CHAPTER 11
Stochastic Methodfs Rooted in Statistical Mechanics

Problem 11.1

By definition, we have

$$p_{ij}^{(n)} = P(X_t = j|X_{t-n} = i)$$

where *t* denotes time and *n* denotes the number of discrete steps. For *n* = 1, we have the one-step transition probability

$$p_{ij}^{(1)} = p_{ij} = P(X_t = j|X_{t-1} = i)$$

For *n* = 2 we have the two-step transition probability

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}$$

where the sum is taken over all intermediate steps *k* taken by the system. By induction, it thus follows that

$$p_{ij}^{(n-1)} = \sum_k p_{ik} p_{kj}^{(n)}$$

Problem 11.2

For $p > 0$, the state transition diagram for the random walk process shown in Fig., P11.2 of the test is irreducible. The reason for saying so is that the system has only one class, namely,
$\{0, \pm 1, \pm 2, ...\}$.

Problem 11.3

The state transition diagram of Fig. P11.3 in the text pertains to a Markov chain with two classes: $\{x_1\}$ and $\{x_1, x_2\}$.

Problem 11.4

The stochastic matrix of the Markov chain in Fig. P11.4 of the text is given by

$$\mathbf{P} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{3}{4} & 0 \end{bmatrix}$$

Let π_1, π_2, and π_3 denote the steady-state probabilities of this chain. We may then write (see Eq. (11.27) of the text)

$$\pi_1 = \pi_1\left(\frac{3}{4}\right) + \pi_2(0) + \pi_3\left(\frac{1}{4}\right)$$

$$\pi_2 = \pi_1\left(\frac{1}{4}\right) + \pi_2\left(\frac{2}{3}\right) + \pi_3\left(\frac{3}{4}\right)$$

$$\pi_3 = \pi_1(0) + \pi_2\left(\frac{1}{3}\right) + \pi_3(0)$$

That is,

$$\pi_1 = \pi_3$$
$$\pi_2 = 3\pi_3$$

We also have, by definition,

$$\pi_1 + \pi_2 + \pi_3 = 1$$

Hence,

$$\pi_3 + 3\pi_3 + \pi_3 = 1$$

or equivalently

$$\pi_3 = \frac{1}{5}$$

and so

$$\pi_1 = \frac{1}{5}$$

$$\pi_2 = \frac{3}{5}$$

Problem 11.6

The Metropolis algorithm and the Gibbs sampler are similar in that they both generate a Markov chain with the Gibbs distribution as the equilibrium distribution.

They differ from each other in the following respect: In the Metropolis algorithm, the transition probabilities of the Markov chain are stationary. In contrast, in the Gibbs sampler, they are nonstationary.

Problem 11.7

Simulated annealing algorithm for solving the travelling salesman problem:

1. Set up an annealing schedule for the algorithm.
2. Initialize the algorithm by picking a tour at random.
3. Choose a pair of cities in the tour and then reverse the order that the cities in-between the selected pairs are visited. This procedure, illustrated in Figure 1 below, generates new tours in a local manner:

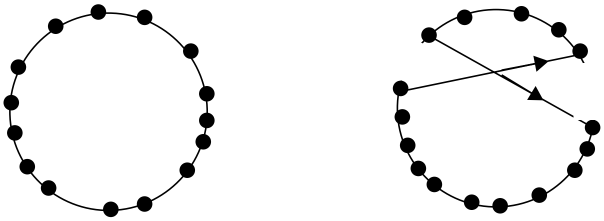

Figure 1: Problem 11.7

4. Calculate the energy difference due to the reversal of paths applied in step 3.
5. If the energy difference so calculated is negative or zero, accept the new tour. If, on the other hand, it is positive, accept the change in the tour with probability defined in accordance with the Metropolis algorithm.
6. Select another pair of cities, and repeat steps 3 to 5 until the required number of iterations is accepted.
7. Lower the temperature in the annealing schedule, and repeat steps 3 to 6.

Problem 11.8

(a) We start with the notion that a neuron j flips from state x_j to $-x_j$ at temperature T with probability

$$P(x_j \rightarrow -x_j) = \frac{1}{1 + \exp(-\Delta E_j / T)} \tag{1}$$

where ΔE_j is the energy difference resulting from such a flip. The energy function of the Boltzman machine is defined by

$$E = -\frac{1}{2}\sum_{i}\sum_{\substack{j \\ i \neq j}} w_{ji}x_i x_j$$

Hence, the energy change produced by neuron j flipping from state x_j to $-x_j$ is

$$\begin{aligned}
\Delta E_j &= \begin{pmatrix}\text{energy with neuron } j \\ \text{in state } x_j\end{pmatrix} - \begin{pmatrix}\text{energy with neuron } j \\ \text{in state } -x_j\end{pmatrix} \\
&= -(x_j)\sum_j w_{ji}x_i + (-x_j)\sum_j w_{ji}x_i \\
&= -2x_j\sum_j w_{ji}x_i \\
&= -2x_j v_j
\end{aligned} \tag{2}$$

where v_i is the induced local field of neuron j.

