Pattern Recognition[1]

Dokuz Eylül University
Computer Engineering Department
24 October 2001 –

Akira Imada
imada@cs.deu.edu.tr

[1] I pick up some of the topics from the following book:

- S. Theodoridis, and K. Koutroumbas (1999) *Pattern Recognition.* Academic Press.

which is available, for example, at http://www.amazon.com.

□ What is Pattern Recognition?

- A typical example:
  - Hand-written character recognition:
    with inputs being pixel values.

- What else?
  - Face recognition;
  - Speech recognition;
  - Robot's eye/ear.

- Can we recognize them with fewer information?

  - Yes!
  - To recognize hand-written characters, e.g.,
    * Perimeter of the boundary:

$$P_i = \sum_{i=1}^{N-1} \|x_{i-1} - x_i\| + \|x_0 - x_1\|. \tag{1}$$

    * Area inside the boundary: A.
    * Roundness ratio:

$$r = p^2/4\pi A. \tag{2}$$

    * Bending energy (defined at a point n):

$$E(n) = \frac{1}{P} \sum_{i=0}^{n-1} |k_i|^2 \tag{3}$$

      where k_i *(curvature of boundary)* is:

$$k_n = \theta_{n+1} - \theta n, \;\; n = 0, 1, \cdots, N-1. \tag{4}$$

      and

$$\theta_n = \tan^{-1} \frac{y_{n+1} - y_n}{x_{n+1} - x_n}, \;\; n = 0, 1, \cdots, N-1. \tag{5}$$

    * Number of halls.
    * Number of corners:

      the number of points where the curvature k_i takes large values (infinity in theory)

  - They are called *features*.
  - Then, how many features are enough?
    (Minimize the number of features.)

In short:

- pattern recognition is a *classifier*.
  - Especially usfull for classification in high-dimensional *feature space*.
- Application:
  - diagnosis.
  - what else?

□ **Statistical Classification:**

Bayes Decision Theory

— for the design of classifiers.

Pattern Recognition is more or less statistical due to:

– statistic variation of patterns;

– statistic nature of feature selection;

– noise in employing sensors.

So, the task is:

– to classify unknown patterns into the *most probable class.*

Then

– what does "the most probable" means?

To answer this question:

– we study here the Bayesian Statistical Theory.

- When we classify an unknown pattern which is represented by a feature vector $\mathbf{x}$, we consider M conditional probabilities:

$$p(\omega_i|\mathbf{x}) \quad i = 1, \cdots, M \tag{6}$$

- Then we define "the most probable class" as the class ω_i which has the highest $p(\omega_i|\mathbf{x})$.

- We have the other possible statistical quantities:

$$p(\omega_i), \quad p(\mathbf{x}|\omega_i), \quad p(\mathbf{x})$$

- The *Bayesian rule* gives a relationship of these quantities:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{\sum_{k=1}^{M} p(\mathbf{x}|\omega_k)p(\omega_k)} \tag{7}$$

  where the denominator is:

$$\sum_{k=1}^{M} p(\mathbf{x}|\omega_k)p(\omega_k) = p(\mathbf{x}). \tag{8}$$

• Example 1.
— to understand the Bayesian Rule

· We have two bags of no difference from its outlook.

· One bag called R has 70 red balls and 30 blue balls.

· The other bag called B has 30 red balls and 70 blue balls.

· When we take one bag at random and pick up 12 balls, returning it to the bag at each time.

· The result was 8 red balls and 4 blue one.

· Then was the bag estimated to be R or B, and how probable the estimate is?

Clearly,

$$p(R) = p(B) = 1/2$$

$$\begin{aligned} p(D|R) &= {}_{12}C_8(0.7)^8(0.3)^4 \\ p(D|B) &= {}_{12}C_8(0.3)^8(0.7)^4 \end{aligned}$$

So we obtain

$$p(R|D) = \frac{(0.7)^8(0.3)^4}{(0.7)^8(0.3)^4 + (0.3)^8(0.7)^4} \approx 0.97$$

(How big it is compared to our intuitive estimation!)

• Example 2.

- Three prisoners (**A**, **B**, and **C**) are in a prison.
- **A** knows that the two out of the three are to be executed tomorrow, and the rest becomes free.
- **A** thought either one of **B** or **C** is sure to be executed.
- Then, **A** asked a guard "even if you tell me which of **B** and **C** is executed, that will not give me any information as for me. So please tell it to me."
- The guard answers that **C** will. $\Rightarrow$ data D
- Now, **A** knows one of **A** or **B** is sure to be free.

Do you guess probability $p(A|D) = 1/2$?

If this is correct, then the answer of the guard had given an information as for A, since probability $p(A) = 1/3$.

You agree that

$$p(A) = p(B) = p(C) = 1/3.$$

Then, try to apply Bayesian rule, i.e., obtain the conditional probability of the data "C will be executed" under the condition that "A will be free tomorrow" And in the same way for B and C. They are:

$$\begin{aligned} p(D|A) &= 1/2. \\ p(D|B) &= 1. \\ p(D|C) &= 0. \end{aligned}$$

In conclusion:

$$p(A|D) = \frac{p(D|A)p(A)}{p(D|A)p(A) + p(D|B)p(B) + p(D|C)p(C)} = 1/3.$$

This shows probability did not change after the information!

- Now it's clear that
  - $p(A)$ and so on are to be called
    - ⋆ *a priori* probabilities;
  - and $p(A|D)$ and so on to be
    - ⋆ *a posteriori* probabilities.

- In the same way,
  - $p(\omega_i)$ is called
    - ⋆ *a priori* probability[2] ;
  - $p(\omega_i|\mathbf{x})$
    - ⋆ *a posteriori* probabilities.

  Furthermore
  - $p(\mathbf{x})$ is called
    - ⋆ p.d.f. of $\mathbf{x}$
  - $p(\mathbf{x}|\omega_i)$
    - ⋆ class-conditional p.d.f.[3]
      - · which describes the distribution of the feature vectors $\mathbf{x}$ in each of the classes ω_i.

[2] Usually given, but if unknown, it can be estimated as N_i/N where N_i is the number of training samples which belong to class ω_i, and N is the total number of training samples

[3] This is also estimated from training data which will be explained later more in detail.

□ The Bayesian Classifier:

Definition 1 (The Bayesian Classifier) *classify* $\mathbf{x}$ *to the class* ω_i *such that* $p(\omega_i|\mathbf{x})$ *takes the maximum value.*

• Two-class case:

$$\star \text{ if } p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x}), \quad \text{then classify } \mathbf{x} \text{ to } \omega_1$$

$$\star \text{ if } p(\omega_1|\mathbf{x}) < p(\omega_2|\mathbf{x}), \quad \text{then classify } \mathbf{x} \text{ to } \omega_2$$

The region where $p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x})$ is the region where

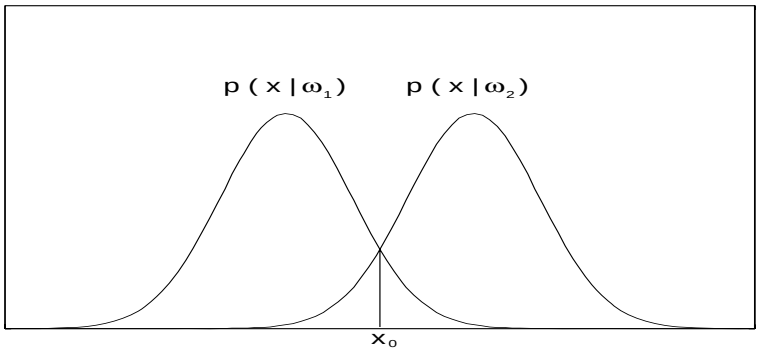
$$\frac{p(\mathbf{x}|\omega_1)p(\omega_1)}{p(\mathbf{x})} > \frac{p(\mathbf{x}|\omega_2)p(\omega_2)}{p(\mathbf{x})} \tag{9}$$

Since $p(\mathbf{x}) > 0$ and $p(\omega_1) = p(\omega_2)$, the region is where

$$p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2) \tag{10}$$

holds. Now assume the two Gaussian distribution cross at $x = x_0$, we can conclude:

$\star$ if $x < x_0$ then classify $\mathbf{x}$ to class ω_1, and vice versa.



