CHAPTER 10
Information-Theoretic Learning Models

Problem 10.1

The maximum entropy distribution of the random variable X is a uniform distribution over the range, [a, b], as shown by

$$f_X(x) = \begin{cases} \dfrac{1}{a-b}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

Hence,

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx \\ &= \int_b^a \frac{1}{a-b} \log(a-b) dx \\ &= \log(a-b) \end{aligned}$$

Problem 10.3

Let

$$\begin{aligned} Y_i &= \mathbf{a}_i^T \mathbf{X}_1 \\ Z_i &= \mathbf{b}_i^T \mathbf{X}_2 \end{aligned}$$

where the vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ have multivariate Gaussian distributions. The correlation coefficient between Y_i and Z_i is defined by

$$\begin{aligned} \rho_i &= \frac{\mathbf{E}[Y_i Z_i]}{\sqrt{\mathbf{E}[Y_i^2]\mathbf{E}[Z_i^2]}} \\ &= \frac{\mathbf{a}_i^T \mathbf{E}[\mathbf{X}_1 \mathbf{X}_2^T] \mathbf{b}_i}{\{(\mathbf{a}_i^T \mathbf{E}[\mathbf{X}_1 \mathbf{X}_1^T] \mathbf{a}_i)(\mathbf{b}_i^T \mathbf{E}[\mathbf{X}_1 \mathbf{X}_2^T] \mathbf{b}_i)\}^{1/2}} \\ &= \frac{\mathbf{a}_i^T \Sigma_{12} \mathbf{b}_i}{\{(\mathbf{a}_i^T \Sigma_{11} \mathbf{a}_i)(\mathbf{b}_i^T \Sigma_{22} \mathbf{b}_i)\}^{1/2}} \end{aligned} \tag{1}$$

where

$$\Sigma_{11} = \mathbf{E}[\mathbf{X}_1\mathbf{X}_1^T]$$
$$\Sigma_{12} = \mathbf{E}[\mathbf{X}_1\mathbf{X}_2^T] = \Sigma_{21}$$
$$\Sigma_{22} = \mathbf{E}[\mathbf{X}_2\mathbf{X}_2^T]$$

The mutual information between Y_i and Z_i is defined by

$$I(Y_i;Z_i) = -\log(1-\rho_i^2)$$

Let r denote the rank of the cross-covariance matrix Σ_{12}. Given the vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, we may invoke the idea of canonical correlations as summarized here:

- Find the pair of random variables $Y_1 = \mathbf{a}_1^T\mathbf{X}_1$ and $Z_1 = \mathbf{b}_1^T\mathbf{X}_2$ that are most highly correlated.
- Extract the pair of random variables $Y_2 = \mathbf{a}_2^T\mathbf{X}_1$ and $Z_2 = \mathbf{b}_2^T\mathbf{X}_2$ in such a way that Y_1 and Y_2 are uncorrelated and so are Z_1 and Z_2.
- Continue these two steps until at most r pairs of variables $\{(Y_1, Z_i), (Y_2, Z_i), ..., (Z_r, Z_r)\}$ have been extracted.

The essence of the canonical correlation described above is to encapsulate the dependence between random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ in the sequence $\{(Y_1, Z_i), (Y_2, Z_i), ..., (Z_r, Z_r)\}$. The uncorrelatedness of the pairs in this serquence, that is,

$$\mathbf{E}[Y_iY_j] = \mathbf{E}[Z_iZ_j] = 0 \qquad \text{for all } j \neq i$$

means that the mutual information between the vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ is the sum of the mutual information measures between the individual elements of the pairs $\{(Y_i, Z_i)\}_{i=1}^r$. That is, we may write

$$I(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{r} I(Y_{ij},Z_i) + \text{constant}$$
$$= -\sum_{i=1}^{r} \log(1-\rho_i^2) + \text{constant}$$

where ρ_i is defined by (1).

