X)=-\int_{-\infty}^{\infty}p_X(x)\log p_X(x)dx$$

1. $p_X(x)\geq 0$
2. $\int_{-\infty}^{\infty}p_X(x)dx=1$
3. $\int_{-\infty}^{\infty}p_X(x)g_i(x)=\alpha_i$ for $i=1,2,\ldots,m$

10.4 Mutual Information (1/3)

Figure 10.1: Relationships embodied in the three lines of Eq. (10.32), involving the mutual information $I(X; Y)$.



$h(X)$ = uncertainty about X before observing Y

$h(X|Y)$ = uncertainty about X after observing Y.

$I(X;Y) = h(X) - h(X|Y)$. The uncertainty about the system input X that is resolved by observing the system output Y.

$$
\begin{aligned}
& I(X;Y) \\
&= h(X) - h(X|Y) \\
&= h(Y) - h(Y|X) \\
&= (h(X) + h(Y)) - h(X,Y) \\
& Eq.\ (10.32)
\end{aligned}
$$

10.4 Mutual Information (2/3)

Joint probability density function of X and Y

$$p_{X,Y}(x,y)=p_{Y|X}(y|x)p_X(x)$$

Joint differential entropy of X and Y

$$h(X,Y)=h(X)+h(Y|X)$$

$$h(X,Y)=h(Y)+h(X|Y)$$

Mutual information (MI) between X and Y

$$I(X;Y)=h(X)-h(X|Y)$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right)dxdy \qquad p_{X,Y}(x,y)=p_{X|Y}(x|y)p_Y(y)$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{X|Y}(x|y)p_Y(y)\log\left(\frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)p_Y(y)}\right)dxdy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{X|Y}(x|y)p_Y(y)\log\left(\frac{p_{X|Y}(x|y)}{p_X(x)}\right)dxdy$$

10.4 Mutual Information (3/3)

Differential entropy is a special case of MI

$$h(X) = I(X;X)$$

Property 1. Nonnegativity

$$I(X;Y) \geq 0$$

Property 2. Symmetry

$$I(Y;X) = I(X;Y)$$

Property 3. Invariance

$$I(Y;X) = I(U;V)$$

with $u = f(x)$, $v = g(y)$

Generalization of MI

$$I(\mathbf{X};\mathbf{Y}) = h(\mathbf{X}) - h(\mathbf{X}|\mathbf{Y})$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) \log\left(\frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})p_{\mathbf{Y}}(\mathbf{y}) \log\left(\frac{p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}\right) d\mathbf{x}d\mathbf{y}$$

10.5 Kullback-Leibler Divergence (1/2)

KL Divergence (KLD) between $p_{\mathbf{X}}(\mathbf{x})$ and $g_{\mathbf{X}}(\mathbf{x})$

$$D_{p||g} = \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{p_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right) d\mathbf{x}$$

$$= \mathbf{E}\left[\log\left(\frac{p_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right)\right]$$

A distance between two probability distributions, but no symmetricity, thus divergence.

$$D_{p||g} \neq D_{g||p}$$

Property 1. Nonnegativity

$$D_{p||g} \geq 0$$

Property 2. Invariance

$$D_{p_{\mathbf{X}}||g_{\mathbf{X}}} = D_{p_{\mathbf{Y}}||g_{\mathbf{Y}}}$$

10.5 Kullback-Leibler Divergence (2/2)

Relationship between KLD and MI

$$I(\mathbf{X};\mathbf{Y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{-\infty} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\log\left(\frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}}(\mathbf{y})}\right)d\mathbf{x}d\mathbf{y}$$

$$I(\mathbf{X};\mathbf{Y}) = D_{p_{\mathbf{X},\mathbf{Y}}||p_{\mathbf{X}}p_{\mathbf{Y}}}$$

Mutual information between a pair of vectors $\mathbf{X}$ and $\mathbf{Y}$ is equal to the KL-divergence between the joint pdf $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$ and the product of the marginal pdfs $p_{\mathbf{X}}(\mathbf{x})$ and $p_{\mathbf{Y}}(\mathbf{y})$.

10.6 Copulas

A measure of statistical dependence between X and Y that is not disturbed by their scaled versions or their variances.

We transform X and Y into two new random variables U and V, respectively, such that the distributions of both U and V are uniform over the interval [0,1].

$$u = P_X(x), \quad v = P_Y(y)$$

The new pair of random variables (U, V) is uniquely determined, and called a copula.

$$P_{X,Y}(x,y) = C_{U,V}(P_X(x), P_Y(y)), \qquad C_{U,V}(u,v) = P(P_X^{-1}(x), P_Y^{-1}(y))$$

The copula, involving the pair of random variables (U, V) is a function that models the statistical dependence between U and V in a distribution-free manner.

Relationship between MI and the copula's entropy

$$I(X;Y) = I(U;V)$$

$$I(U;V) = h_C(U) + h_C(V) - h_C(U,V)$$

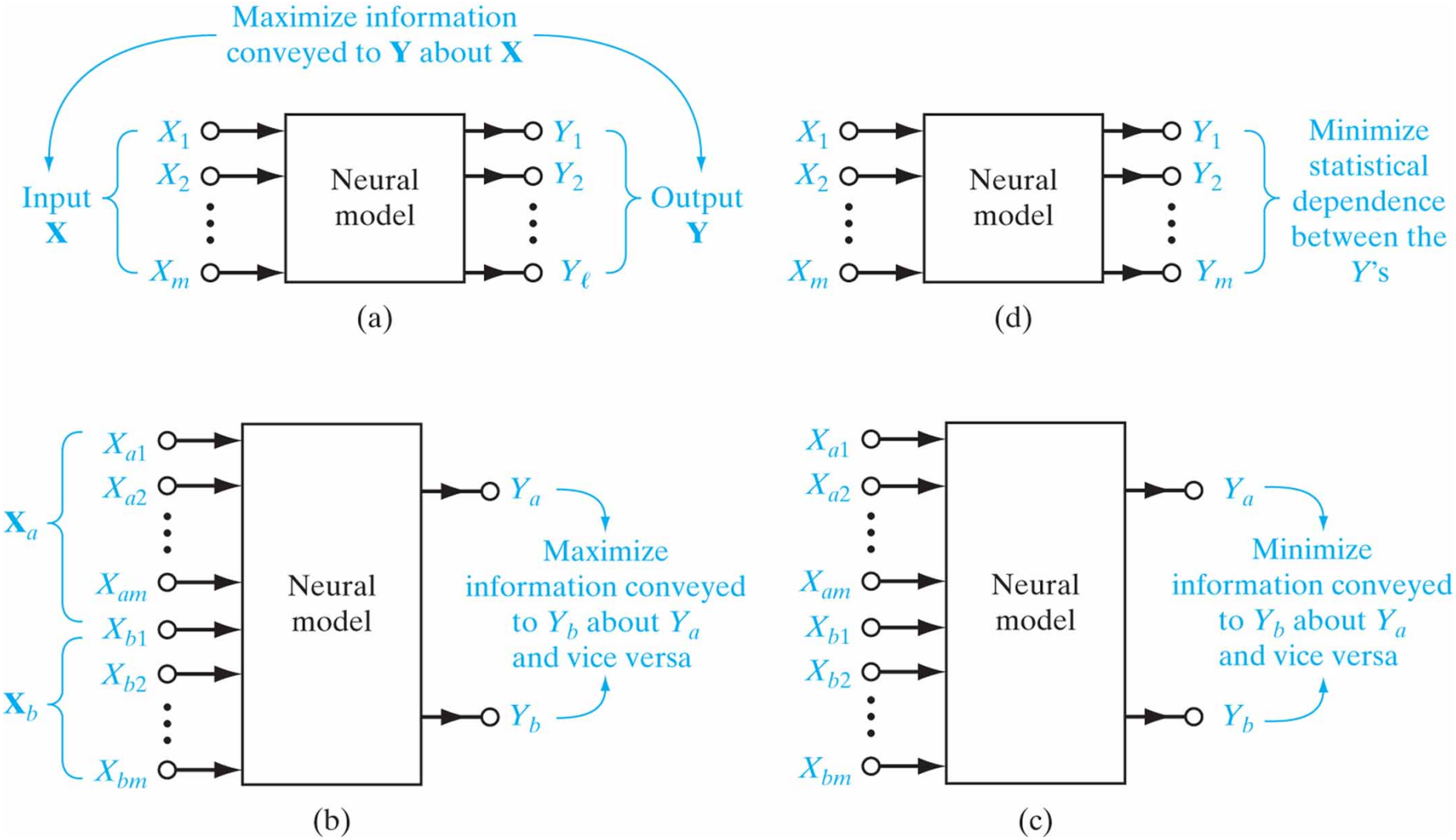
Section 10.2
p. 508 Example 1

Since $h_C(U) = 0$ and $h_C(V) = 0$ (U, V are uniformly distributed over [0,1]),

$$I(U;V) = -h_C(U,V) = \mathbf{E}\left[\log C_{U,V}(u,v)\right]$$

10.7 MI as an Objective Function

Figure 10.2: Four basic scenarios that lend themselves to the application of information maximization and its three variants.