Excersise 1 *If we assume that* $p(\omega_1|\mathbf{x}) = N(0, 1/2)$[4] , $p(\omega_2|\mathbf{x}) = N(1, 1/2)$ *and* $p(\omega_1) = p(\omega_2) = 1/2$*, then how is our classification like?*

[4] $N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}$

□ l-dimensional Gaussian distribution:

- What does it mean by Gaussian distribution of a vector $\mathbf{x}$?

$$\mathbf{x} = (x_1, x_2, \cdots, x_l)$$

Assume we have p samples each of x_i has a value x_{ik} $(k = 1, 2, \cdots, p)$ such as:

	x_{1k}	x_{2k}	$\cdots$	x_{lk}
$k=1$	3.2	7.4		9.4
$k=2$	2.3	8.1		8.7
			$\cdots$	
$k=p$	2.9	9.2		7.9
μ_i				
σ_i^2				

You may fill the blancs for μ_i and σ_i.

Then you define the pdf as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)) \tag{11}$$

where

$$\mu = (\mu_1, \mu_2, \cdots, \mu_l)$$

and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2l} \\ & & \cdots & \\ \sigma_{l1} & \sigma_{l2} & \cdots & \sigma_l^2 \end{pmatrix} \tag{12}$$

In summary,

- $\mathbf{x} = (x_1, x_2, \cdots, x_l)$;
- μ is a l-dimensional vector whose i-th element μ_i is the mean value of x_i, i.e.,

$$\mu_i = \frac{\sum_{k=1}^{p} x_{ik}}{p}$$

- σ_i^2 is the variance of x_i, i.e.,

$$\sigma_i^2 = \frac{\sum_{k=1}^{p} (x_{ik} - \mu_i)^2}{p}$$

- σ_{ij} is the covariance between x_i and x_j, i.e.,

$$\sigma_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} - \mu_i)(x_{jk} - \mu_j)}{p}$$

- $|\Sigma|$ is determinant of the matrix Σ;
- Σ^{-1} is inverse matrix of Σ[5] ;
- $\mathbf{x}^T$ is transverse of vector $\mathbf{x}$[6] .

[5] If we assume

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{13}$$

then

$$|A| = ad - bc \tag{14}$$

and

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{15}$$

[6] That is, when $\mathbf{x} = (x_1, x_2)$

$$\mathbf{x}^T = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{16}$$

• 2-dimensional Gaussian

We can calculate $l = 2$ version from the Equation (11) as:

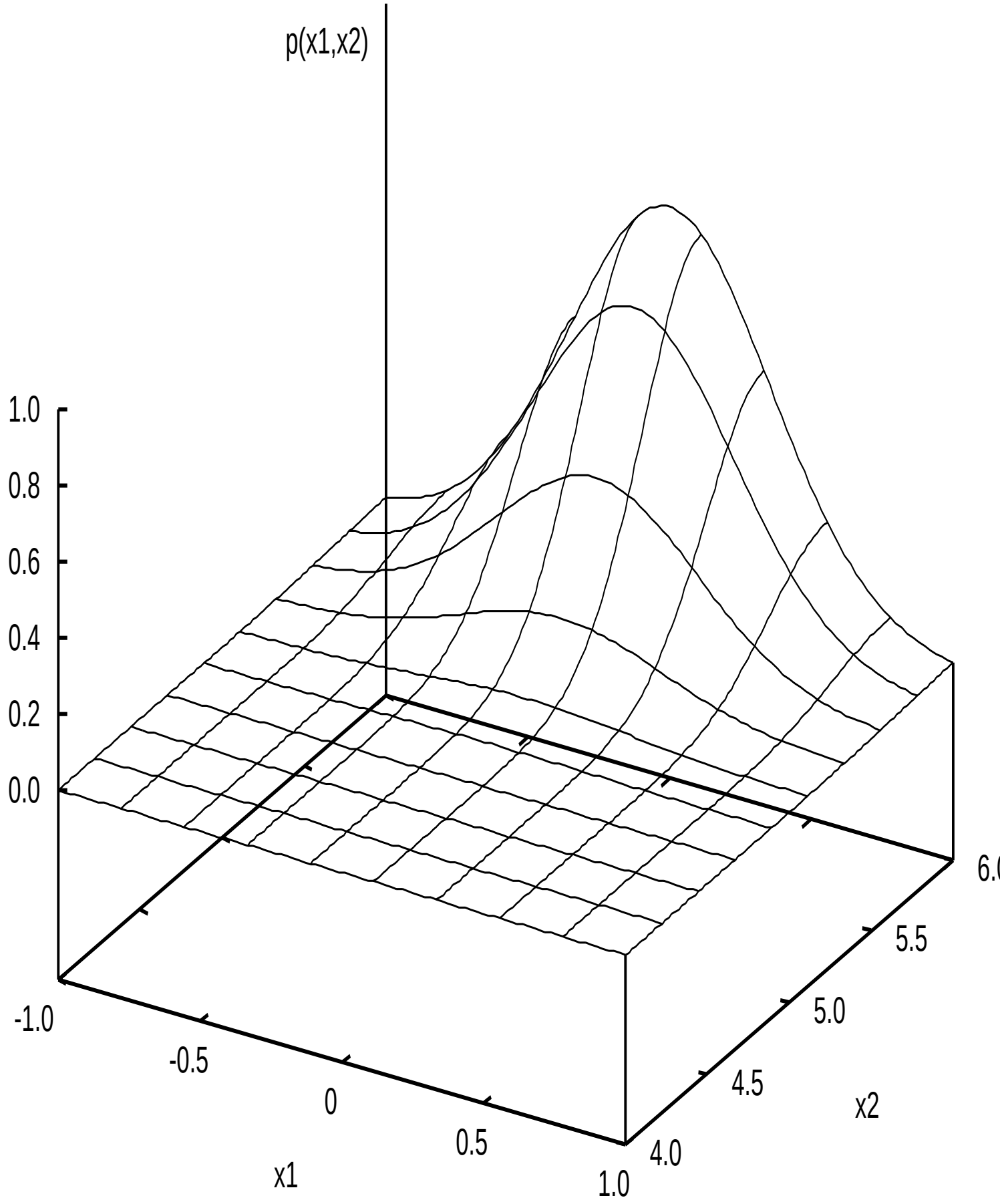
$$p(x_1 \ , \ x_2) = \frac{1}{\sqrt{2\pi}\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\{(-\frac{1}{2(1-\rho^2)} \cdot \left((\frac{x_1-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x_1-\mu_1}{\sigma_1})(\frac{x_2-\mu_2}{\sigma_2}) + (\frac{x_2-\mu_2}{\sigma_2})^2\right)\} \tag{17}$$

where ρ is correlation coefficient defined as:

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{18}$$

Excersise 2 *Try the following two problems.*

1. *Derive the equation (17) from the equation (11).*
2. *By giving your own set of five parameters:*
   μ_1, μ_2, σ_1, σ_2, *and* σ_{12}
   Draw the Gaussian surface on (x_1, x_2) *coordinate. Then explore a couple of configurations of these five parameters.*


p(x1,x2)
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.5
0
0.5
1.0
x1
4.0
4.5
5.0
5.5
6.0
x2

□ **Decision Surfaces and Discriminant Functions:**
— to partition the feature space into M regions.

- If regions R_i and R_j are contiguous

$$p(\omega_i|\mathbf{x}) - p(\omega_j|\mathbf{x}) = 0 \tag{19}$$

determines the surface that partition R_i and R_j.