(b) In light of the result in (2), we may rewrite (1) as

$$P(x_j \rightarrow -x_j) = \frac{1}{1 + \exp(2x_j v_j / T)}$$

This means that for an initial state $x_j = -1$, the probability that neuron j is flipped into state +1 is

$$\frac{1}{1 + \exp(-2v_j / T)} \tag{3}$$

(c) For an initial state of $x_j = +1$, the probability that neuron j is flipped into state -1 is

$$\frac{1}{1 + \exp(+2v_j / T)} = 1 - \frac{1}{1 + \exp(-2v_j / T)} \tag{4}$$

The flipping probability in (4) and the one in (3) are in perfect agreement with the following probabilistic rule

$$x_j = \begin{cases} +1 & \text{with probability } P(v_j) \\ \text{-1} & \text{with probability } 1\text{-}P(v_j) \end{cases}$$

where $P(v_j)$ is itself defined by

$$P(v_j) = \frac{1}{1 + \exp(-2v_j / T)}$$

Problem 11.9

The log-likelihood function $L(\mathbf{w})$ is (see (11.48) of the text)

$$L(\mathbf{w}) = \sum_{\mathbf{x}_\alpha \in T} \log \sum_{\mathbf{x}_\beta} \exp\left(-\frac{E(\mathbf{x})}{T}\right) - \log \sum_{\mathbf{x}} \exp\left(-\frac{E(\mathbf{x})}{T}\right)$$

Differentiating $L(\mathbf{w})$ with respect to weight w_{ji}:

$$\frac{\partial L(\mathbf{w})}{\partial w_{ji}} = \frac{1}{T} \sum_{\mathbf{x}_\alpha \in T} \frac{\partial E(\mathbf{x})}{\partial w_{ji}} \left(-\frac{1}{\sum_{\mathbf{x}_\beta} \exp\left(-\frac{E(\mathbf{x})}{T}\right)} + \frac{1}{\sum_{\mathbf{x}} \exp\left(-\frac{E(\mathbf{x})}{T}\right)} \right)$$

The energy function $E(\mathbf{x})$ is defined by (see (11.39) of the text)

$$E(\mathbf{x}) = -\frac{1}{2} \sum_i \sum_{\substack{j \\ i \neq j}} w_{ji} x_i x_j$$

Hence,

$$\frac{\partial E(\mathbf{x})}{\partial w_{ji}} = -x_i x_j, \qquad i \neq j \tag{1}$$

We also note the following:

$$P(\mathbf{X}_\beta = \mathbf{x}_\beta | \mathbf{X}_\alpha = \mathbf{x}_\alpha) = \frac{1}{\sum_{\mathbf{x}_\beta} \exp\left(-\frac{E(\mathbf{x})}{T}\right)} \tag{2}$$

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{\sum_{\mathbf{x}} \exp\left(-\frac{E(\mathbf{x})}{T}\right)} \tag{3}$$

Accordingly, using the formulas of (1) to (3), we may redefine the derivative $\partial L(\mathbf{w})/\partial w_{ji}$ as follows:

$$\frac{\partial L(\mathbf{w})}{\partial w_{ji}} = \frac{1}{T}\sum_{\mathbf{x}_\alpha \in T}\left(\sum_{\mathbf{x}_\beta} P(\mathbf{X}_\beta = \mathbf{x}_\beta | \mathbf{X}_\alpha = \mathbf{x}_\alpha) x_j x_i - \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}) x_j x_i\right)$$

which is the desired result.

Problem 11.10

(a) Factoring the transition process from state *i* to state *j* into a two-step process, we may express the transition probability p_{ji} as

$$p_{ji} = \tau_{ji} q_{ji} \qquad \text{for } j \neq i \tag{1}$$

where τ_{ji} is the probability that a transition from state *j* to state *i* is attempted, and q_{ji} is the conditional probability that the attempt is successful given that it was attempted. When $j = i$, the property that each row of the stochastic matrix must add to unity implies that

$$\begin{aligned} p_{ii} &= 1 - \sum_{j \neq i} p_{ij} \\ &= 1 - \sum_{j \neq i} \tau_{ij} q_{ij} \end{aligned}$$