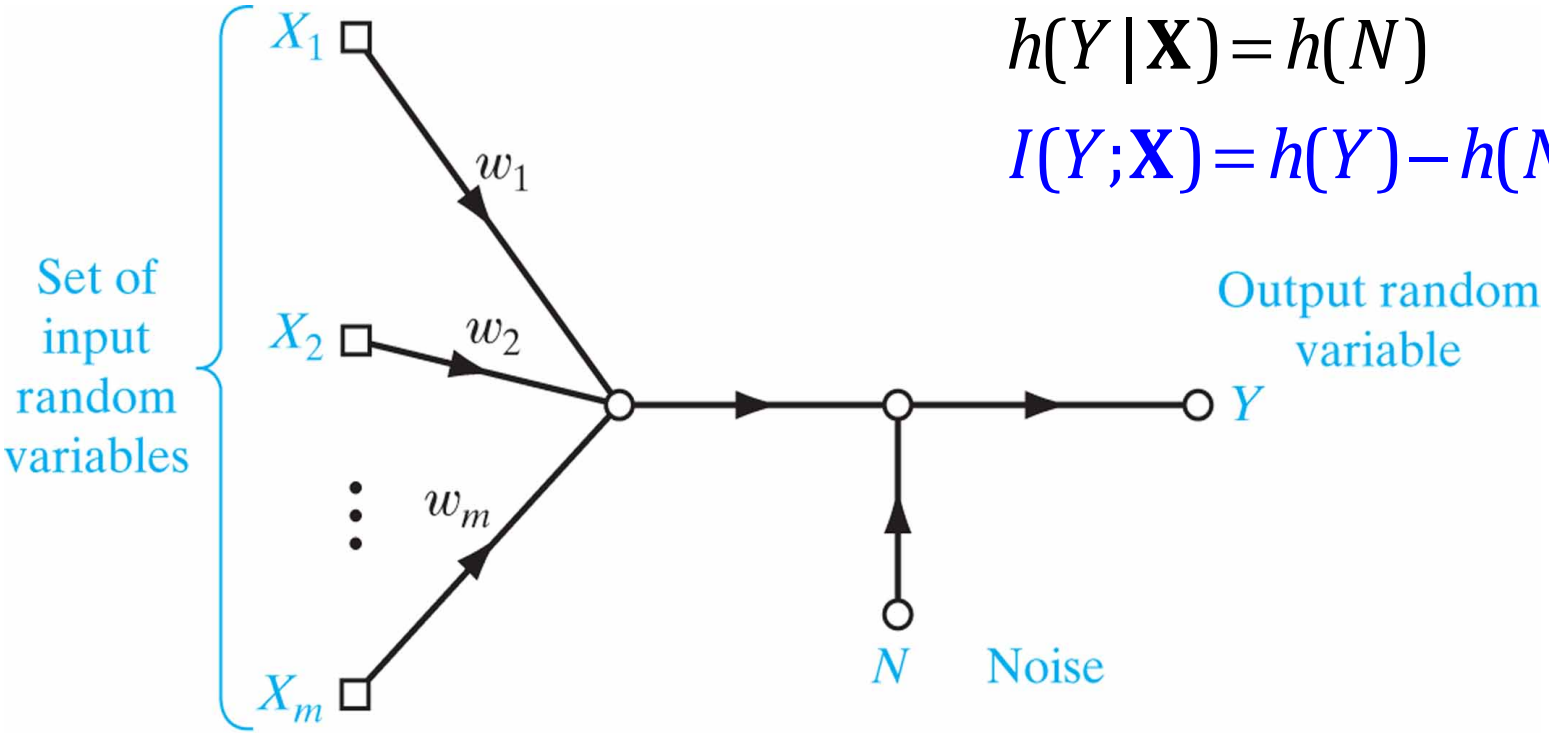
10.8 Maximum Mutual Information Principle (Infomax) (1/3)

Figure 10.3: Signal-flow graph of a noisy neuron.

$$I(Y;\mathbf{X}) = h(Y) - h(Y|\mathbf{X})$$

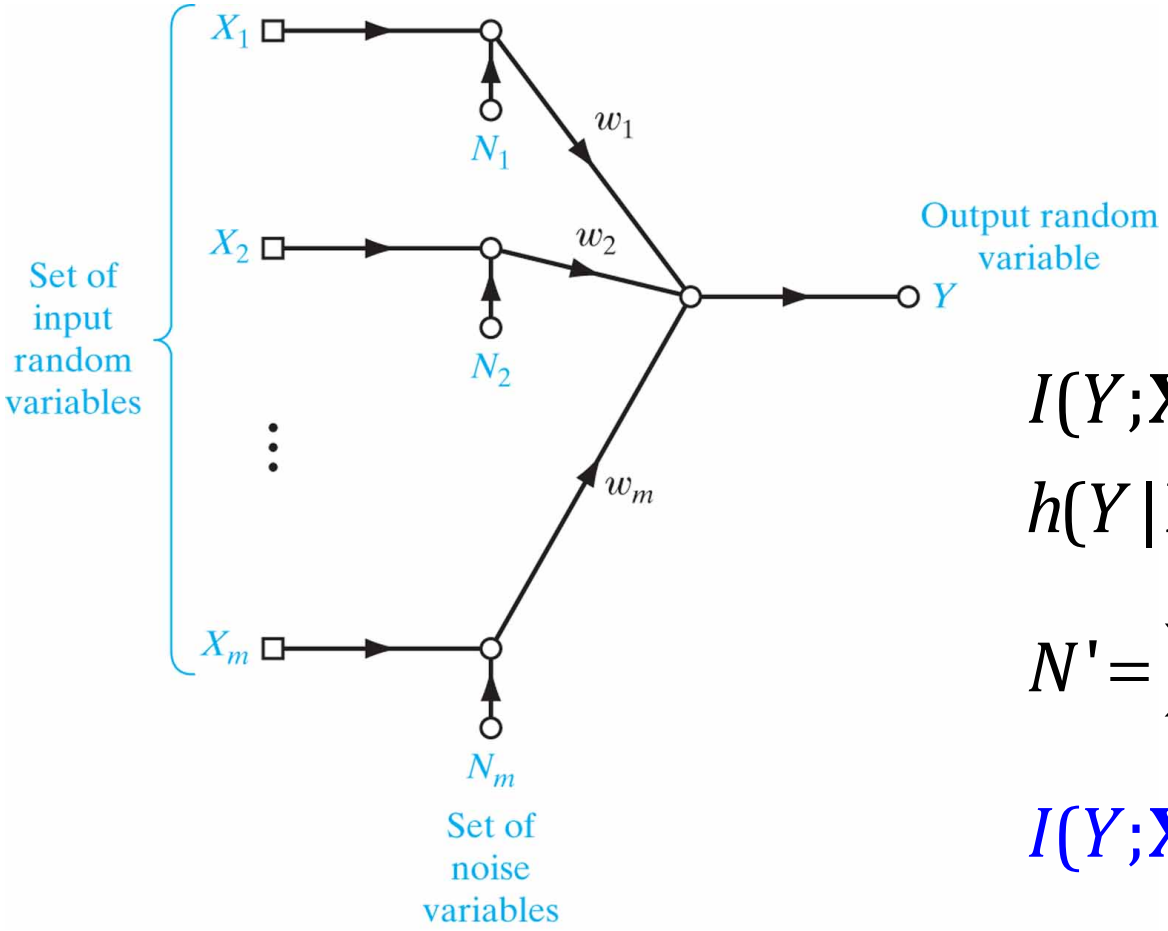
$$h(Y|\mathbf{X}) = h(N)$$

$$I(Y;\mathbf{X}) = h(Y) - h(N)$$



10.8 Maximum Mutual Information Principle (Infomax) (2/3)

Figure 10.4: Another noisy model of the neuron.



$$I(Y;\mathbf{X}) = h(Y) - h(Y|\mathbf{X})$$

$$h(Y|\mathbf{X}) = h(N')$$

$$N' = \sum_{i=1}^{m} w_i N_i$$

$$I(Y;\mathbf{X}) = h(Y) - h(N')$$

10.8 Maximum Mutual Information Principle (Infomax) (3/3)

Noiseless network

$$I(\mathbf{Y};\mathbf{X}) = h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{X})$$

With the noiselss mapping from $\mathbf{X}$ to $\mathbf{Y}$, the conditional differential entropy $h(\mathbf{Y}|\mathbf{X})$ attains the lowest possible value (diverges to -∞)

Since conditional entropy $h(\mathbf{Y}|\mathbf{X})$ is independent of $\mathbf{W}$, we can write

$$\frac{\partial I(\mathbf{Y};\mathbf{X})}{\partial \mathbf{W}} = \frac{\partial h(\mathbf{Y})}{\partial \mathbf{W}}$$

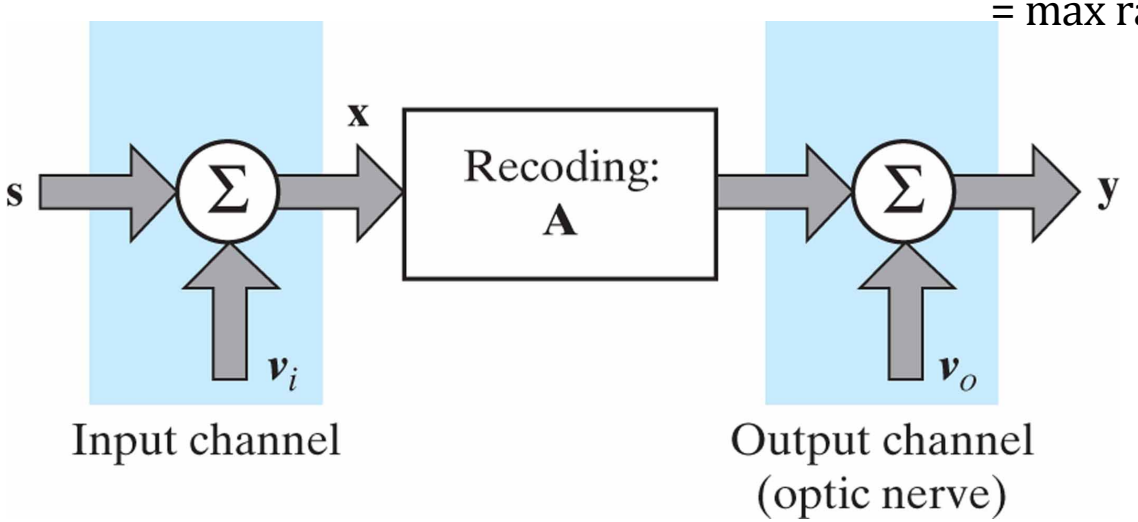
For a noiseless mapping network, maximizing the differential entropy of the network output ***Y*** is equivalent to maximizing the MI between ***Y*** and the network input ***X***, with both maximizations being performed w.r.t. the weight matrix ***W*** of the mapping network.