- Or, if necessary, using a monotonic function $f(x)$[7] , we define:

$$g_i(x) = f(p(\omega_i|\mathbf{x})) \tag{20}$$

and we can say

$$\text{if } g_i(x) > g_j(x) \text{ classify } \mathbf{x} \text{ to } \omega_i. \tag{21}$$

Hence

$$g_{ij} \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0 \tag{22}$$

determines the *decision surfaces* separating contiguous regions and called *discriminant function.*

[7] e.g., $f(x) = \ln(x)$ for the Gaussian distribution.

□ **Examples of Decision Surface:**

• 1-D Gaussian case.

The Equation (19) in 1-D Caussian cases leads:

$$p(\omega_1|\mathbf{x}) = p(\omega_2|\mathbf{x}) \tag{23}$$

That is,

$$\frac{1}{\sqrt{2\pi}\sigma_1}\exp\{-\frac{(x-\mu_1)^2}{2{\sigma_1}^2}\} = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\{-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}\}. \tag{24}$$

Solving this equation w.r.t. x you can obtain decision surface, actually a point in this case, x_0. [8]

Excersise 3 *Obtain the decision boundary x_0 when the two classes follow the Gaussian distributions with $N(1, \frac{1}{2})$ and $N(3, \frac{1}{2})$ respectively.*

• General Gaussian case.

The Equation (11) with taking the function *ln* as $f(\cdot)$ in the Equation (20), we obtain our discriminating function g_i as:

$$\begin{aligned} g_i(\mathbf{x}) &= \ln(p(\mathbf{x}|\omega_i)p(\omega_i)) \\ &= \ln p(\mathbf{x}|\omega_i) + \ln p(\omega_i) \\ &= -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln p(\omega_i) + c_i \end{aligned}$$

where

$$c_i = -(l/2)\ln(2\pi) - (1/2)\ln|\Sigma_i| \tag{25}$$

[8] See also the Eqations (9) and (10).

• 2-D Gaussian case.

Our discriminat function (25) for 2-D Gaussian pdf is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln p(\omega_i) + c_i \tag{26}$$

where

$$c_i = -(1/2)\ln(2\pi) - (1/2)\ln|\Sigma_i|. \tag{27}$$

This is simplified

⋆ if

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

then

$$g_i = \frac{(x_1-\mu_1)^2}{-2\sigma_1^2} + \frac{(x_2-\mu_2)^2}{-2\sigma_2^2} - \frac{1}{2}\ln(\sigma_1^2\sigma_2^2) + \ln p(\omega_i) - \frac{1}{2}\ln(2\pi) \tag{28}$$

Note that the last two terms in the right hand side of the above equation will be cancelled when obtaining the border $g_i - g_j = 0$.

- This kind of two distributions construct a *quadratic* decision surface.
  - · *ellipsoid*
  - · *hyperbola*
  - · *parabola*
  - · *pair of lines*
- If expanded to $l > 2$ case, then it is called *hyper-quadratic.*

- If $\sigma_1 = \sigma_2$ holds in both Σ_i and Σ_j then the decision surface is a *circle.*
- Furthermore, if $\Sigma_i = \Sigma_j$ holds then the decision surface is a *straight line.*

Excersise 4 *When* $p(\omega_1) = p(\omega_2)$, $\mu_1 = (0,0)$ *and* $\mu_2 = (1,1)$ *obtain decision surface in the following four cases.*

$$(1) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}$$

$$(2) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.10 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.20 \end{pmatrix}$$

$$(3) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.15 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.25 \end{pmatrix}$$

$$(4) \quad \Sigma_1 = \begin{pmatrix} 0.20 & 0 \\ 0 & 0.10 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.30 & 0 \\ 0 & 0.10 \end{pmatrix}$$

$$(5) \quad \Sigma_1 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.15 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.15 & 0 \\ 0 & 0.10 \end{pmatrix}$$

Excersise 5 *When* $p(\omega_1) = p(\omega_2)$ *again, but* $\mu_1 = \mu_2 = (0,0)$ *this time, obtain decision surface.*

$$(6) \quad \Sigma_1 = \begin{pmatrix} 0.30 & 0 \\ 0 & 0.30 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.10 & 0 \\ 0 & 0.90 \end{pmatrix}$$

Excersise 6 *When* $p(\omega_1) = p(\omega_2)$, *what condition is needed for the border of two classes to be the following decision surface?*

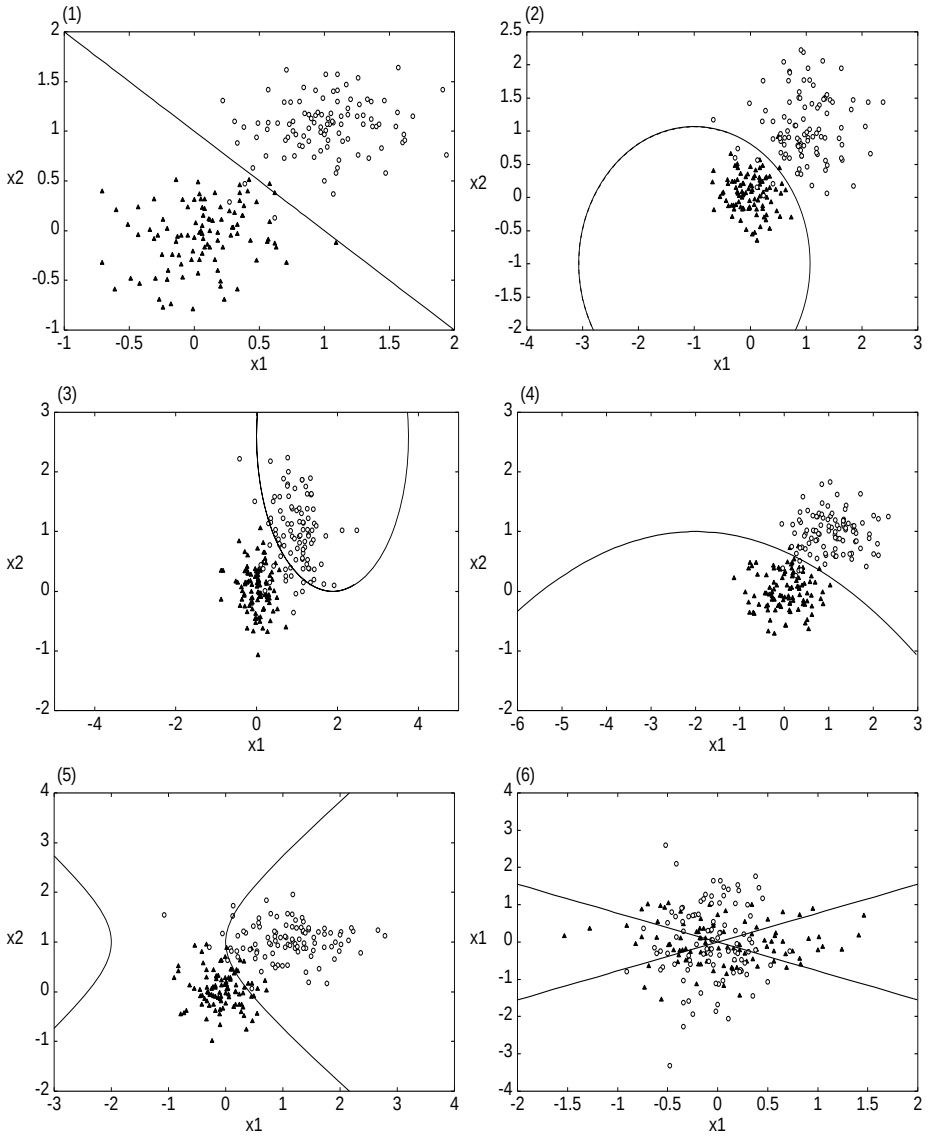
(1) ellipse;

(2) parabola;

(3) hyperbola;

(4) two straight lines.

• Examples of border — from Excersise 4 and 5.