(b) We require that the attempt-rate matrix be symmetric:

$$\tau_{ji} = \tau_{ij} \qquad \text{for all } i \neq j \tag{2}$$

and that it satisfies the normalization condition

$$\sum_j \tau_{ji} = 1 \qquad \text{for all } i \neq j$$

We also require the property of complementary conditional transition probability

$$q_{ji} = 1 - q_{ij} \tag{3}$$

For a stationary distribution, we have

$$\pi_i = \sum_j \pi_j p_{ji} \qquad \text{for all } i \tag{4}$$

Hence, using (1) to (3) in (4):

$$\begin{aligned}\pi_i &= \sum_j \pi_j \tau_{ji} p_{ji} \\ &= \sum_j \pi_j \tau_{ji}(1 - q_{ij}) \qquad (5)\end{aligned}$$

Next, recognizing that

$$\sum_j p_{ij} = 1 \qquad \text{for all } i$$

we may go on to write

$$\begin{aligned}\pi_i &= \pi_i \sum_j p_{ij} \\ &= \sum_j \pi_i p_{ij} \\ &= \sum_j \pi_i \tau_{ij} q_{ij} \qquad (6)\end{aligned}$$

Hence, combining (5) and (6), using the symmetry property of (2), and then rearranging terms:

$$\sum_j \tau_{ji}(\pi_i q_{ij} + \pi_j q_{ij} - \pi_j) = 0 \qquad (7)$$

(c) For $\tau_{ji} \neq 0$ the condition of (7) can only be satisfied if

$$\pi_i q_{ij} + \pi_j q_{ij} - \pi_j = 0$$

which, in turn, means that q_{ij} is defined by

$$q_{ij} = \frac{1}{1 + (\pi_i / \pi_j)} \qquad (8)$$

(d) Make a change of variables:

$$E_i = -T \log \pi_i + T^*$$

where T and T^* are arbitrary constants. We may then express π_i in terms of E_i as

$$\pi_i = \frac{1}{Z}\exp\left(-\frac{E_i}{T}\right)$$

where

$$Z = \exp\left(-\frac{T^*}{T}\right)$$

Accordingly, we may reformulate (8) in the new form

$$q_{ij} = \frac{1}{1 + \exp\left(-\frac{1}{T}(E_i - E_j)\right)}$$

$$= \frac{1}{1 + \exp(-\Delta E/T)} \tag{9}$$

where $\Delta E = E_i - E_j$. To evaluate the constant Z, we note that

$$\sum_i \pi_i = 1$$

and therefore

$$Z = \sum_i \exp(-E_i/T)$$

(e) The formula of (9) is the only possible distribution for state transitions in the Boltzmann machine; it is recognized as the Gibbs distribution.

Problem 11.11

We start with the Kullback-Leibler divergence

$$D_{p^+||p^-} = \sum_\alpha p_\alpha^+ \log\left(\frac{p_\alpha^+}{p_\alpha^-}\right) \tag{1}$$

The probability distribution p_α^+ in the clamped condition is naturally independent of the synaptic weights w_{ji} in the Boltzman machine, whereas the probability distribution p_α^- is dependent on w_{ji}. Hence differentiating (1) with respect to w_{ji}:

$$\frac{\partial D_{p^{+}||p^{-}}}{\partial w_{ji}} = -\sum_{\alpha} \frac{p^{+}_{\alpha}}{p^{-}_{\alpha}} \frac{p^{-}_{\alpha}}{\partial w_{ji}} \tag{2}$$

To minimize $D_{p^{+}||p^{-}}$, we use the method of gradient descent:

$$\begin{aligned} \Delta w_{ji} &= -\varepsilon \frac{D_{p^{+}||p^{-}}}{\partial w_{ji}} \\ &= \varepsilon \sum_{\alpha} \frac{p^{+}_{\alpha}}{p^{-}_{\alpha}} \frac{p^{-}_{\alpha}}{\partial w_{ji}} \end{aligned} \tag{3}$$

where ε is a positive constant.

Let $p^{-}_{\alpha\beta}$ denote the joint probability that the visible neurons are in state α and the hidden neurons are in state β, given that the network is in its clamped condition. We may then write

$$p^{-}_{\alpha} = \sum_{\beta} p^{-}_{\alpha\beta}$$

Assuming that the network is in thermal equilibrium, we may use the Gibbs distribution

$$p^{-}_{\alpha\beta} = \frac{1}{Z} \exp\left(-\frac{E_{\alpha\beta}}{T}\right)$$

to write

$$p^{-}_{\alpha} = \frac{1}{Z} \sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right) \tag{4}$$

where $E_{\alpha\beta}$ is the energy of the network when the visible neurons are in state α and the hidden neurons are in state β. The partition function Z is itself defined by

$$Z = \sum_{\alpha} \sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right)$$