10.9 Infomax and Redundancy Reduction

Figure 10.5: Model of a perceptual system. The signal vector s and noise vectors $v\downarrow i$ and $v\downarrow o$ are values of the random vectors S, $N\downarrow i$, and $N\downarrow o$, respectively.

$C(\mathbf{Y})$: channel capacity

= max rate of info flow possible



$$\mathbf{X} = \mathbf{S} + \mathbf{N}_i \qquad \mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}_o$$

Redundancy measure

$$R = 1 - \frac{I(\mathbf{Y};\mathbf{S})}{C(\mathbf{Y})}$$

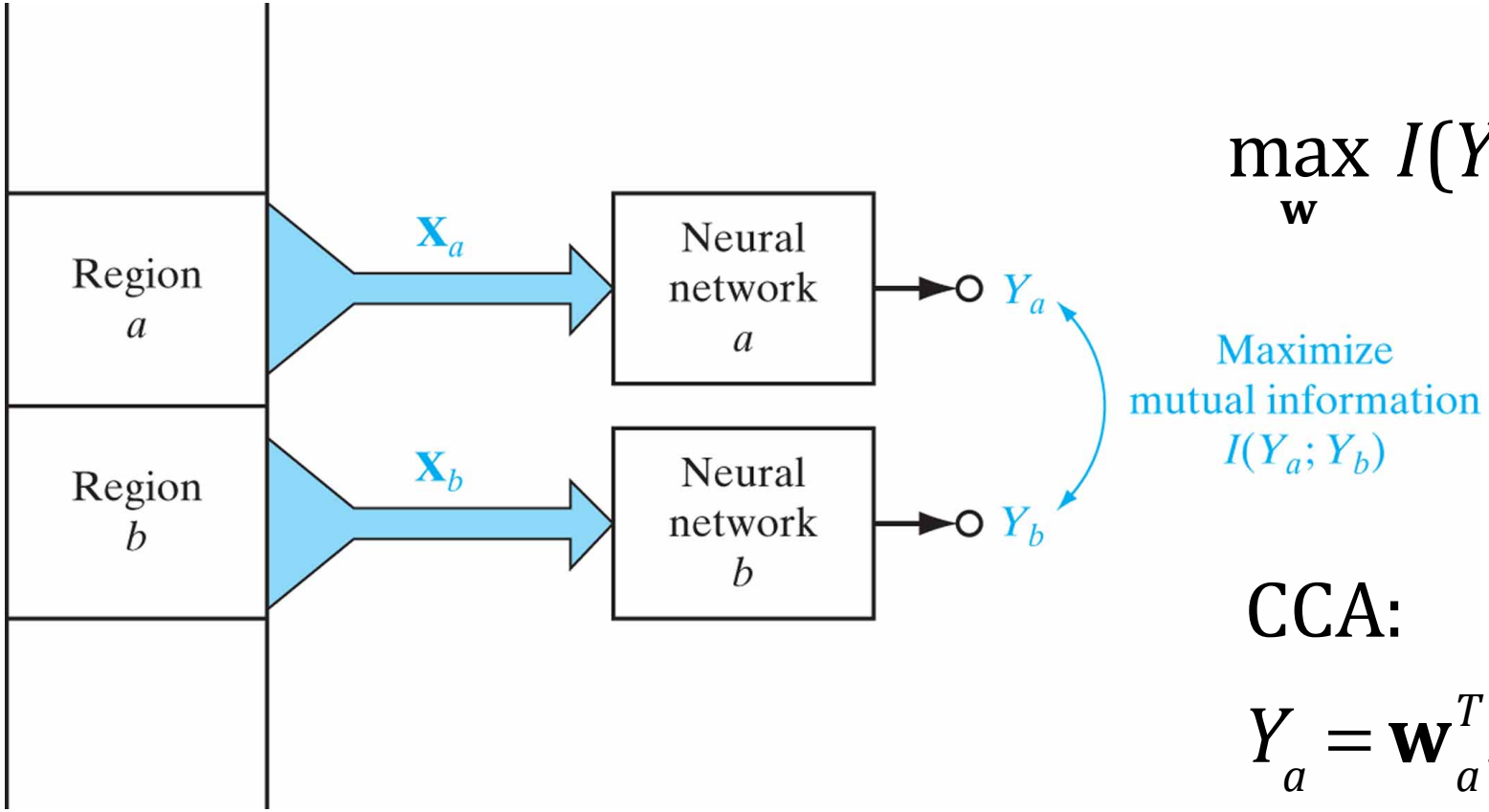
Minimize (Min redundancy)

$$F_1(\mathbf{Y};\mathbf{S}) = C(\mathbf{Y}) - \lambda I(\mathbf{Y};\mathbf{S})$$

Maximize (Infomax)

$$F_2(\mathbf{Y};\mathbf{S}) = I(\mathbf{Y};\mathbf{S}) + \lambda C(\mathbf{Y})$$

10.10 Spatially Coherent Features



$$\max_{\mathbf{w}} I(Y_a;Y_b)$$

CCA:

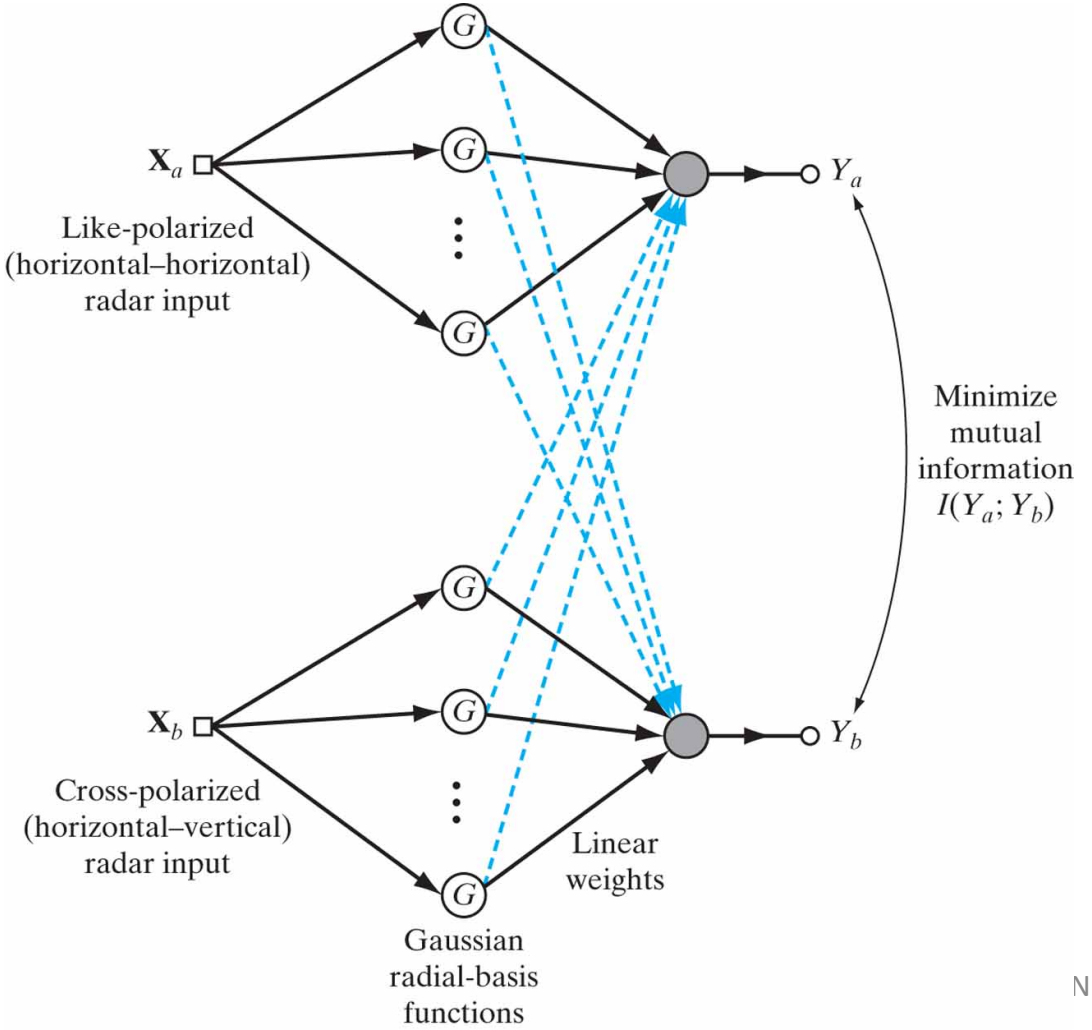
$$Y_a = \mathbf{w}_a^T \mathbf{X}_a$$

$$Y_b = \mathbf{w}_b^T \mathbf{X}_b$$

Figure 10.6: Processing of two neighboring regions of an image in accordance with the Imax principle.