Problem 10.4

Consider a multilayer perceptron with a single hidden layer. Let w_{ji} denote the synaptic weight of hidden neuron j connected to source node i in the input layer. Let $x_{i|\alpha}$ denote the ith component of the input vector $\mathbf{x}$, given example α. Then the induced local field of neuron j is

$$v_{j|\alpha} = \sum_i w_{ji} x_{i|\alpha} \tag{1}$$

Correspondingly, the output of hidden neuron j for example α is given by

$$y_{j|\alpha} = \varphi(v_{j|\alpha}) \tag{2}$$

where $\varphi(\cdot)$ is the logistic function

$$\varphi(v) = \frac{1}{1 + e^{-v}}$$

Consider next the output layer of the network. Let w_{kj} denote the synaptic weight of output neuron k connected to hidden neuron j. The induced local field of output neuron k is

$$v_{k|\alpha} = \sum_i w_{kj} y_{j|\alpha} \tag{3}$$

The kth output of the network is therefore

$$y_{k|\alpha} = \varphi(v_{k|\alpha}) \tag{4}$$

The output $y_{k|\alpha}$ is assigned a probabilistic interpretation by writing

$$p_{k|\alpha} = y_{k|\alpha} \tag{5}$$

Accordingly, we may view $y_{k|\alpha}$ as an estimate of the conditional probability that the proposition k is true, given the example α at the input. On this basis, we may interpret

$$1 - y_{k|\alpha} = 1 - p_{k|\alpha}$$

as the estimate of the conditional probability that the proposition k is false, given the input example α. Correspondingly, let $q_{k|\alpha}$ denote the actual (true) value of the conditional probability that the proposition k is true, given the input example α. This means that 1 - $q_{k|\alpha}$ is the actual

value of the conditional probability that the proposition k is false, given the input example α. Thus, we may define the Kullback-Leibler divergence for the multilayer perceptron as

$$D_{p||q} = \sum_{\alpha} p_{\alpha} \sum_{k} \left[q_{k|\alpha} \log\left(\frac{q_{k|\alpha}}{p_{k|\alpha}}\right) + (1 - q_{k|\alpha}) \log\left(\frac{1 - q_{k|\alpha}}{1 - p_{k|\alpha}}\right) \right]$$

where p_{α} is the a priori probability of occurrence of example α at the input.

To perform supervised training of the multilayer perceptron, we use gradient descent on $D_{p||q}$ in weight space. First, we use the chain rule to express the partial derivative of $D_{p||q}$ with respect to the synaptic weight w_{kj} of output neuron k as follows:

$$\begin{aligned} \frac{\partial D_{p||q}}{\partial w_{kj}} &= \frac{\partial D_{p||q}}{\partial p_{k|\alpha}} \frac{\partial p_{k|\alpha}}{\partial y_{k|\alpha}} \frac{\partial y_{k|\alpha}}{\partial v_{k|\alpha}} \frac{\partial v_{k|\alpha}}{\partial w_{kj}} \\ &= -\sum_{\alpha} p_{\alpha} (q_{k|\alpha} - p_{k|\alpha}) y_{j|\alpha} \end{aligned} \tag{6}$$

Next, we express the partial derivative of $D_{p||q}$ with respect to the synaptic weight w_{ji} of hidden neuron j by writing

$$\frac{\partial D_{p||q}}{\partial w_{ji}} = -\sum_{\alpha} p_{\alpha} \sum_{\alpha} \left(\frac{q_{k|\alpha}}{p_{k|\alpha}} - \frac{1 - q_{k|\alpha}}{1 - p_{k|\alpha}} \right) \frac{\partial p_{k|\alpha}}{\partial w_{ji}} \tag{7}$$

Via the chain rule, we write

$$\begin{aligned} \frac{\partial p_{k|\alpha}}{\partial w_{ji}} &= \frac{\partial p_{k|\alpha}}{\partial y_{k|\alpha}} \frac{\partial y_{k|\alpha}}{\partial v_{k|\alpha}} \frac{\partial v_{k|\alpha}}{\partial y_{j|\alpha}} \frac{\partial y_{j|\alpha}}{\partial v_{j|\alpha}} \frac{\partial v_{j|\alpha}}{\partial w_{ji}} \\ &= \varphi'(v_{k|\alpha}) w_{kj} \varphi'(v_{j|\alpha}) x_{i|\alpha} \end{aligned} \tag{8}$$