The equations of the above borders are:

(1)
$$x_1 + x_2 = 1.$$

(2)
$$(x_1 + 1)^2 + (x_2 + 1)^2 = 4 - 0.2\ln(1/4).$$

(3)
$$15(x_1 + 1)^2 + 8(x_2 + 3/2)^2 = 93/2 - 6\ln(3/10).$$

(4)
$$x_1^2 + 4x_1 + 12x_2 - 8 + 0.6ln(2/3) = 0.$$

(5)
$$(x_1 + 1)^2 - (x_2 - 3/2)^2 = 3/2.$$

(6)
$$(\sqrt{3}x_1 + \sqrt{5})(\sqrt{3}x_1 - \sqrt{5}) = 0.$$

⋆ An example of more general case:

That is, our Σ is no more of the form

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

but, instead, of the more general form

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}.$$

Then the Equation (26) leads an equation of the form

$$c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2 + c_4 x_1 + c_5 x_5 + c_6 = 0.$$

Excersise 7 *When $p(\omega_1) = p(\omega_2)$, $\mu_1 = (0,0)$ and $\mu_2 = (1,1)$ still holds, obtain decision surface in the following case.*

$$\Sigma_1 = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.2 \end{pmatrix}$$

The border of the Excersise 7 will be

$$18x_1^2 - 67x_1x_2 - 52x_2^2 = (11/10)\ln 11.$$

Excersise 8 *But how can we draw the graph of this equation?*

□ **Classification by distance from mean:**

Recalling that we classify $\mathbf{x}$ to ω_i such that g_i is maximized, the Equation (25) implies we minimize:

$$(\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) \tag{29}$$

that is,

- for diagonal $\Sigma = \sigma^2 I$, minimize

$$d_E \equiv \|\mathbf{x} - \mu_i\| \tag{30}$$

- for no-diagonal Σ, minimize

$$d_M \equiv (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) \tag{31}$$

  which is called *Mahalanobis distance.*

Note that:

- $d_E = const. \Rightarrow$ *hyper-sphere*
- $d_M = const. \Rightarrow$ *hyper-ellipse*

Excersise 9 *About a two-class, two dimensional Bayesian classifier where*

$$p(\omega_1) = p(\omega_2), \quad \mu_1 = (0, 0), \quad \mu_2 = (1, 0)$$

and

$$\Sigma_1 = \Sigma_2 = \Sigma = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$$

answer the following questions.

(1) Obtain d_M *of* $\mathbf{x} = (1.0, 2.2)$ *from* μ_1 *and* μ_2*, respectively.*

(2) Should $\mathbf{x}$ *classified to* ω_1 *or* ω_2*?*
(Compare the result when we use d_E *instead of* d_M*.*

□ **Diagonalization of Variance Matrix:**

— to learn the shape of the ellipses

Definition

- *Eigenvalue* of Σ are λ's that satisfy

$$|\Sigma - \lambda I| = 0 \tag{32}$$

- *Eigenvector* of *Sigma* are $\mathbf{x}$'s that satisfy

$$\Sigma \mathbf{x} = \lambda \mathbf{x} \tag{33}$$

Then we have the following relation:

$$\Sigma = \Phi \Lambda \Phi^T \tag{34}$$

where Λ is the diagonal matrix whose elements are the eigenvalues of Σ, Φ is a matrix whose columns are corresponing eignevectors of Σ, and due to its symmetry $\Phi^T = \Phi^{-1}$.

Hence, our ellipse $d_M = c^2$ becomes:

$$(\mathbf{x} - \mu)^T \Phi \Lambda \Phi^T (\mathbf{x} - \mu) = c^2. \tag{35}$$

This is interpreted as:

$$(\mathbf{X} - \mu)^T \Lambda (\mathbf{X} - \mu) = c^2 \tag{36}$$

on the (rotated) new coordinate:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{X} = \Phi^T \mathbf{x} \tag{37}$$

where the ellipse has the equation of the form:

$$a_1 X_1^2 + a_2 X_2^2 = c$$

instead of the form:

$$a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 x_2 = c$$

Excersise 10 *When*

$$\Sigma = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$$

answer the following questions.

(1) Obtain Eigenvalues of Σ.

(2) Obtain Eigenvectors of Σ.

(3) Solve the equation w.r.t. x_1 *and* x_2.

$$\mathbf{x}^T \Sigma^{-1} \mathbf{x} = 1.$$

(4) Obtain Φ *and* Λ*, and ascertain the relation:*

$$\Sigma = \Phi^{-1} \Lambda \Phi.$$

(5) Obtain the equation of the projected ellipse w.r.t. X_1 *and* X_2

$$\mathbf{x}^T (\Phi \Lambda^{-1} \Phi^T) \mathbf{x} = 1$$

that is,

$$\mathbf{X}^T \Lambda^{-1} \mathbf{X} = 1.$$

• Examples of border — from Excersise 10

Since now our Σ is

$$\Sigma = \begin{pmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{pmatrix}$$

$$\mathbf{x}^T\Sigma\mathbf{x} = (x_1 \;\; x_2)\begin{pmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \cdots = 0.95x_1^2 - 0.30x_1x_2 + 0.55x_2.$$

So our ellipse is, e.g.,

$$0.95x_1^2 - 0.30x_1x_2 + 0.55x_2 = 1.$$

In order to learn the shape of the ellipse we transform it with $\Sigma = \Phi\Lambda\Phi^T$ where

$$\Phi = \frac{1}{\sqrt{10}}\begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}$$

and

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

That is, our ellipse $\mathbf{x}^T\Sigma^{-1}\mathbf{x} = 1$ is expressed as $\mathbf{x}^T\Phi\Lambda^{-1}\Phi^T\mathbf{x} = 1$ where

$$\mathbf{x}^T\Phi = (x_1 \;\; x_2)\begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix}/\sqrt{10} = (3x_1 - x_2 \;\; x_1 + 3x_2)/\sqrt{10}$$

and

$$\Phi^T\mathbf{x} = \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}/\sqrt{10} = (3x_1 - x_2 \;\; x_1 + 3x_2)/\sqrt{10}$$

Hence, if we use the new coordinate:

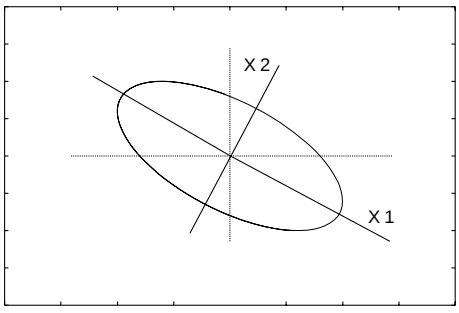
$$(X_1 \;\; X_2) = (3x_1 - x_2 \;\; x_1 + 3x_2)/\sqrt{10}$$