10.11 Spatially Incoherent Features (1/2)

Figure 10.7: Block diagram of a neural processor, the goal of which is to suppress background clutter using a pair of polarimetric, noncoherent radar inputs; clutter suppression is attained by minimizing the mutual information between the outputs of the two modules.



$$\min_{\mathbf{w}} I(Y_a;Y_b)$$

$$C = (\mathrm{tr}[\mathbf{W}^T\mathbf{W}]-1)^2$$

$$F = I(Y_a;Y_b)+\lambda C$$

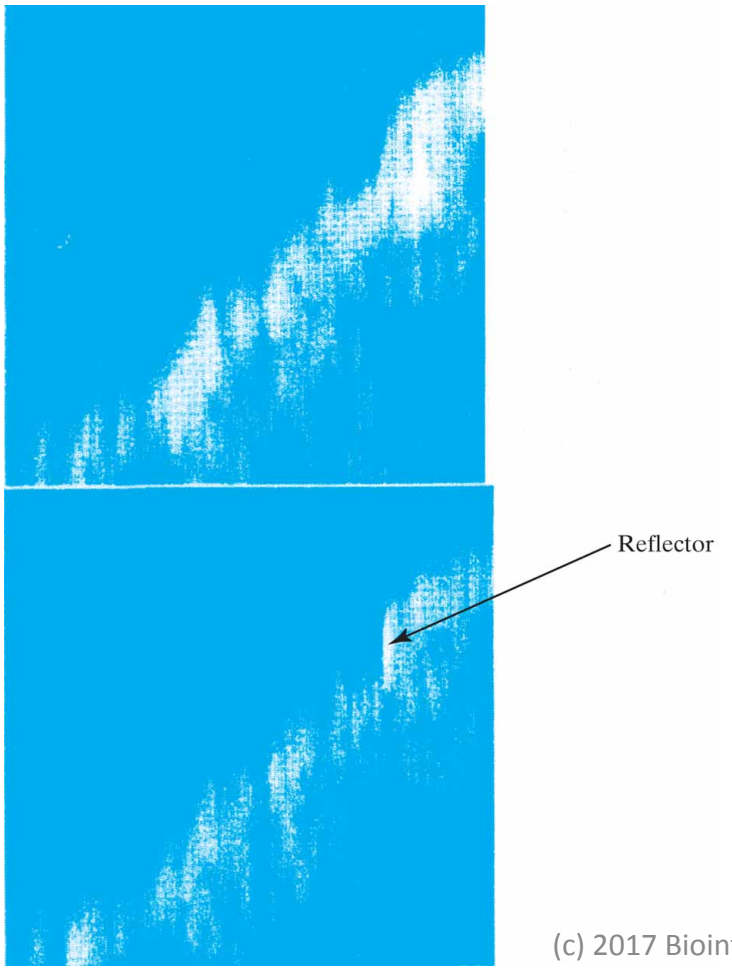
$$\frac{\partial F}{\partial \mathbf{W}} = \mathbf{0}$$

$$\frac{\partial I(Y_b;Y_b)}{\partial \mathbf{W}} + \lambda \frac{\partial C}{\partial \mathbf{W}} = \mathbf{0}$$

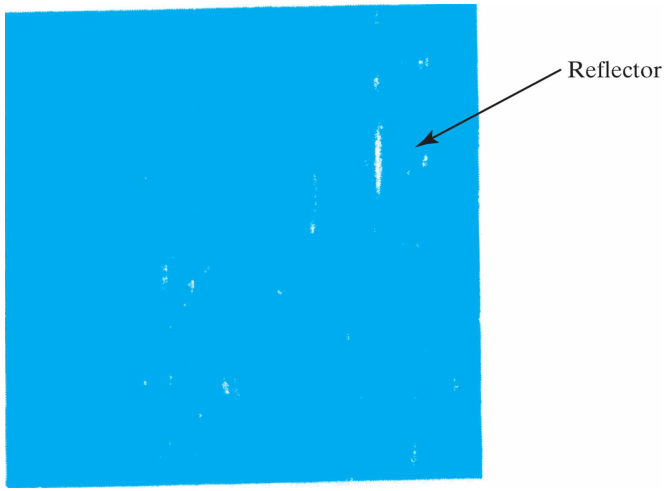
10.11 Spatially Incoherent Features (2/2)

Figure 10.8: Application of the Imin principle to radar polarimetry. (a) Raw B-scan radar images (azimuth plotted versus range) for horizontal– horizontal polarization (top) and horizontal–vertical (bottom) polarization. (b) Composite image computed by minimizing the mutual information between the two polarized radar images of part (a).

(a)



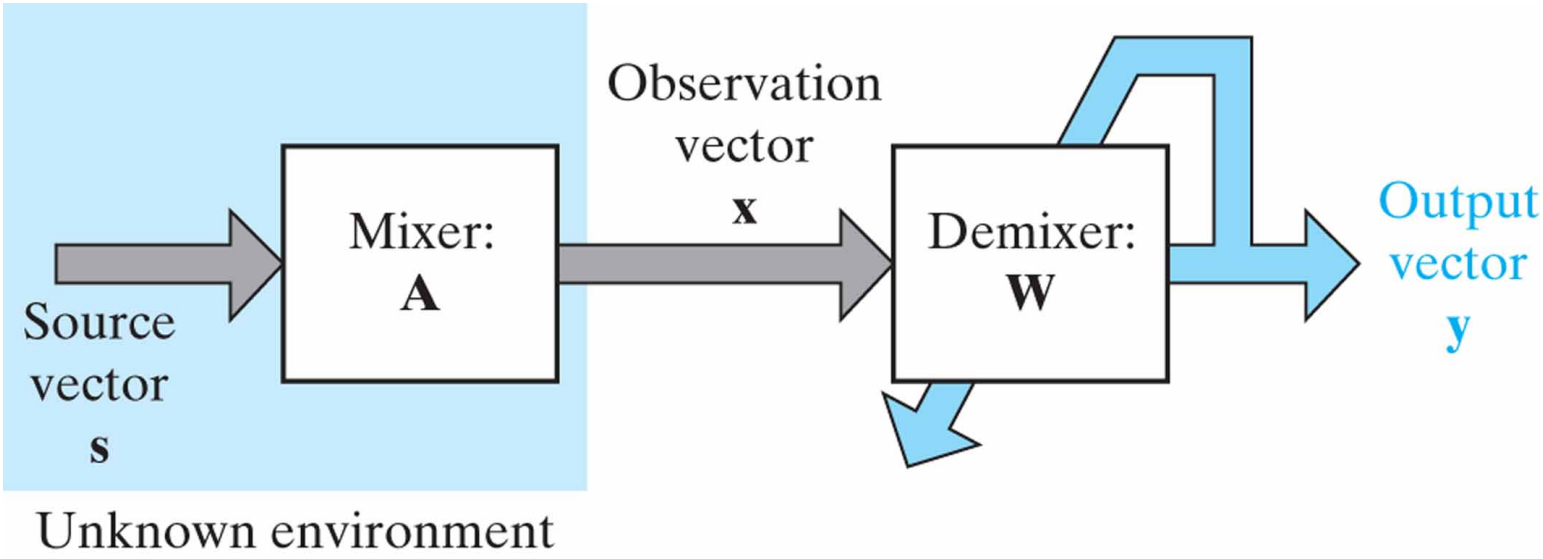
(b)



$$\min_{\mathbf{w}} I(Y_a; Y_b)$$

10.12 Independent-Components Analysis (1/3)

Figure 10.9: Block diagram of the processor for solving the blind source separation problem. The vectors s, x, and y are values of the respective random vectors S, X, and Y.