But

$$\begin{aligned} \varphi'(v_{k|\alpha}) &= y_{k|\alpha}(1 - y_{k|\alpha}) \\ &= p_{k|\alpha}(1 - p_{k|\alpha}) \end{aligned} \tag{9}$$

Hence, using (8) and (9) we may simplify (7) as

$$\frac{\partial D_{p||q}}{\partial w_{ji}} = -\sum_{\alpha} p_{\alpha} x_{i|\alpha} \varphi'\left(\sum_{i} w_{ji} x_{i|\alpha} \right) \sum_{k} (p_{k|\alpha} - q_{k|\alpha}) w_{kj}$$

where $\varphi'(\cdot)$ is the derivative of the logistic function $\varphi(\cdot)$ with respect to its argument.

Assuming the use of the learning-rate parameter η for all weight changes applied to the network, we may use the method of steepest descent to write the following two-step probabilistic algorithm:

1. For output neuron k, compute

$$\Delta w_{kj} = -\eta \frac{\partial D_{p||q}}{\partial w_{kj}}$$

$$= \eta \sum_{\alpha} p_{\alpha}(q_{k|\alpha} - P_{k|\alpha}) y_{j|\alpha}$$

2. For hidden neuron j, compute

$$\Delta w_{ji} = -\eta \frac{\partial D_{p||q}}{\partial w_{ji}}$$

$$= \eta \sum_{\alpha} p_{\alpha} x_{i|\alpha} \varphi'\left(\sum_{i} w_{ji} x_{i|\alpha}\right) \sum_{k} (p_{k|\alpha} - q_{k|\alpha}) w_{kj}$$

Problem 10.9

We first note that the mutual information between the random variables X and Y is defined by

$$I(X;Y) = h(X) + h(Y) - h(X, Y)$$

To maximize the mutual information $I(X;Y)$ we need to maximize the sum of the differential entropy $h(X)$ and the differential entropy $h(\hat{Y})$ and also minimize the joint differential entropy $h(X,Y)$. From the definition of differential entropy, both $h(X)$ and $h(Y)$ attain their maximum value of 0.5 when X and Y occur with probability 1/2. Moreover $h(X,Y)$ is minimized when the joint probability of X and Y occupies the smallest possible region in the probability space.

Problem 10.10

The outputs Y_1 and Y_2 of the two neurons in Fig. P10.6 in the text are respectively defined by

$$Y_1 = \left(\sum_{i=1}^{m} w_{1i} x_i\right) + N_1$$

$$Y_2 = \left(\sum_{i=1}^{L} w_{2i} x_i\right) + N_2$$

where are w_{1i} the synaptic weights of output neuron 1, and the w_{2i} are synaptic weights of output neuron 2. The mutual information between the output vector $\mathbf{Y} = [Y_1, Y_2]^T$ and the input vector $\mathbf{X} = [X_1, X_2, .., X_m]^T$ is

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}) &= h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{X}) \\ &= h(\mathbf{Y}) - h(\mathbf{N}) \end{aligned} \tag{1}$$

where $h(\mathbf{Y})$ is the differential entropy of the output vector $\mathbf{Y}$ and $h(\mathbf{N})$ is the differential entropy of the noise vector $\mathbf{N} = [\mathbf{N}_1, \mathbf{N}_2]^T$.

Since the noise terms $\mathbf{N}_1$ and $\mathbf{N}_2$ are Gaussian and uncorrelated, it follows that they are statistically independent. Hence,

$$\begin{aligned} h(\mathbf{N}) &= h(\mathbf{N}_1, \mathbf{N}_2) \\ &= h(\mathbf{N}_1) + h(\mathbf{N}_2) \\ &= 1 + \log(2\pi\sigma_N^2) \end{aligned} \tag{2}$$