our ellipse is

$$(X_1 \;\; X_2)\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = 1$$

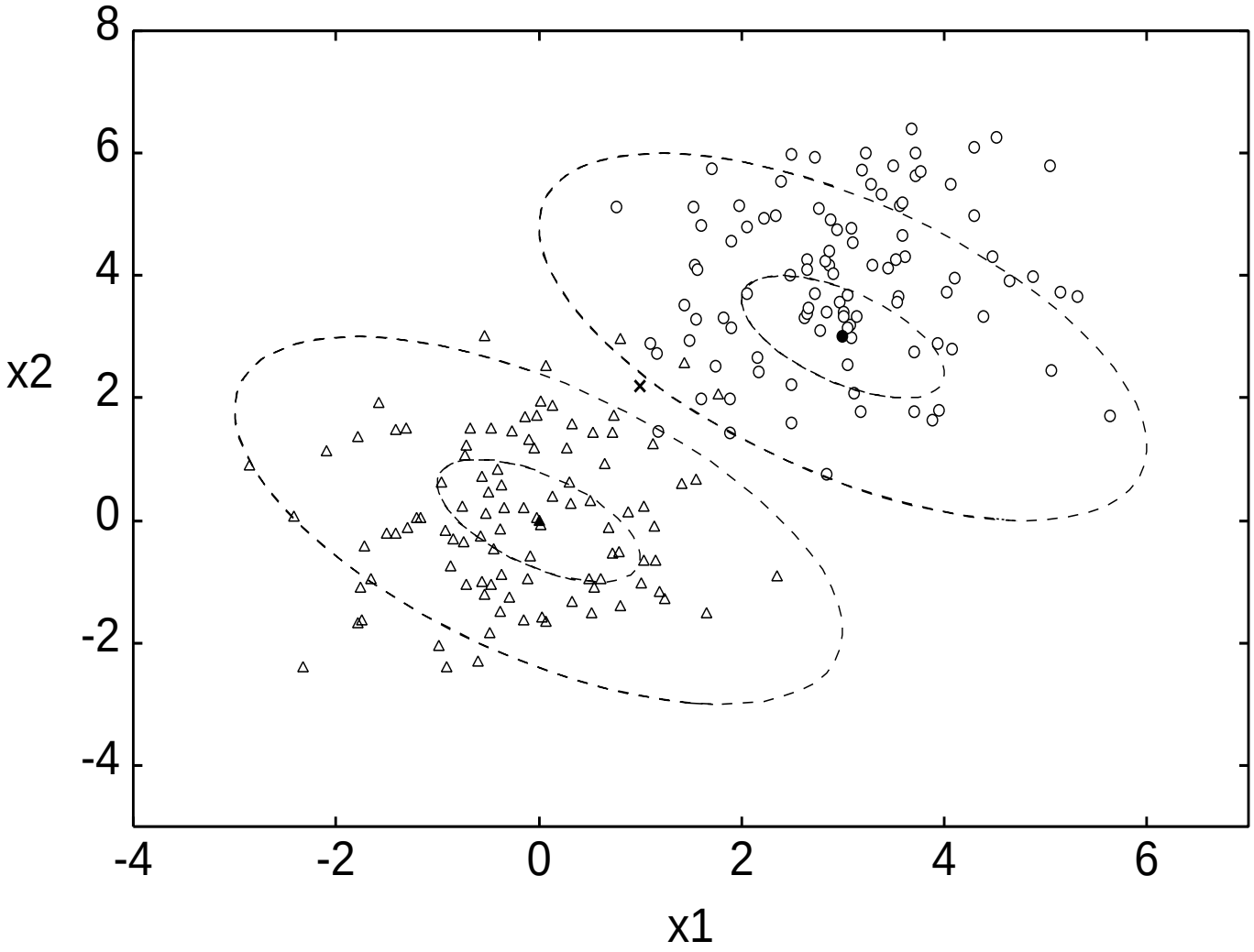
That is

$$X_1^2 + 2X_2^2 = 1$$



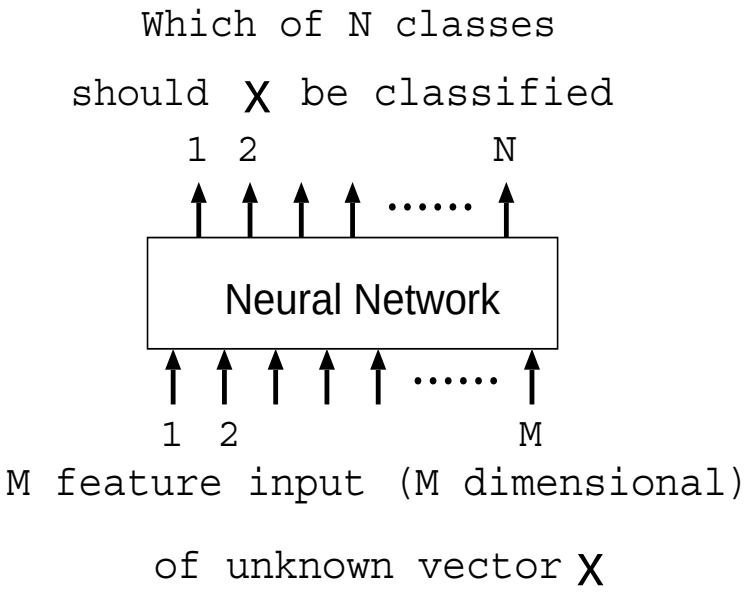
- From Excersise 9 and 10.

  · Points whose Mahalanobis distance is identical from center (mean) of the two classes respectively.

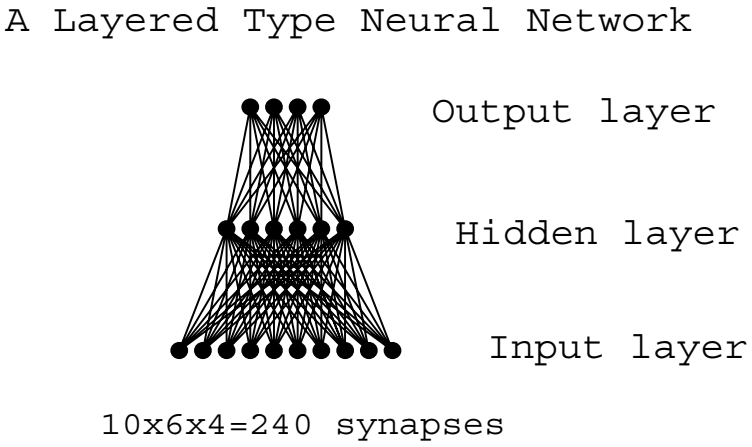


□ Pattern Classification by Neural Network

- We can use *Neural Networks* to classify for any number of features to any number of classes.



- What is *Neural Networks*?
  - Neuron
  - Synaps
  - Synaptic weight and Transfer function
  - Learning



- Definition of variables:
  - L layers of neurons.
  - k_r neurons in the r-th layer. $\quad r = 1, 2, \cdots, L.$
  - M input neurons ($k_1 = M$) and N output neurons ($k_L = N$).
  - w_{ij}^r: weight from the j-th neuron in the r-th layer to the i-th neuron in the r-th layer.
  - p training pairs: $(\mathbf{x}(\mu), \mathbf{z}(\mu)), \quad i = 1, 2, \cdots, p.$
  - $y_i^r(\mu)$: the output of the i-th neuron in the r-th layer when the μ-th training sample is given.
  - fan-in's to j-th neuron in the r-th layer when the μ-th training sample is given.

$$h_j^r(\mu) \equiv \sum_{k=1}^{k_{r-1}} w_{jk}^r y_k^{r-1}(i) + w_{j0}^r = \sum_{k=0}^{k_{r-1}} w_{jk}^r y_k^{r-1}(i), \tag{38}$$

where

$$y_0^{r-1} = 1$$

and

$$(y_1^1(\mu), \cdots, y_M^1(\mu)) = (x_1(\mu), \cdots, x_M(\mu)) = \mathbf{x}(\mu).$$

  - Each output neuron's error:

$$e_i(\mu) \equiv y_i(\mu) - z_i(\mu), \;\; i = 1, 2, \cdots, N. \tag{39}$$

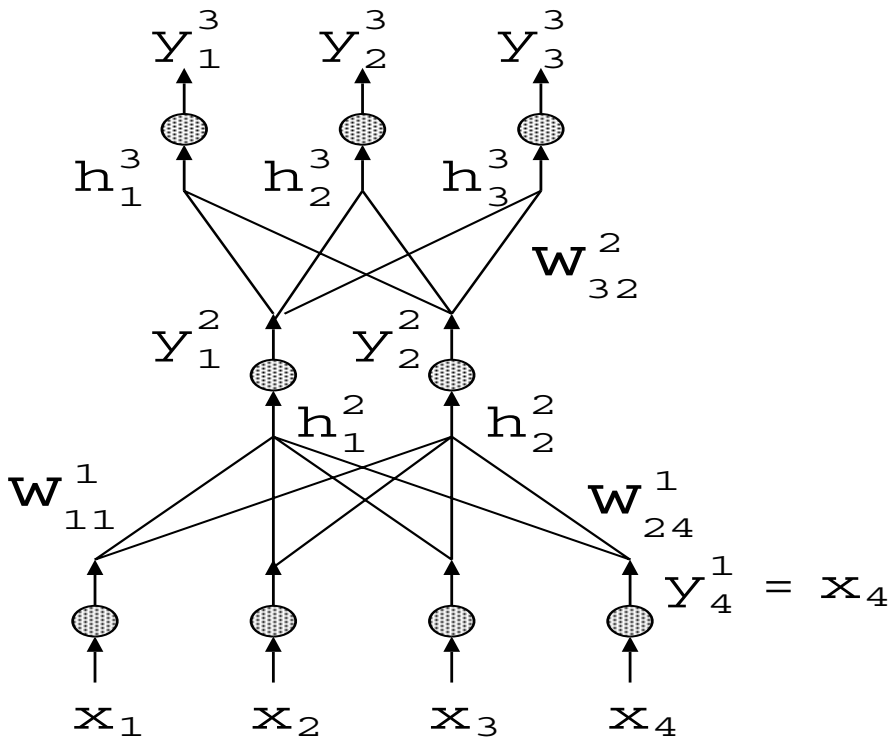
where $(z_1(\mu), \cdots, z_N(\mu)) = \mathbf{z}(\mu)$

  - Error function to be minimized:

$$D = \frac{1}{2} \sum_{\mu=1}^{p} \sum_{i=1}^{N} e_i(\mu)^2 \tag{40}$$

• An example of the notations

- $L = 3, \quad M = k_1 = 4, \quad N = k_3 = 3 \quad$ and $\quad k_2 = 2.$



- e.g.

  ⋆ $h_2^2(\mu) = w_{21}^1 \cdot x_1(\mu) + w_{22}^1 \cdot x_2(\mu) + w_{23}^3 \cdot x_3(\mu) + w_{24}^1 \cdot x_4(\mu).$

  ⋆ $h_2^3(\mu) = w_{21}^2 \cdot y_1^2(\mu) + w_{22}^2 \cdot y_2^2(\mu).$

□ Back Propagation

— Learning for multi-layered analog neural networks.

• The algorithm:

1. *Initialization:*

   · Initialize all the weights with small random values.

2. *Forward computations:*

   · For each of the training vectors

   $$\mathbf{x}(i), \quad \mu = 1, 2, \cdots, p$$

   compute all the

   $$y_j^r(\mu) = f(h_j^r(\mu)) \tag{41}$$

   for $j = 1, 2, \cdots, k_r$ and $r = 1, 2, \cdots, L$.

3. *Backward computations:*

   · For each $\mu = 1, 2, \cdots, p$ and $j = 1, 2, \cdots, N$ compute

   $$\delta_j^L(\mu) = e_j(\mu) f'(h_j^L(\mu)). \tag{42}$$

   · Then compute

   $$\delta_j^{r-1}(\mu) = e_j^{r-1}(\mu) f'(h_j^{r-1}(\mu)) \tag{43}$$

   for $r = L, L-1, \cdots, 2$ and $j = 1, 2, \cdots, k_r$ where

   $$e_j^{r-1}(\mu) = \sum_{k=1}^{k_r} \delta_k^r(\mu) w_{kj}^{r-1}. \tag{44}$$

4. Update the weights for $r = 1, 2, \cdots, L$ and $j = 1, 2, \cdots, k_r$

   $$w_{jk}^r\,[\text{new}] = w_{jk}^r\,[\text{old}] + \triangle w_{jk}^r \tag{45}$$

   where

   $$\triangle w_{jk}^r = -\epsilon \delta_j^r(\mu) y_k^r(\mu). \tag{46}$$

   until no change occurs during the training cycle of $\mu = 1, 2, \cdots, p$.

(⋆ Forward calculation in our previous toy example: From the bottom to top.)

$$y_1^1(\mu) = x_1(\mu), \quad y_2^1(\mu) = x_2(\mu), \quad y_3^1(\mu) = x_3(\mu), \quad y_4^1(\mu) = x_4(\mu).$$

$$\Downarrow$$

$$h_1^2(\mu) = w_{11}^1 \cdot y_1^1 + w_{12}^1 \cdot y_2^1 + w_{13}^1 \cdot y_3^1 + w_{14}^1 \cdot y_4^1 \qquad \text{from (38)}$$

$$\Downarrow$$

$$\text{similarly } h_2^2(\mu).$$

$$\Downarrow$$

$$y_1^2 = f(h_1^2(\mu)) \qquad \text{from (41)}$$

$$\Downarrow$$

$$\text{similarly } y_2^2(\mu).$$

$$\Downarrow$$

$$h_1^3(\mu) = w_{11}^2 \cdot y_1^2 + w_{12}^2 \cdot y_2^2 \qquad \text{from (38) again}$$

$$\Downarrow$$

$$\text{similarly } h_2^3(\mu) \quad \text{and} \quad h_3^3(\mu).$$

$$\Downarrow$$

$$y_1^3 = f(h_1^3(\mu)) \qquad \text{from (41) again}$$

$$\Downarrow$$

$$\text{similarly } y_2^3(\mu) \quad \text{and} \quad y_3^3(\mu).$$

(⋆ Backward calculation in our previous toy example: From the top to bottom.)

$$e_1^{(3)} = y_1^3 - z_1(\mu) \qquad \text{from (39)}$$

$$\Downarrow$$

$$\text{similarly } e_2^{(3)}(\mu) \qquad \text{and} \qquad e_3^{(3)}(\mu).$$

$$\Downarrow$$

$$\delta_1^3(\mu) = e_1^{(3)} \cdot f'(h_1^3(\mu)) \qquad \text{from (42)}$$

$$\Downarrow$$

$$\text{similarly } \delta_2^3(\mu) \qquad \text{and} \qquad \delta_3^3(\mu).$$

$$\Downarrow$$

$$e_1^2(\mu) = \delta_1^3(\mu) \cdot w_{11}^2 + \delta_2^3(\mu) \cdot w_{21}^2 + \delta_3^3(\mu) \cdot w_{31}^2$$

$$\Downarrow$$

$$\text{similarly } e_2^2(\mu) \qquad \text{from (44).}$$

$$\Downarrow$$

$$\delta_1^2(\mu) = e_1^2(\mu) \cdot f'(h_1^2(\mu)) \qquad \text{from (43)}$$

$$\Downarrow$$

$$\text{similarly } \delta_2^2(\mu).$$

$$\Downarrow$$

$$\begin{array}{c} e_1^1(\mu) = \delta_1^2(\mu) \cdot w_{11}^1 + \delta_2^2(\mu) \cdot w_{21}^1 \\ e_2^1(\mu) = \delta_1^2(\mu) \cdot w_{12}^1 + \delta_2^2(\mu) \cdot w_{22}^1 \\ e_3^1(\mu) = \delta_1^2(\mu) \cdot w_{13}^1 + \delta_2^2(\mu) \cdot w_{23}^1 \\ e_4^1(\mu) = \delta_1^2(\mu) \cdot w_{14}^1 + \delta_2^2(\mu) \cdot w_{24}^1 \\ \text{from (44) again} \end{array}$$