$$\mathbf{X} = \mathbf{AS} = \sum_{i=1}^{m} \mathbf{a}_i S_i$$

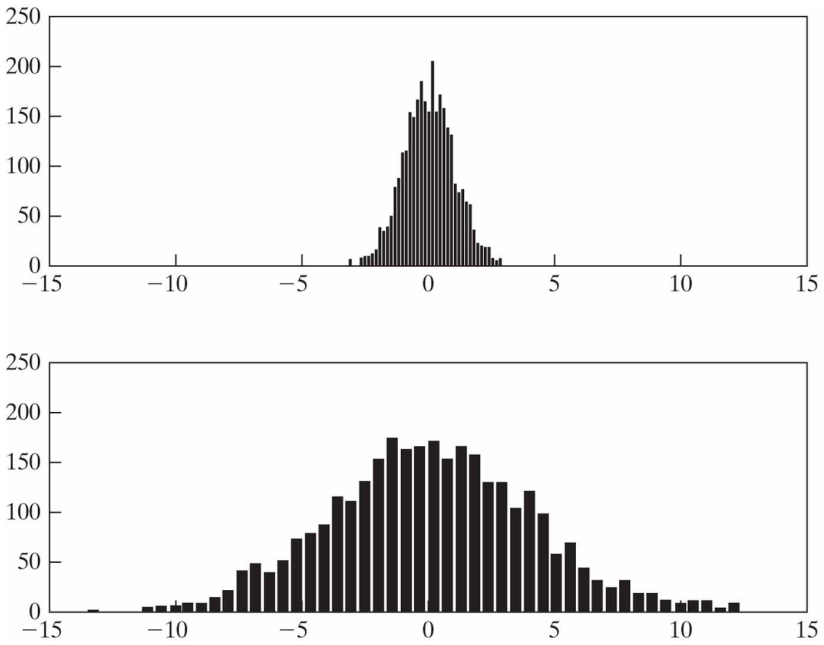
$$\mathbf{Y} = \mathbf{WX}$$

Solution to BSS by ICA

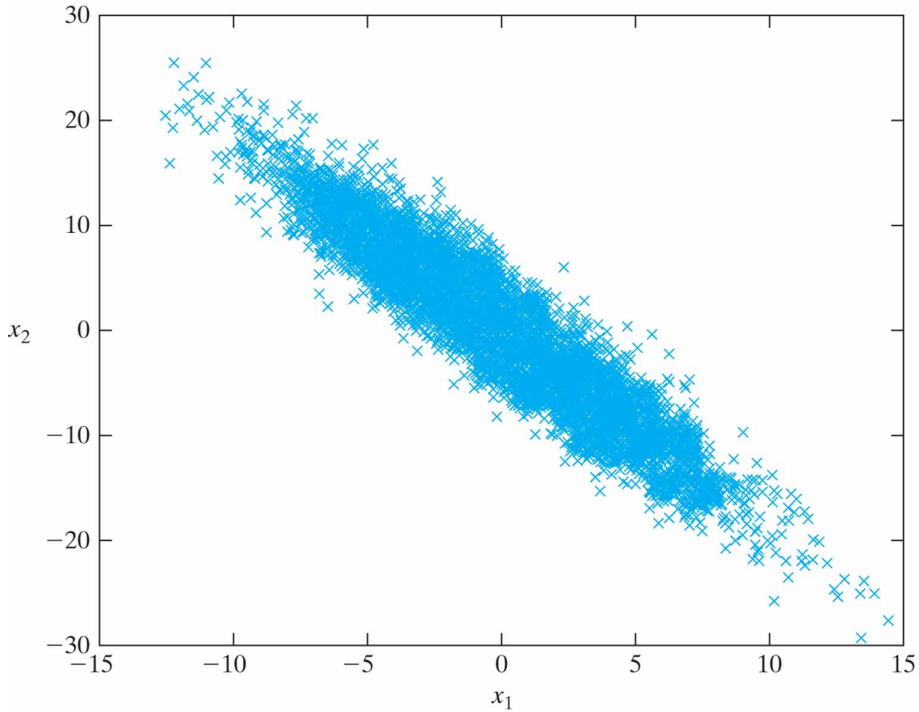
$$\mathbf{y} = \mathbf{Wx} = \mathbf{WAs} = \mathbf{DPs}$$

10.12 Independent-Components Analysis (2/3)

Figure 10.10: Two Gaussian distributed processes.



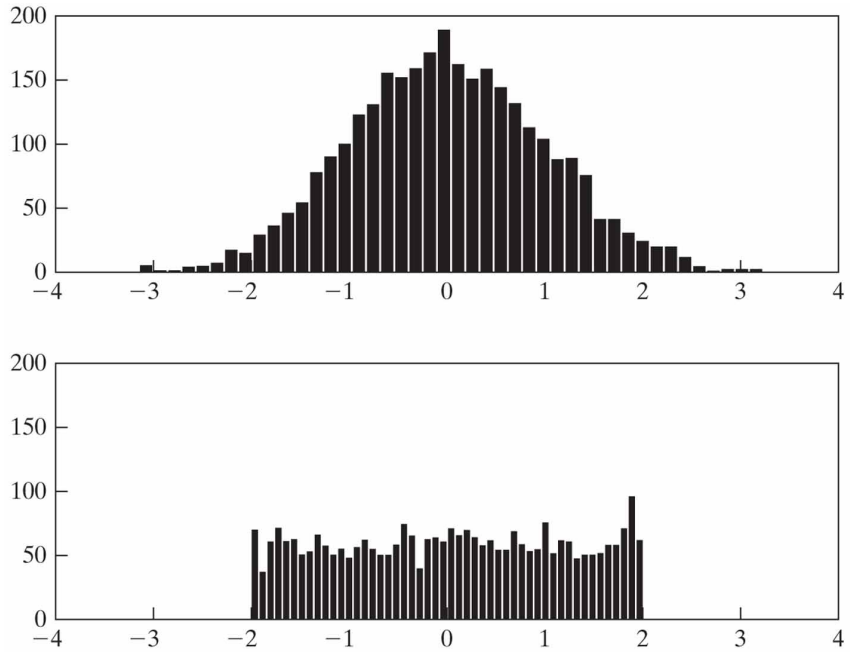
(a) Histograms of the two processes: The top histogram refers to Gaussian signal source S_1 of zero mean and veriance $\sigma_1^2 = 1$; the bottom one refers to Gaussian source signal S_2 of zero mean and veriance $\sigma_2^2 = 16$.



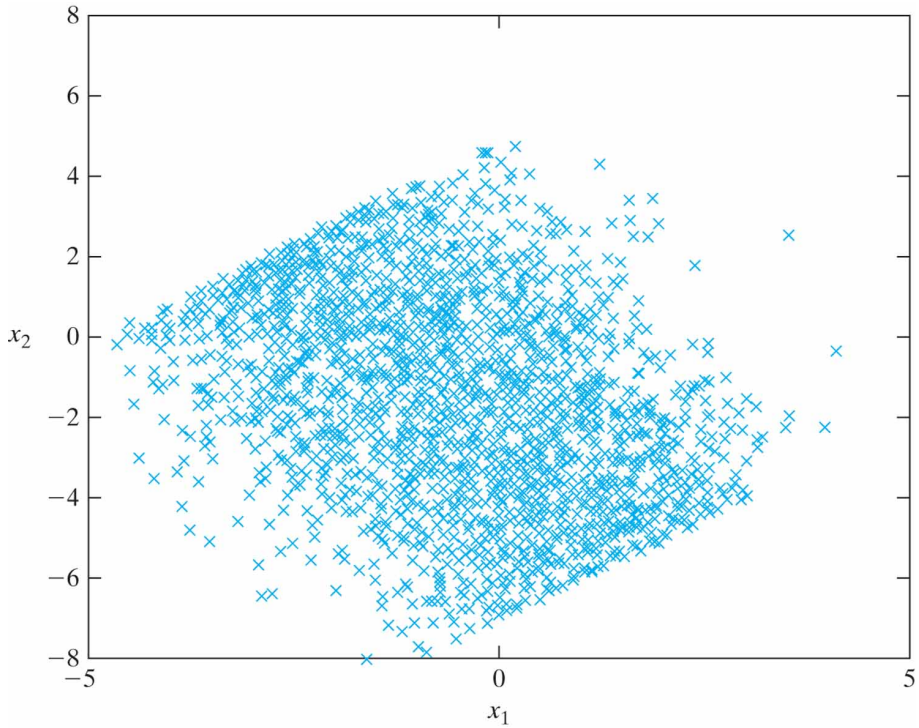
(b) Two-dimensional distribution of the linearly mixed signals X_1 and X_2.

10.12 Independent-Components Analysis (3/3)

Figure 10.11: Gaussian- and uniformly-distributed processes.



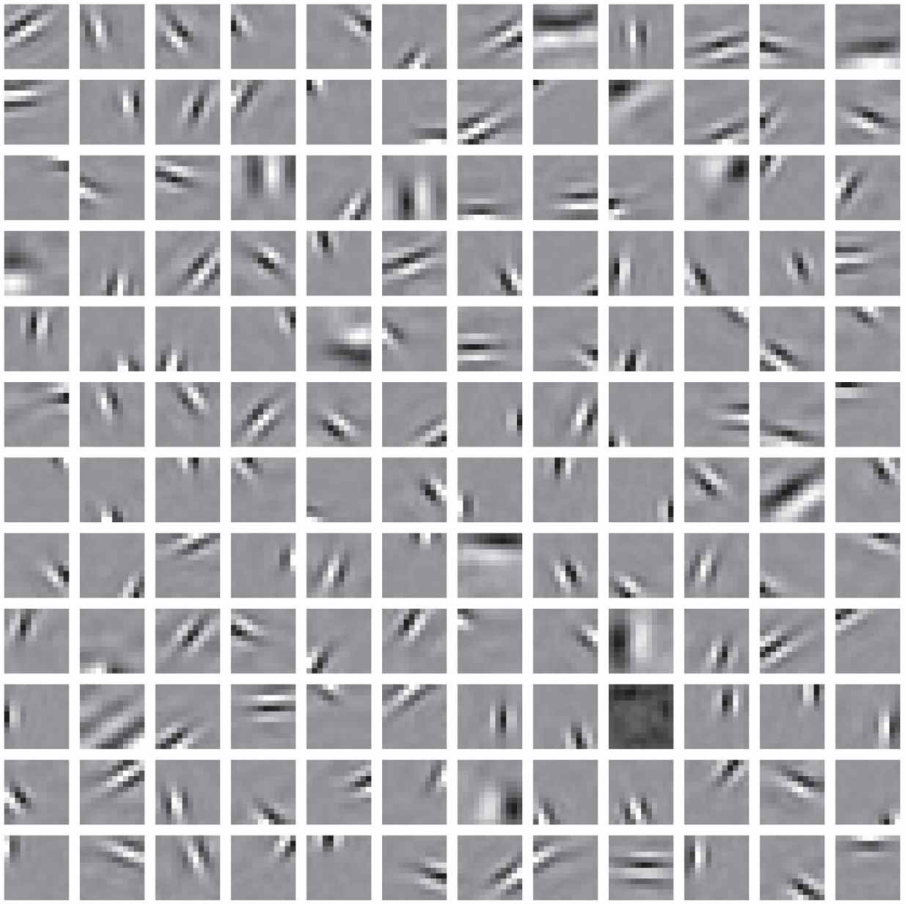
(a) Histograms of the two processes: The top histogram refers to Gaussian source signal S_1 of zero mean and veriance σ_1^2; the bottom one refers to uniformly distributed source signal S_2 uniformly distributed over the interval $[-2, 2]$.



(b) Two-dimensional distribution of the linearly mixed signals X_1 and X_2.