The differential entropy of the output vector $\mathbf{Y}$ is

$$\begin{aligned} h(\mathbf{Y}) &= h(Y_1, Y_2) \\ &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\log f_{Y_1,Y_2}(y_1, y_2)dy_1 dy_2 \end{aligned}$$

where $f_{Y_1,Y_2}(y_1, y_2)$ is the joint pdf of Y_1 and Y_2. Both Y_1 and Y_2 are dependent on the same set of input signals, and so they are correlated with each other. Let

$$\begin{aligned} \mathbf{R} &= \mathbf{E}[\mathbf{Y}\mathbf{Y}^T] \\ &= \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \end{aligned}$$

where

$$r_{ij} = \mathbf{E}[Y_i Y_j], \qquad i, j = 1, 2$$

The individual element of the correlation matrix $\mathbf{R}$ are given by

$$\begin{aligned} r_{11} &= \sigma_1^2 + \sigma_N^2 \\ r_{12} &= r_{21} = \sigma_1\sigma_1\rho_{12} \end{aligned}$$

$$r_{22} = \sigma_2^2 + \sigma_N^2$$

where σ_1^2 and σ_2^2 are the respective variances of Y_1 and Y_2 in the absence of noise, and ρ_{12} is their correlation coefficient also in the absence of noise. For the general case of an N-dimensional Gaussian distribution, we have

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{N/2}(\det \mathbf{R})^{1/2}} \exp\left(\frac{1}{2}\mathbf{y}^T \mathbf{R}^{-1}\mathbf{y}\right)$$

Correspondingly, the differential entropy of the N-dimensional vector $\mathbf{Y}$ is described as

$$h(\mathbf{Y}) = \log((2\pi e)^{N/2}\det(\mathbf{R}))$$

where e is the base of the natural logarithm. For the problem at hand, we have $N = 2$ and so

$$\begin{aligned} h(\mathbf{Y}) &= \log(2\pi e \det(\mathbf{R})) \\ &= 1 + \log(2\pi \det(\mathbf{R})) \end{aligned}$$

Hence, the use of (2) and (3) in (1) yields

$$I(\mathbf{X};\mathbf{Y}) = \log\left(\frac{\det(\mathbf{R})}{\sigma_N^2}\right) \tag{4}$$

For a fixed noise variance σ_N^2, the mutual information I($\mathbf{X}$;$\mathbf{Y}$) is maximized by maximizing the determinant det($\mathbf{R}$). By definition,

$$\det(\mathbf{R}) = r_{11}r_{22} - r_{12}r_{21}$$

That is,

$$\det(\mathbf{R}) = \sigma_N^4 + \sigma_N^2(\sigma_1^2 + \sigma_2^2) + \sigma_1^2\sigma_2^2(1 - \rho_{12}^2) \tag{5}$$

Depending on the value of noise variance σ_N^2, we may identify two distinct situations:

1. **Large noise variance**. When σ_N^2 is large, the third term in (5) may be neglected, obtaining
   $$\det(\mathbf{R}) \approx \sigma_N^4 + \sigma_N^2(\sigma_1^2 + \sigma_2^2)$$

In this case, maximizing det(**R**) requires that we maximize $(\sigma_1^2 + \sigma_2^2)$. This requirement may be satisfied simply by maximizing the variance σ_1^2 of output Y_1 or the variance σ_2^2 of output Y_2, separately. Since the variance of output Y_i : i = 1, 2, is equal to σ_i^2 in the absence of noise and $\sigma_1^2 + \sigma_N^2$ in the presence of noise, it follows from the Infomax principle that the optimum solution for a fixed noise variance is to maximize the variance of either output, Y_1 or Y_2.

2. **Low noise variance**. When the noise variance σ_N^2 is small, the third term $\sigma_1^2\sigma_2^2(1 - \rho_{12}^2)$ in (5) becomes important relative to the other two terms. The mutual information $I(\mathbf{X};\mathbf{Y})$ is then maximized by making an optimal tradeoff between two options: keeping the output variances σ_1^2 and σ_2^2 large, and making the outputs Y_1 and Y_2 of the two neurons uncorrelated.

Based on these observations, we may now make the following two statements:

- A high-noise level favors redundancy of response, in which case the two output neurons compute the same linear combination of inputs. Only one such combination yields a response with maximum variance.
- A low-noise level favors diversity of response, in which case the two output neurons compute different linear combinations of inputs even though such a choice may result in a reduced output variance.