$$\Downarrow$$

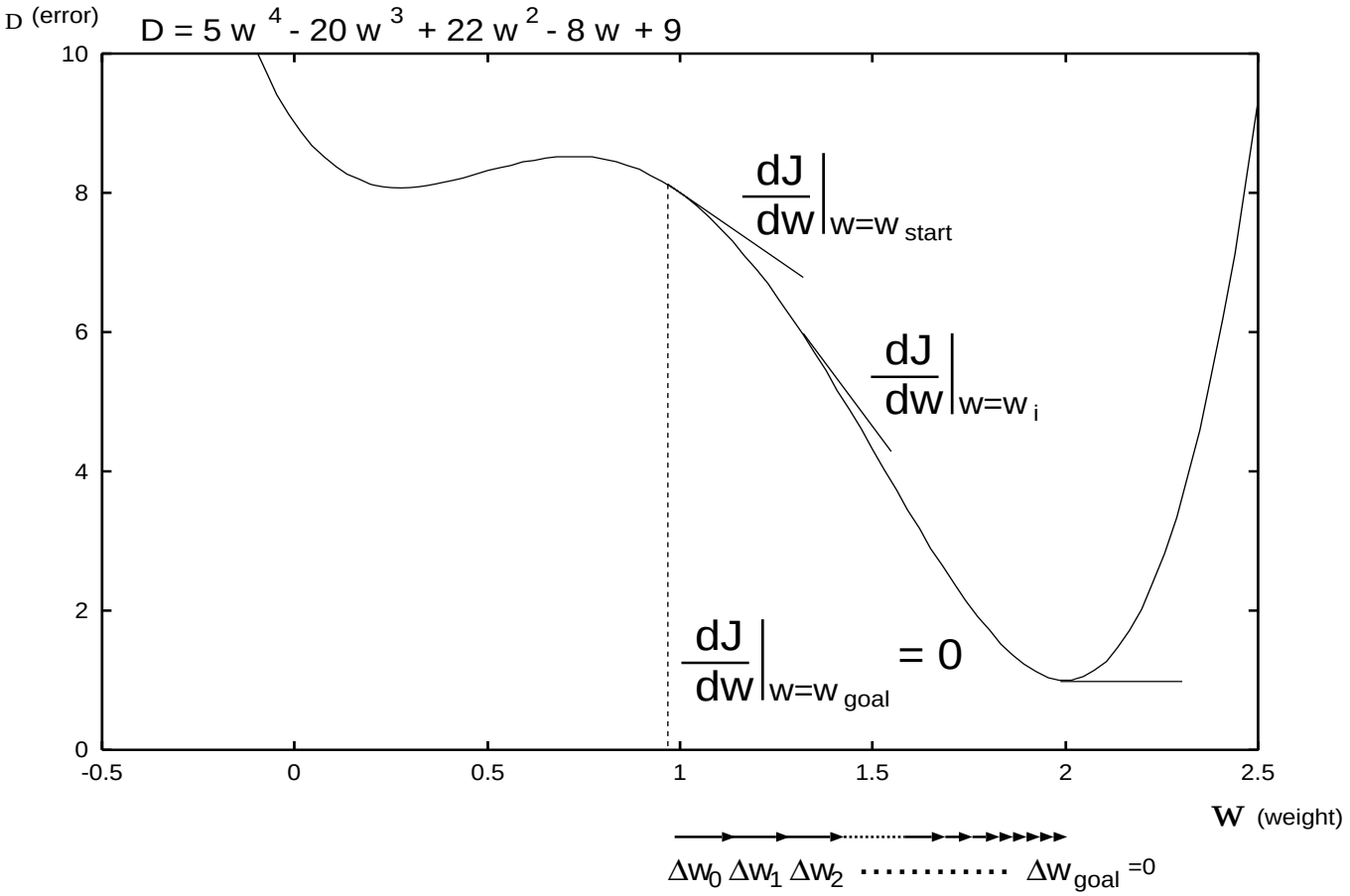
$$\delta_1^1(\mu) = e_1^1(\mu) \cdot f'(h_1^1(\mu)) \qquad \text{from (43) again}$$

$$\Downarrow$$

$$\text{similarly } \delta_2^1(\mu), \quad \delta_3^1(\mu), \quad \delta_4^1(\mu).$$

• Rationale

⋆ Thought Experiment: 1-D weight



Modification:

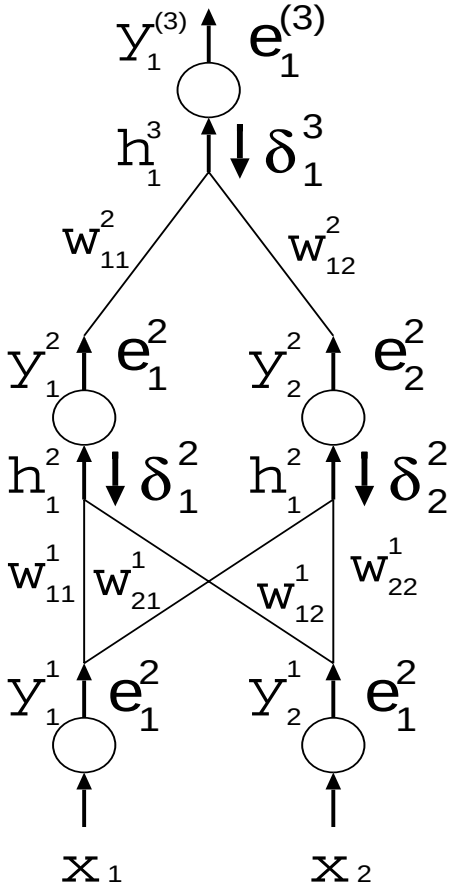
$$w^{\text{new}} = w^{\text{old}} + \triangle w \tag{47}$$

where

$$\triangle w = -\epsilon \cdot \frac{dD}{dw}|_{w=w^{\text{old}}} \tag{48}$$

means that starting from w_0, w gradually approaches the point where D takes a minimum value with δw being decreased.

⋆ An example: XOR (revisited)



In this example,

$$
\begin{aligned}
D &= \frac{1}{2}\{e_1^{(3)}\}^2 \\
&= \frac{1}{2}\{y_i^{(3)} - z_i\}^2 \\
&= \frac{1}{2}\{(f(h_i^3) - z_i\}^2 \\
&= \frac{1}{2}\{f(w_{11}^2 \cdot y_1^2 + w_{12}^2 \cdot y_2^2) - z_1\}^2 \\
&= \frac{1}{2}\{f(w_{11}^2 \cdot f(h_1^2) + w_{12}^2 \cdot f(h_2^2) - z_1\}^2 \\
&= \frac{1}{2}\{f(w_{11}^2 \cdot f(w_{11}^1 \cdot y_1^1 + w_{12}^1 \cdot y_2^1) + w_{12}^2 \cdot f(w_{21}^1 \cdot y_1^1 + w_{22}^1 \cdot y_2^1) - z_1\}^2 \\
&= \frac{1}{2}\{f(w_{11}^2 \cdot f(w_{11}^1 \cdot x_1 + w_{12}^1 \cdot x_2) + w_{12}^2 \cdot f(w_{21}^1 \cdot x_1 + w_{22}^1 \cdot x_2) - z_1\}^2
\end{aligned}
$$

You can see that D is a function of only all the weights (6-D in this case).

⋆ Where are the definitions of δ_j^r and e_j^{r-1} in (43) and (44) from?

We have studied that

$$D = \frac{1}{2}\sum_{i=1}^{N} {e_i}^2 \tag{49}$$

$$= \frac{1}{2}\sum_{i=1}^{N}(y_i - z_i)^2 \tag{50}$$

$$= \frac{1}{2}\sum_{i=1}^{N}(f(h_i^L) - z_i)^2 \tag{51}$$

where

$$h_i^{r+1} = \sum_{k=1}^{k_r} w_{ik}^r \cdot y_k^r \tag{52}$$

So, D is a function of all the w_{ik}^r and

$$\frac{\partial D}{\partial w_{jk}^r} = \frac{\partial D}{\partial h_j^{r+1}} \cdot \frac{\partial h_j^{r+1}}{\partial w_{jk}^r} \tag{53}$$

holds. From (52)

$$\frac{\partial h_j^{r+1}}{\partial w_{jk}^r} = y_k^r \tag{54}$$

and we put

$$\delta_j^{r+1} \equiv \frac{\partial D}{\partial h_j^{r+1}} \tag{55}$$

Since D is a function of h_i^L and

$$h_i^{r+1} = \sum_{k=1}^{k_r} w_{ik}^r f(h_k^r), \tag{56}$$

recursive use of this equation D can be expressed as a function of

$$\sum_{k=1}^{k_r} w_{ik}^r f(h_k^r)$$

for any r. So δ_i^{r-1} is defined as

$$\delta_i^{r-1} \equiv \frac{\partial D}{\partial h_i^{r-1}} = \sum_{i=1}^{k_r} \frac{\partial D}{\partial h_k^{r-1}} \cdot \frac{\partial h_k^r}{\partial h_k^{r-1}} = \sum_{i=1}^{k_r} \delta_k^r \cdot \frac{\partial h_k^r}{\partial h_k^{r-1}} \tag{57}$$