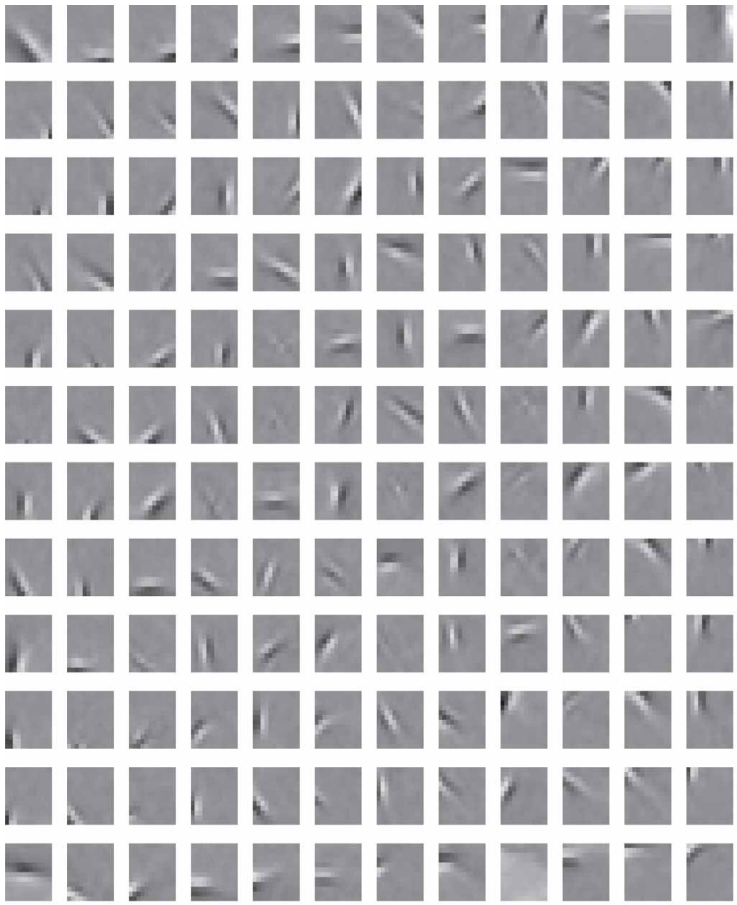
10.13 Sparse Coding of Natural Images (1/2)

Figure 10.12: The result of applying the sparse-coding algorithm to a natural image. (The figure is reproduced with the permission of Dr. Bruno Olshausen.)



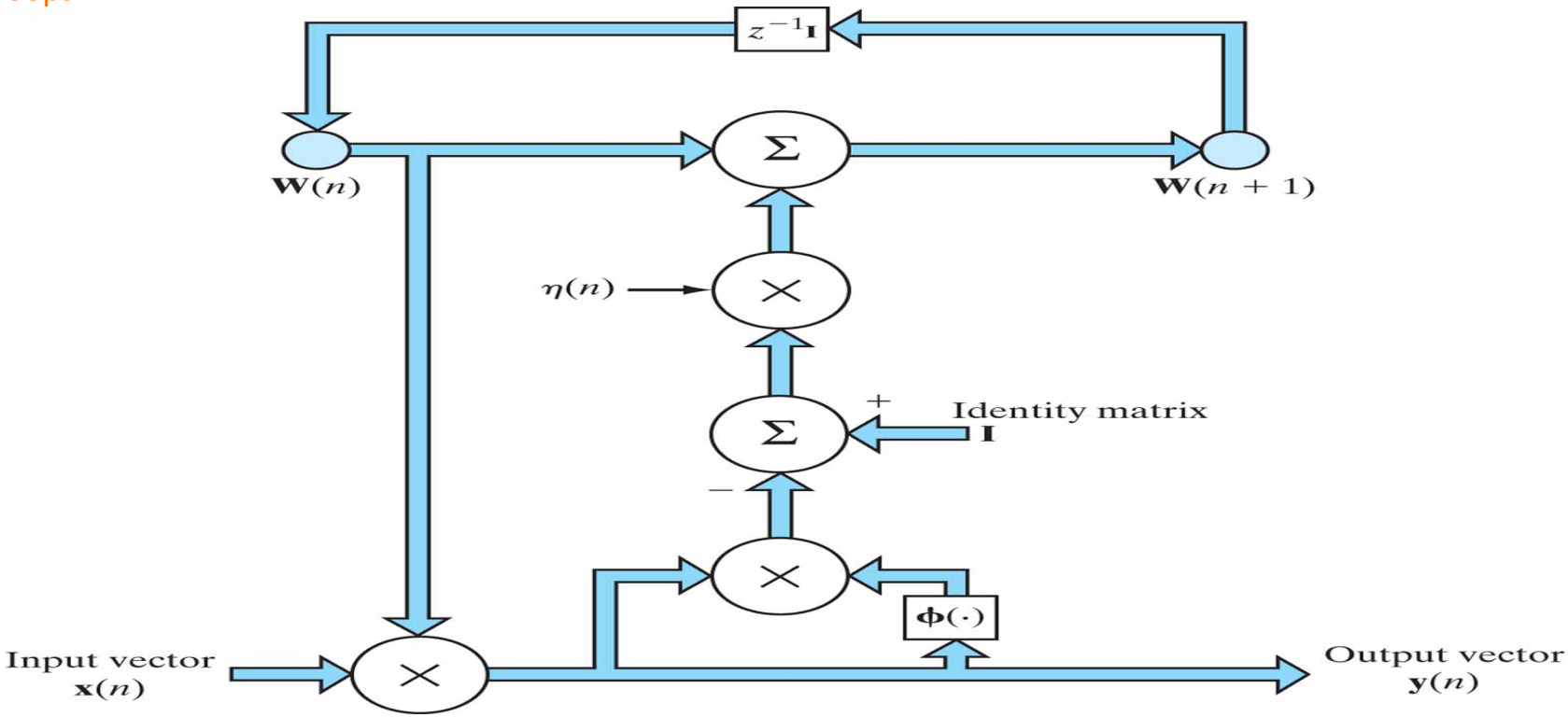
10.13 Sparse Coding of Natural Images (2/2)

Figure 10.13: The result of applying the Infomax algorithm for ICA to another natural image. (The figure is reproduced with the permission of Dr. Anthony Bell.)



10.14 Natural Gradient Learning for ICA

Figure 10.14: Signal-flow graph of the blind source separation learning algorithm described in Eqs. (10.85) and (10.104): The block labeled z–1I represents a bank of uni-time delays. The graph embodies a multiplicity of feedback loops.



$$\mathbf{W}(n+1)=\mathbf{W}(n)+\eta(n)\left[\mathbf{I}-\Phi(\mathbf{y}(n))\mathbf{y}^{T}(n)\right]\mathbf{W}(n)$$

10.19 Rate Distortion Theory and Information Bottleneck (1/3)

$\mathbf{T}$: compressed version of $\mathbf{X}$

Mutual information btw $\mathbf{X}$ and $\mathbf{T}$

$$I(\mathbf{X};\mathbf{T}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) \log\left(\frac{q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})}{p_{\mathbf{T}}(\mathbf{t})}\right) d\mathbf{x}d\mathbf{t}$$

Expected distortion

$$\mathbf{E}[d(\mathbf{x},\mathbf{t})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) d(\mathbf{x},\mathbf{t}) d\mathbf{x}d\mathbf{t}$$

Rate distortion theory

Find the rate distortion function

$$R(D) = \min_{q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})} I(\mathbf{X};\mathbf{T})$$

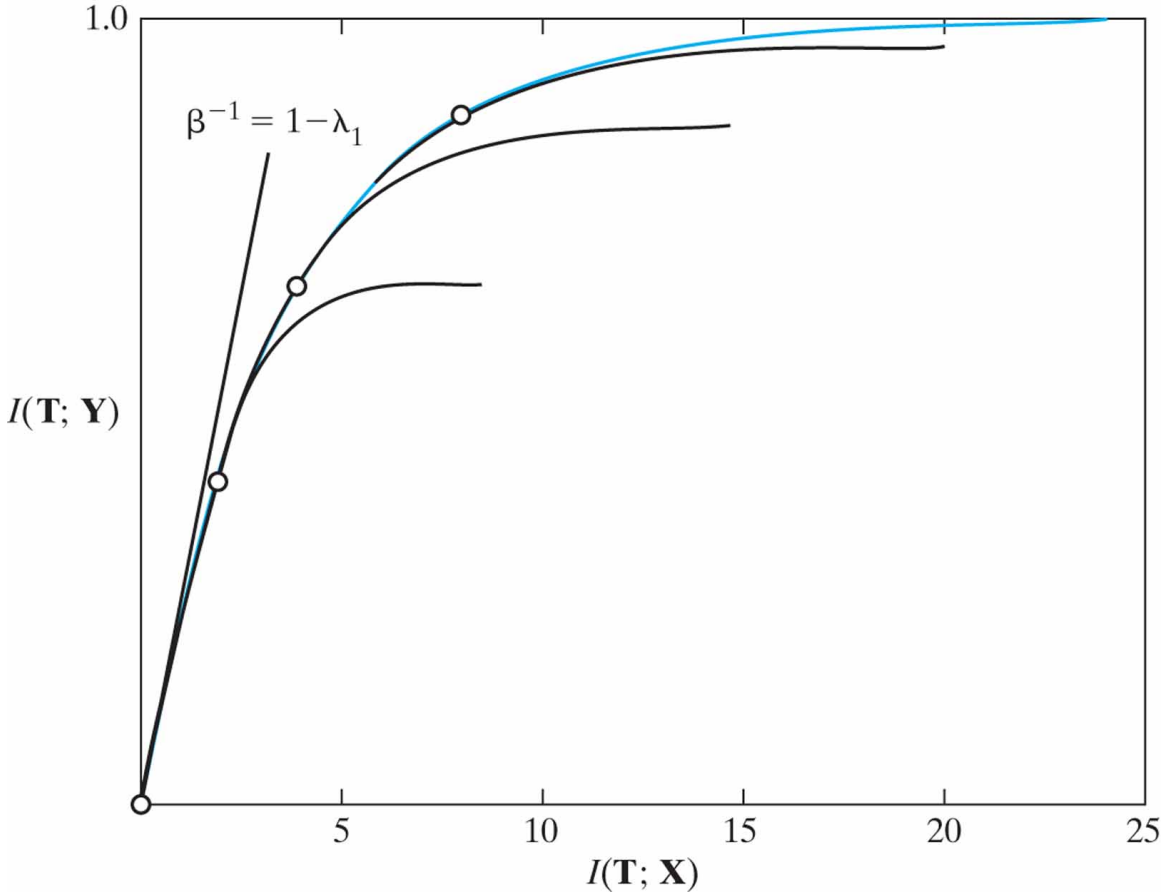
subjet to the distortion constaint

$$\mathbf{E}[d(\mathbf{x},\mathbf{t})] \le D$$

Minimize the mutual information between the source ***X*** and its representation ***T***, subject to a prescribed distortion constraint. (constrained optimization problem)