Problem 10.11

(a) We are given

$$Y_a = S + N_a$$
$$Y_b = S + N_b$$

Hence,

$$\frac{Y_a + Y_b}{2} = S + \frac{1}{2}(N_a + N_b)$$

The mutual information between $\frac{1}{2}(Y_a + Y_b)$ and the signal component S is

$$I\left(\frac{Y_a + Y_b}{2}; S\right) = h\left(\frac{Y_a + Y_b}{2}\right) - h\left(\frac{Y_a + Y_b}{2}\middle| S\right) \tag{1}$$

The differential entropy of $\frac{Y_a + Y_b}{2}$ is

$$h\left(\frac{Y_a + Y_b}{2}\right) = \frac{1}{2}\left[1 + \log\left(\frac{\pi}{2}\text{var}[Y_a + Y_b]\right)\right] \tag{2}$$

The conditional differential entropy of $\frac{Y_a + Y_b}{2}$ given S is

$$\begin{aligned} h\left(\frac{Y_a + Y_b}{2}\middle| S\right) &= h\left(\frac{N_a + N_b}{2}\right) \\ &= \frac{1}{2}\left[\log\left(\frac{\pi}{2}\text{var}[N_a + N_b]\right)\right] \end{aligned} \tag{3}$$

Hence, the use of (2) and (3) in (1) yields (after the simplification of terms)

$$I\left(\frac{Y_a + Y_b}{2}; S\right) = \log\left(\frac{\text{var}[Y_a + Y_b]}{\text{var}[N_a + N_b]}\right)$$

(b) The signal component S is ordinarily independent of the noise components N_a and N_b. Hence with

$$Y_a + Y_b = 2S + N_a + N_b$$

it follows that

$$\text{var}[Y_a + Y_b] = 4\text{var}[S] + \text{var}[N_a + N_b]$$

The ratio $(\text{var}[Y_a + Y_b])/(\text{var}[N_a + N_b])$ in the expression for the mutual information $I\left(\frac{Y_a + Y_b}{2}; S\right)$ may therefore be interpreted as a signal-plus-noise to noise ratio.

Problem 10.12

Principal components analysis (PCA) and independent-components analysis (ICA) share a common feature: They both linearly transform an input signal into a fixed set of components.

However, they differ from each other in two important respects:

1. PCA performs decorrelation by minimizing second-order moments; higher-order moments are not involved in this computation. On the other hand, ICA performs statistical independence by using higher-order moments.

2. The output signal vector resulting from PCA has a diagonal covariance matrix. The first principal component defines a direction in the original signal space that captures the maximum possible variance; the second principal component defines another direction in the remaining orthogonal subspace that captures the next maximum possible variance, and so on. On the other hand, ICA does not find the directions of maximum variances but rather interesting directions where the term "interesting" refers to "deviation from Gaussianity".

Problem 10.13

Independent components analysis may be used as a preprocessing tool before signal detection and pattern classification. In particular, through a change of coordinates resulting from the use of ICA, the probability density function of multichannel data may be expressed as a product of marginal densities. This change, in turn, permits density estimation with shorter observations.

Problem 10.14

Consider m random variables $X_1, X_2, ..., X_m$ that are defined by

$$X_i = \sum_{j=1}^{N} a_{ij} U_j, \qquad i = 1, 2, ..., N$$

where the U_j are independent random variables. The Darmois' theorem states that if the X_i are independent, then the variables U_j for which $a_{ij} \neq 0$ are all Gaussian.

For independent-components analysis to work, at most a single X_i can be Gaussian. If all the X_i are independent to begin with, there is no need for the application of independent-components analysis. This, in turn, means that all the X_i must be Gaussian. For a finite N, this condition can only be satisfied if all the U_j are not only independent but also Gaussian.

Problem 10.15

The use of independent-components analysis results in a set of components that are as statistically independent of each other as possible. In contrast, the use of decorrelation only addresses second-order statistics and there is therefore no guarantee of statistical independence.