Here,

$$\frac{\partial h_k^r}{\partial h_k^{r-1}} = \frac{\partial\left(\sum_{m=1}^{k_{r-1}} w_{km}^r \cdot f(h_m^{r-1})\right)}{\partial h_k^{r-1}} = w_{kj}^r \cdot f'(h_j^{r-1}) \tag{58}$$

Hence,

$$\delta_i^{r-1} = \sum_{k=1}^{k_r} \delta_k^r \cdot w_{kj}^r \cdot f'(h_j^{r-1}) \tag{59}$$

In conclusion, if we define

$$e_j^{r-1} = \sum_{k=1}^{k_r} \delta_k^r \cdot w_{kj}^r \tag{60}$$

we obtain

$$\delta_j^{r-1} = e_j^{r-1} \cdot f'(h_j^{r-1}) \tag{61}$$

□ Linear Transformation

— to reduce number of features.

• Discrete Fourier Transform (DFT)

$$y(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \exp(-j\frac{2\pi}{N}kn) \tag{62}$$

· Inverse DFT

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} y(k) \exp(j\frac{2\pi}{N}kn) \tag{63}$$

- DFT transforms $x(0), x(1), \cdots, x(N-1)$ to $y(0), y(1), \cdots, y(N-1)$.
- For example, we can send transformed data $\mathbf{y}$ instead of real data $\mathbf{x}$.
- The number of data is not reduced directly, but usually most of the energy lies in the low-frequency region.
- Matrix form of the transformation e.g. $N = 4$

$$\frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{pmatrix}$$

• Kalhunen-Loeve Transform (KL)

K-L Transform is calculated from the correlation matrix R of the data, so that the n-th raw is the eigen vector of R corresponding to the n-th largest eigen value.

Excersise 11 *Obtain the K-L Transform for the correlation matrix:*

$$R = \begin{pmatrix} 0.3 & 0.1 & 0.1 \\ 0.1 & 0.3 & -0.1 \\ 0.1 & -0.1 & 0.3 \end{pmatrix}$$

- **Hadamar Transform:**

  · For a 2^n-dimensional vector $\mathbf{x}$ (2^n feature vector), the transform and its inverse transform are:

$$\mathbf{y} = H_n\mathbf{x}, \qquad \mathbf{x} = H_n\mathbf{y} \tag{64}$$

  where

$$H_n = H_1 \otimes H_{n-1} \tag{65}$$

  and

$$H_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{66}$$

  Here $\otimes$ denotes the Kronecker product for two matrices:[9]

$$[A \otimes B]_{ij} = A_{ij}B \tag{70}$$

- **Haar Transform:**

  · For a 2^n-dimensional vector $\mathbf{x}$:

  (1) Define Haar function $h_k(z)$, which is continuous and are defined in $[0, 1]$ as follows:

  - Decompose k into two integers p and q, such that

$$k = 2^p + q - 1, \qquad k = 0, 1, \cdots, L-1 \tag{71}$$

  which is unique when

$$\begin{cases} q = 0 \quad \text{or} \quad 1 & \text{if} \quad p = 0 \\ 0 \le p \le n-1, \quad 0 < q \le 2^p & \text{if} \quad p \ne 0 \end{cases} \tag{72}$$

[9] for example, if

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \tag{67}$$

and

$$B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \tag{68}$$

then we define

$$A \otimes B = \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{pmatrix} \tag{69}$$

- then Haar functions are

$$h_0(z) \equiv h_{00}(z) = \frac{1}{\sqrt{L}} \tag{73}$$

and

$$h_k(z) \equiv h_{pq}(z) = \frac{1}{\sqrt{L}} \begin{cases} 2^{\frac{p}{2}} & \text{if } \frac{q-1}{2^p} \le z < \frac{q-\frac{1}{2}}{2^p} \\ -2^{\frac{p}{2}} & \text{if } \frac{q-\frac{1}{2}}{2^p} \le z < \frac{q}{2^p} \\ 0 & \text{otherwise in } [0,1] \end{cases} \tag{74}$$

(2) The k-th raw of the Haar transform matrix of order $L = 2^n$ is:

$$z = \frac{m}{L}, \qquad m = 0, 1, 2, \cdots L-1.$$

Excersise 12 *Obtain the 8 × 8 transform matrix.*

· An example of $n = 3$, that is, $L = 8$.

$$k = 2^p + q - 1, \qquad k = 0, 1, \cdots, 7 \tag{75}$$

where

$$\begin{cases} q = 0 \quad \text{or} \quad 1 & \text{if} \quad p = 0 \\ 0 \le p \le 2, \quad 0 < q \le 2^p & \text{if} \quad p \ne 0. \end{cases} \tag{76}$$

So,

p=0	q=0	k=0
	q=1	k=1
	$(0 < q \le 2)$	$(k = 1 + q)$
p=1	q=1	k=2
	q=2	k=3
	$(0 < q \le 4)$	$(k = 3 + q)$
p=2	q=1	k=4
	q=2	k=5
	q=3	k=6
	q=4	k=7

Hence,

k	0	1	2	3	4	5	6	7
p	0	0	1	1	2	2	2	2
k	0	1	1	2	1	2	3	4

Then the Haar functions are:

$$h_1(z) \equiv h_{01}(z) = \frac{1}{\sqrt{8}} \begin{cases} 1 & \text{if} \quad -1 \le z < \frac{1}{2} \\ -1 & \text{if} \quad \frac{1}{2} \le z < 1 \end{cases} \tag{77}$$

$$h_2(z) \equiv h_{11}(z) = \frac{1}{\sqrt{8}} \begin{cases} \sqrt{2} & \text{if} \quad 0 \le z < \frac{1}{4} \\ -\sqrt{2} & \text{if} \quad \frac{1}{4} \le z < \frac{1}{2} \\ 0 & \text{otherwise in } [0,1] \end{cases} \tag{78}$$

$$h_3(z) \equiv h_{12}(z) = \frac{1}{\sqrt{8}} \begin{cases} \sqrt{2} & \text{if} \quad \frac{1}{2} \le z < \frac{3}{4} \\ -\sqrt{2} & \text{if} \quad \frac{3}{4} \le z < 1 \\ 0 & \text{otherwise in } [0,1] \end{cases} \tag{79}$$

. . . . . .

$$h_7(z) \equiv h_{24}(z) = \frac{1}{\sqrt{8}} \begin{cases} 2 & \text{if} \quad \frac{3}{4} \le z < \frac{7}{8} \\ -2 & \text{if} \quad \frac{7}{8} \le z < 1 \\ 0 & \text{otherwise in } [0,1] \end{cases} \tag{80}$$

and

$h_0(z)$	1	1	1	1	1	1	1	1
$h_1(z)$	1	1	1	1	–1	–1	–1	–1
$h_2(z)$	$\sqrt{2}$	$\sqrt{2}$	$-\sqrt{2}$	$-\sqrt{2}$	0	0	0	0
$h_3(z)$	0	0	0	0	$\sqrt{2}$	$\sqrt{2}$	$-\sqrt{2}$	$-\sqrt{2}$
$h_4(z)$	2	–2	0	0	0	0	0	0
$h_5(z)$	0	0	2	–2	0	0	0	0
$h_6(z)$	0	0	0	0	2	–2	0	0
$h_7(z)$	0	0	0	0	0	0	2	–2

Since

$$H_{ij} = \frac{1}{\sqrt{L}} h_j(\frac{j}{L}), \tag{81}$$

we obtain

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{pmatrix} \tag{82}$$