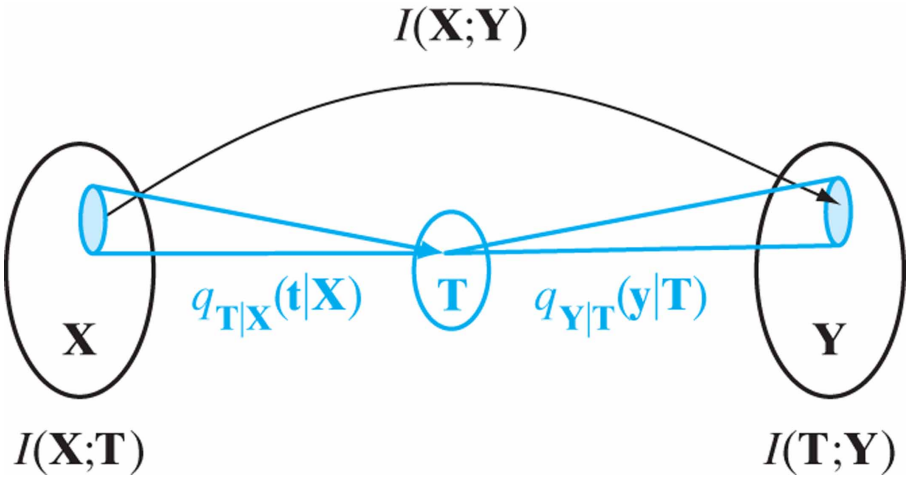
10.19 Rate Distortion Theory and Information Bottleneck (2/3)

Figure 10.21: The information curve for multivariate Gaussian variables. The envelope (blue curve) is the optimal compression–prediction tradeoff, captured by varying the Lagrange multiplier β from zero to infinity. The slope of the curve at each point is given by 1/β.There is always a critical lower value of β that determines the slope at the origin, below which there are only trivial solutions. The suboptimal (black) curves are obtained when the dimensionality of T is restricted to fixed lower values. (This figure is reproduced with the permission of Dr. Naftali Tishby.)



10.19 Rate Distortion Theory and Information Bottleneck (3/3)

Figure 10.22: An illustration of the information bottleneck method. The bottleneck T captures the relevant portion of the original random vector X with respect to the relevant variable Y by minimizing the information I(X;T) while maintaining I(T;Y) as high as possible. The bottleneck T is determined by the three distributions , , and , which represent the solution of the bottleneck equations (10.170) to (10.172).



$$q_{T|X}(t|x) = \frac{q_T(t)}{Z(x,\beta)} \exp(-D_{p||q})$$

$$q_T(t) = \sum_x q_{T|X}(t|x) p_X(x)$$

$$q_{Y|T}(y|t) = \sum_x q_{Y|T}(y,x|t)$$

$$= \sum_x q_{Y|T}(y|t) q_{X|T}(x|t)$$

$$= \sum_x q_{Y|T}(y|t) q_{T|X}(t|x) \left(\frac{p_X(x)}{q_T(t)} \right)$$

Information bottleneck method:

Find representation **T** that maximizes

$$J(q_{T|X}(t|x)) = I(X;T) - \beta I(T;Y)$$

Bayes rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

10.20 Optimal Manifold Representation of Data (1/7)

$q_{\mathrm{M|X}}(\boldsymbol{\mu}|\mathbf{x})$: conditional pdf of points on the manifold

Stochastic map

$$P_{\mathrm{M}}:\mathbf{x}\rightarrow q_{\mathrm{M|X}}(\boldsymbol{\mu}|\mathbf{x})$$

Distance measure

$$d(\mathbf{x},\boldsymbol{\mu})=\|\mathbf{x}-\boldsymbol{\mu}\|^2$$

Expected distortion

$$\mathbf{E}[d(\mathbf{x},\boldsymbol{\mu})]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{\mathrm{X}}(\mathbf{x})q_{\mathrm{M|X}}(\boldsymbol{\mu}|\mathbf{x})\|\mathbf{x}-\boldsymbol{\mu}\|^2\,d\mathbf{x}d\boldsymbol{\mu}$$

Mutual information between the manifold M and the datas set X

$$I(\mathrm{X};\mathrm{M})=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{\mathrm{X}}(\mathbf{x})q_{\mathrm{M|X}}(\boldsymbol{\mu}|\mathbf{x})\log\left(\frac{q_{\mathrm{M|X}}(\boldsymbol{\mu}|\mathbf{x})}{q_{\mathrm{M}}(\boldsymbol{\mu})}\right)d\mathbf{x}d\boldsymbol{\mu}$$

i.e. the number of bits required to encode a data point $\mathbf{x}$ into a point $\boldsymbol{\mu}$ on the manifold M.

10.20 Optimal Manifold Representation of Data (2/7)

Tradeoff:

1. Faithful representation of data: **minimize distortion**
2. Good compression of data: **maximize MI**

The manifold is optimal if the channel capacity $I(\mathrm{X};\mathrm{M})$ is maximized while the expected distortion $\mathbf{E}[d(\mathbf{x},\boldsymbol{\mu})]$ is fixed at some prescribed value.

Constrained optimization problem: minimize F

$$F(\mathrm{M},P_{\mathrm{M}}) = \mathbf{E}[d(\mathbf{x},\boldsymbol{\mu})] + \lambda I(\mathrm{X};\mathrm{M})$$

Parameterize the manifold and introduce the bottleneck vector $\mathbf{T}$

$$\boldsymbol{\gamma}(\mathbf{t}):\ \mathbf{t} \rightarrow \mathrm{M}$$

$\boldsymbol{\gamma}(\mathbf{t})$: descriptor of the manifold M

New distance measure

$$d(\mathbf{x},\boldsymbol{\gamma}(\mathbf{t})) = \| \mathbf{x} - \boldsymbol{\gamma}(\mathbf{t}) \|^2$$

10.20 Optimal Manifold Representation of Data (3/7)

Expected distortion and MI (channel capacity)

$$\mathbf{E}[d(\mathbf{x},\boldsymbol{\gamma}(\mathbf{t}))] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) \|\mathbf{x}-\boldsymbol{\gamma}(\mathbf{t})\|^2 \, d\mathbf{x} d\mathbf{t}$$

$$I(\mathbf{X};\mathbf{T}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) \log\left(\frac{q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})}{q_{\mathbf{T}}(\mathbf{t})}\right) d\mathbf{x} d\mathbf{t}$$

Functional F to be minimized

$$F(\boldsymbol{\gamma}(\mathbf{t}), q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})) = \mathbf{E}[d(\mathbf{x},\boldsymbol{\gamma}(\mathbf{t}))] + \lambda I(\mathbf{X};\mathbf{T})$$

To find the optimal manifold, we consider two conditions

1. $\dfrac{\partial F}{\partial \boldsymbol{\gamma}(\mathbf{t})} = \mathbf{0}$ for $q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})$ fixed
2. $\dfrac{\partial F}{\partial q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})} = 0$ for $\boldsymbol{\gamma}(\mathbf{t})$ fixed

10.20 Optimal Manifold Representation of Data (4/7)

Applying condition 1, we obtain

$$\frac{\partial F}{\partial \boldsymbol{\gamma}(\mathbf{t})} = \frac{\partial \mathbf{E}[d(\mathbf{x}, \boldsymbol{\gamma}(\mathbf{t}))]}{\partial \boldsymbol{\gamma}(\mathbf{t})} = \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})(-2\mathbf{x} + 2\boldsymbol{\gamma}(\mathbf{t}))d\mathbf{x} = \mathbf{0}$$

From this we obtain

$$\boldsymbol{\gamma}(\mathbf{t}) = \frac{1}{q_{\mathbf{T}}(\mathbf{t})} \int_{-\infty}^{\infty} \mathbf{x} p_{\mathbf{X}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) d\mathbf{x}$$

$$q_{\mathbf{T}}(\mathbf{t}) = \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) d\mathbf{x}$$

To apply the condition 2, we have the additional constraint

$$\int_{-\infty}^{\infty} q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) d\mathbf{t} = 1 \qquad \text{for all } \mathbf{x}$$

To satisfy this additional constraint, we introduce the new Lagrangean multiplier $\beta(\mathbf{x})$.