Problem 10.16

The Kullback-Leibler divergence between the joint pdf $f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})$ and the factorial pdf $\tilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})$ is the multifold integral

$$D_{f_Y || \tilde{f}_Y} = \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w}) \left(\log \frac{f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})}{\tilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})} \right) d\mathbf{y} \tag{1}$$

Let

$$d\mathbf{y} = dy_i dy_j d\mathbf{y}'$$

where $\mathbf{y}'$ excludes y_i and y_j. We may then rewrite (1) as

$$\begin{aligned}
D_{f_Y||\tilde{f}_Y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dy_i dy_j \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w}) \log f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w}) d\mathbf{y}' \\
&\quad - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dy_i dy_j \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w}) \log \tilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{w}) d\mathbf{y}' \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_i,Y_j}(y_i, y_j, \mathbf{w}) \log f_{Y_i,Y_j}(y_i, y_j, \mathbf{w}) dy_i dy_j \\
&\quad - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_i,Y_j}(y_i, y_j, \mathbf{w}) \log(f_{Y_i}(y_i, \mathbf{w}) f_{Y_j}(y_j, \mathbf{w})) dy_i dy_j \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_i,Y_j}(y_i, y_j) \log\left(\frac{f_{Y_i,Y_j}(y_i, y_j, \mathbf{w})}{f_{Y_i}(y_i, \mathbf{w}) f_{Y_j}(y_j, \mathbf{w})}\right) dy_i dy_j \\
&= I(Y_i; Y_j)
\end{aligned}$$

That is, the Kullback-Leibler divergence between the joint pdf $f_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})$ and the factorial pdf distribution $\tilde{f}_{\mathbf{Y}}(\mathbf{y}, \mathbf{w})$ is equal to the mutual information between the components Y_i and Y_j of the output vector $\mathbf{Y}$ for any pair (i, j).

Problem 10.18

Define the output matrix

$$\mathbf{Y} = \begin{bmatrix} y_1(0) & y_1(1) & \cdots & y_1(N-1) \\ y_2(0) & y_2(1) & \cdots & y_2(N-1) \\ \vdots & \vdots & & \vdots \\ y_m(0) & y_m(1) & \cdots & y_m(N-1) \end{bmatrix} \tag{1}$$

where m is the dimension of the output vector $\mathbf{y}(n)$ and N is the number of samples used in computing the matrix $\mathbf{Y}$. Correspondingly, define the m-by-N matrix of activation functions

$$\Phi(\mathbf{Y}) = \begin{bmatrix} \varphi(y_1(0)) & \varphi(y_1(1)) & \dots & \varphi(y_1(N-1)) \\ \varphi(y_2(0)) & \varphi(y_2(1)) & \dots & \varphi(y_2(N-1)) \\ \vdots & \vdots & & \vdots \\ \varphi(y_m(0)) & \varphi(y_m(1)) & \dots & \varphi(y_m(N-1)) \end{bmatrix}$$

In the batch mode, we define the average weight adjustment (see Eq. (10.100) of the text)

$$\begin{aligned} \Delta\mathbf{W} &= \frac{1}{N}\sum_{n=0}^{N-1}\Delta\mathbf{W}(n) \\ &= \eta\left(\mathbf{I} - \frac{1}{N}\left(\sum_{n=0}^{N-1}\phi(\mathbf{y}(n))\mathbf{y}^T(n)\right)\right)\mathbf{W} \end{aligned}$$

Equivalently, using the matrix definitions introduced in (2), we may write

$$\Delta\mathbf{W} = \eta\left(\mathbf{I} - \frac{1}{N}\Phi(\mathbf{Y})\mathbf{Y}^T\right)\mathbf{W}$$

which is the desired formula.