10.20 Optimal Manifold Representation of Data (5/7)

The new expanded functional F

$$F(\boldsymbol{\gamma}(\mathbf{t}), q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big\{p_{\mathbf{X}}(\mathbf{x})q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})\|\mathbf{x}-\boldsymbol{\gamma}(\mathbf{t})\|^2 + \lambda p_{\mathbf{X}}(\mathbf{x})q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})\log\left(\frac{q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})}{q_{\mathbf{T}}(\mathbf{t})}\right) + \beta(\mathbf{x})q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})\Big\}d\mathbf{t}d\mathbf{x}$$

$\mathbf{E}[d(\mathbf{x},\boldsymbol{\gamma}(\mathbf{t}))]$

$I(\mathbf{X};\mathbf{T})$

Applying condition 2, we obtain

$$\frac{1}{\lambda}\|\mathbf{x}-\boldsymbol{\gamma}(\mathbf{t})\|^2 + \log\left(\frac{q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})}{q_{\mathbf{T}}(\mathbf{t})}\right) + \frac{\beta(\mathbf{x})}{\lambda p_{\mathbf{X}}(\mathbf{x})} = 0$$

Setting $\dfrac{\beta(\mathbf{x})}{\lambda p_{\mathbf{X}}(\mathbf{x})} = \log Z(\mathbf{x},\lambda)$ and solving for $q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x})$, we get

$$q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) = \frac{q_{\mathbf{T}}(\mathbf{t})}{Z(\mathbf{x},\lambda)}\exp\left(-\frac{1}{\lambda}\|\mathbf{x}-\boldsymbol{\gamma}(\mathbf{t})\|^2\right)$$

$$Z(\mathbf{x},\lambda) = \int_{-\infty}^{\infty} q_{\mathbf{T}}(\mathbf{t})\exp\left(-\frac{1}{\lambda}\|\mathbf{x}-\boldsymbol{\gamma}(\mathbf{t})\|^2\right)d\mathbf{t}$$

10.20 Optimal Manifold Representation of Data (6/7)

Discrete approximation with $\delta(\cdot)$ the Dirac delta funcrtion

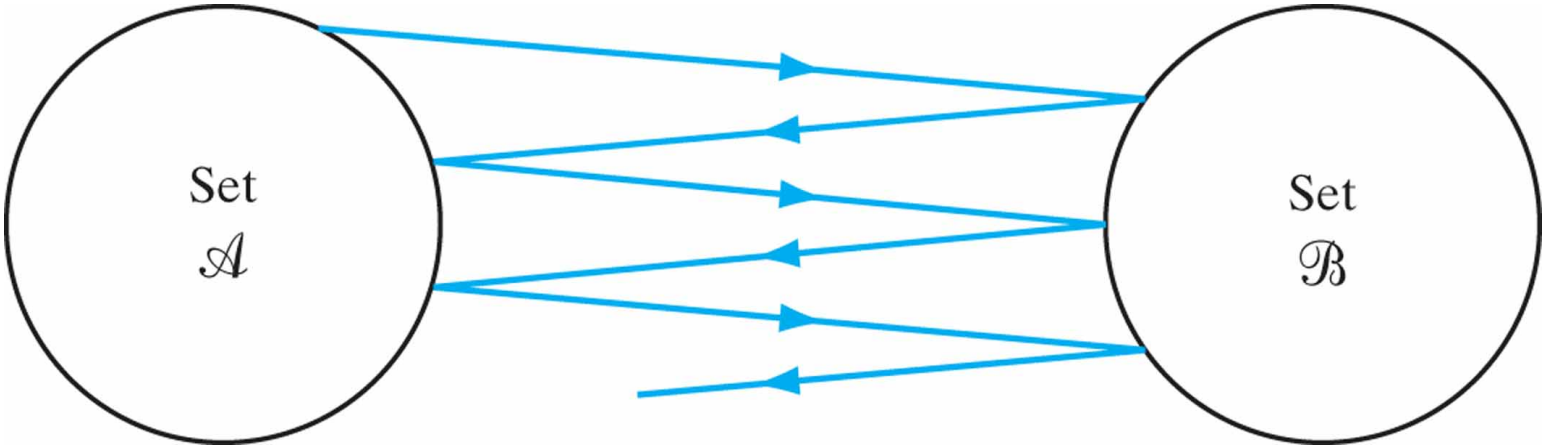
$$p_{\mathbf{x}}(\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{x}-\mathbf{x}_i)$$

Using the L-point discrete set $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_L\}$ to model the manifold represented by the continuous variable $\mathbf{t}$.

We model the manifold M by the discrete set $\mathrm{T} = \{\mathbf{t}_j\}_{j=1}^{L}$

$$\boldsymbol{\gamma}(t) \Rightarrow \boldsymbol{\gamma}_j, \quad q_{\mathbf{T}|\mathbf{X}}(\mathbf{t}|\mathbf{x}) \Rightarrow q_j(\mathbf{x}_j), \quad q_{\mathbf{T}}(\mathbf{t}) \Rightarrow q_j$$

Figure 10.23: Illustrating the alternating process of computing the distance between two convex sets A and B.



10.20 Optimal Manifold Representation of Data (7/7)

Iterative algorithm for computing the discrete model fo the manifold:

$$p_j(n)=\frac{1}{N}\sum_{i=1}^{N}p_j(\mathbf{x}_i,n)$$

$$\gamma_{j,\alpha}(n)=\frac{1}{p_j(n)}\cdot\frac{1}{N}\sum_{i=1}^{N}x_{i,\alpha}p_j(\mathbf{x}_i,n),\qquad \alpha=1,2,...,m$$

$$Z(\mathbf{x}_i,\lambda,n)=\sum_{j=1}^{L}p_j(n)\exp\left(-\frac{1}{\lambda}\|\mathbf{x}-\boldsymbol{\gamma}_j(n)\|^2\right)$$

$$p_j(\mathbf{x}_i,n+1)=\frac{p_j(n)}{Z(\mathbf{x}_i,\lambda,n)}\exp\left(-\frac{1}{\lambda}\|\mathbf{x}-\boldsymbol{\gamma}_j(n)\|^2\right)$$

Initialization

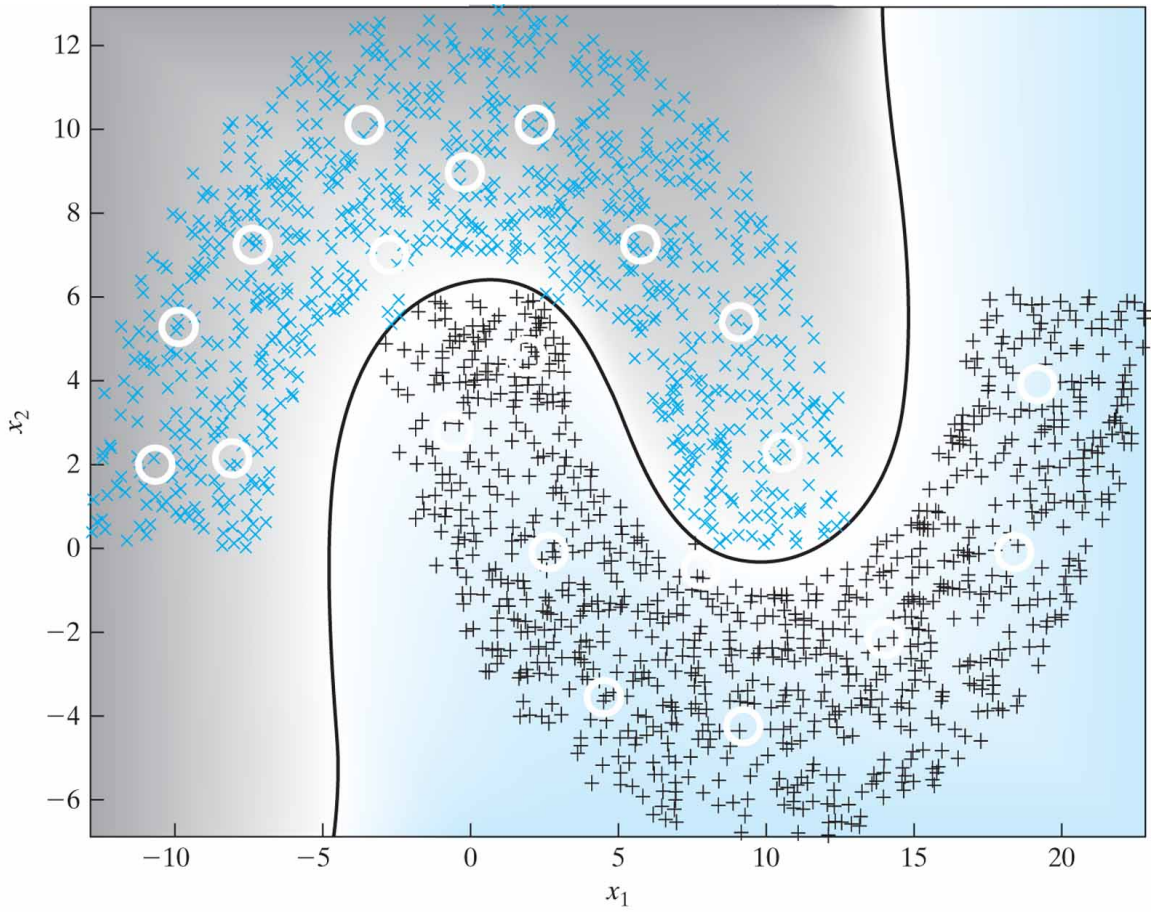
$$\boldsymbol{\gamma}_j=x_{i,j},\qquad p_j(0)=1/L,\qquad j=1,2,...,L$$

Termination condition

$$\max_j|\boldsymbol{\gamma}_j(n)-\boldsymbol{\gamma}_j(n-1)|<\varepsilon$$

10.21 Computer Experiments: Pattern Classification

Figure 10.24: Pattern classification of the double-moon configuration of Fig. 1.8, using the optimal manifold + LMS algorithm with distance d = –6 and 20 centers.



Summary and Discussion (Ch. 10)

- Information theory and entropy
  - Uncertainty, probability, information, entropy
  - The maximum entropy principle (Max Ent)
  - Mutual information (MI)
  - Kullback-Leibler divergence (KL)
- Mutual information as the objective function of self-organization
  1. The Infomax principle
  2. The principle of minimum redundancy
  3. The Imax principle
  4. The Imin principle
- Applications to machine learning
  - Independent-Components Analysis
  - Information Bottleneck
  - Manifold Learning