Problem 10.19

(a) Let $q(\mathbf{y})$ denote a pdf equal to the determinant $\det(\mathbf{J})$ with the elements of the Jacobian $\mathbf{J}$ being as defined in Eq. (10.115). Then using Eq. (10.116) we may express the entropy of the random vector $\mathbf{Z}$ at the output of the nonlinearity in Fig. 10.16 of the text as

$$h(\mathbf{Z}) = -D_{f||q}$$

Invoking the pythogorean decomposition of the Kullback-Leibler divergence, we write

$$D_{f||q} = D_{f||\tilde{f}} + D_{\tilde{f}||q}$$

Hence, the differential entropy

$$h(\mathbf{Z}) = D_{f||\tilde{f}} - D_{\tilde{f}||q} \tag{1}$$

(b) If $q(\mathbf{y}_i)$ happens to equal the source pdf $f_U(\mathbf{y}_i)$ for all i, we then find that $D_{\tilde{f}||q} = 0$. In such a case, (1) reduces to

$$h(\mathbf{Z}) = -D_{f||\tilde{f}}$$

That is, the entropy $h(\mathbf{Z})$ is equal to the negative of the Kullback-Leibler divergence between the pdf $f_{\mathbf{Y}}(\mathbf{y})$ and the corresponding factorial distribution $\tilde{f}_{\mathbf{Y}}(\mathbf{y})$.

Problem 10.20

(a) From Eq. (10.124) in the text,

$$\Phi = \log|\det(\mathbf{A})| + \log|\det(\mathbf{W})| + \sum_i \log\left(\frac{\partial z_i}{\partial y_i}\right)$$

The matrix $\mathbf{A}$ of the linear mixer is fixed. Hence differentiating Φ with respect to $\mathbf{W}$:

$$\frac{\partial \Phi}{\partial \mathbf{W}} = \mathbf{W}^{-T} + \sum_i \frac{\partial}{\partial \mathbf{W}} \log\left(\frac{\partial z_i}{\partial y_i}\right) \tag{1}$$

(b) From Eq. (10.126) pf the text,

$$z_i = \frac{1}{1 + e^{-y_i}}$$

Differentiating z_i with respect to y_i:

$$\begin{aligned} \frac{\partial z_i}{\partial y_i} &= \frac{e^{-y_i}}{(1 + e^{-y_i})^2} \\ &= z_i - z_i^2 \end{aligned} \tag{2}$$

Hence, differentiating $\log\left(\frac{\partial z_i}{\partial y_i}\right)$ with respect to the demixing matrix $\mathbf{W}$, we get

$$\frac{\partial}{\partial \mathbf{W}} \log\left(\frac{\partial z_i}{\partial y_i}\right) = \frac{\partial}{\partial \mathbf{W}} \log(z_i - z_i^2)$$

$$= \frac{\partial z_i}{\partial \mathbf{W}} \frac{\partial}{\partial z_i} \log(z_i - z_i^2)$$

$$= \frac{\partial z_i}{\partial \mathbf{W}} \frac{1}{(z_i - z_i^2)}(1 - 2z_i)$$

$$= \frac{\partial z_i}{\partial y_i} \frac{\partial y_i}{\partial \mathbf{W}} \frac{1}{(z_i - z_i^2)}(1 - 2z_i) \tag{3}$$

But from (2) we have

$$\frac{\partial z_i}{\partial y_i}\left(\frac{1}{z_i - z_i^2}\right) = 1$$

Hence, we may simplify (3) to

$$\frac{\partial}{\partial \mathbf{W}} \log\left(\frac{\partial z_i}{\partial y_i}\right) = \frac{\partial y_i}{\partial \mathbf{W}}(1 - 2z_i)$$

We may thus rewrite (1) as

$$\frac{\partial \Phi}{\partial \mathbf{W}} = \mathbf{W}^{-T} + \sum_i \frac{\partial y_i}{\partial \mathbf{W}}(1 - 2z_i)$$

Putting this relation in matrix form and recognizing that the demixer output **y** is equal to **Wx** where **x** is the observation vector, we find that the adjustment applied to **W** is defined by

$$\Delta \mathbf{W} = \eta \frac{\partial \Phi}{\partial \mathbf{W}}$$

$$= \eta(\mathbf{W}^{-T} + (\mathbf{1} - 2\mathbf{z})\mathbf{x}^T)$$

where η is the learning-rate parameter and **1** is a vector of ones.