The energy $E_{\alpha\beta}$ is defined in terms of the synaptic weights w_{ji} by

$$E_{\alpha\beta} = -\frac{1}{2}\sum_{\substack{i \\ i \neq j}} \sum_{j} w_{ji} x_{j|\alpha\beta} x_{i|\alpha\beta} \tag{5}$$

where $x_{i|\alpha\beta}$ is the state of neuron i when the visible neurons are in state α and the hidden neurons are in state β. Therefore, using (4):

$$\frac{\bar{p}_{\alpha}}{\partial w_{ji}} = -\frac{1}{ZT}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right)\frac{\partial E_{\alpha\beta}}{\partial w_{ji}} - \frac{1}{Z^2}\ \frac{\partial Z}{\partial w_{ji}}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right) \tag{6}$$

From (5) we have (remembering that in a Boltzmann machine $w_{ji} = w_{ij}$)

$$\frac{\partial E_{\alpha\beta}}{\partial w_{ji}} = -x_{j|\alpha\beta} x_{i|\alpha\beta} \tag{7}$$

The first term on the right-hand side of (6) is therefore

$$\begin{aligned} -\frac{1}{ZT}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right)\frac{\partial E_{\alpha\beta}}{\partial w_{ji}} &= +\frac{1}{ZT}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right) x_{j|\alpha\beta} x_{i|\alpha\beta} \\ &= \frac{1}{T}\sum_{\beta} \bar{p}_{\alpha\beta} x_{j|\alpha\beta} x_{i|\alpha\beta} \end{aligned}$$

where we have made use of the Gibbs distribution

$$\bar{p}_{\alpha\beta} = \frac{1}{Z}\exp\left(-\frac{E_{\alpha\beta}}{T}\right)$$

as the probability that the visible neurons are in state α and the hidden neurons are in state β in the free-running condition. Consider next the second term on the right-hand side of (6). Except for the minus sign, we may express this term as the product of two factors:

$$\frac{1}{Z^2}\ \frac{\partial Z}{\partial w_{ji}}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right) = \left[\frac{1}{Z}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right)\right]\left[\frac{1}{Z}\frac{\partial Z}{\partial w_{ji}}\right] \tag{8}$$

The first factor in (8) is recognized as the Gibbs distribution $\bar{p}_{\alpha}$ defined by

$$\bar{p}_{\alpha} = \frac{1}{Z}\sum_{\beta} \exp\left(-\frac{E_{\alpha\beta}}{T}\right) \tag{9}$$

To evaluate the second factor in (8), we write

$$\begin{aligned}\frac{1}{Z}\ \frac{\partial Z}{\partial w_{ji}} &= \frac{1}{Z}\ \frac{\partial}{\partial w_{ji}}\sum_{\alpha}\sum_{\beta}\exp\left(-\frac{E_{\alpha\beta}}{T}\right) \\ &= -\frac{1}{TZ}\ \sum_{\alpha}\sum_{\beta}\exp\left(-\frac{E_{\alpha\beta}}{T}\right)\frac{\partial E_{\alpha\beta}}{\partial w_{ji}} \\ &= \frac{1}{TZ}\ \sum_{\alpha}\sum_{\beta}\exp\left(-\frac{E_{\alpha\beta}}{T}\right)x_{j|\alpha\beta}x_{i|\alpha\beta} \\ &= \frac{1}{T}\ \sum_{\alpha}\sum_{\beta}\overset{-}{p}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta} \end{aligned} \tag{10}$$

Using (9) and (10) in (8):

$$\frac{1}{Z^2}\ \frac{\partial Z}{\partial w_{ji}}\sum_{\beta}\exp\left(-\frac{E_{\alpha\beta}}{T}\right) = \frac{\overset{-}{p}_{\alpha}}{T}\sum_{\alpha}\sum_{\beta}\overset{-}{p}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta} \tag{11}$$

We are now ready to revisit (6) and thus write

$$\frac{\partial \overset{-}{p}_{\alpha}}{\partial w_{ji}} = \frac{1}{T}\sum_{\beta}\overset{-}{p}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta} - \frac{\overset{-}{p}_{\alpha}}{T}\sum_{\alpha}\sum_{\beta}\overset{-}{p}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta}$$

We now make the following observations:

1. The sum of probability $\overset{+}{p}_{\alpha}$ over the states α is unity, that is,

$$\sum_{\alpha}\overset{+}{p}_{\alpha} = 1 \tag{12}$$

2. The joint probability

$$\overset{-}{p}_{\alpha\beta} = \overset{-}{p}_{\beta|\alpha}\overset{-}{p}_{\alpha} \tag{13}$$

Similarly

$$\overset{+}{p}_{\alpha\beta} = \overset{+}{p}_{\beta|\alpha}\overset{+}{p}_{\alpha} \tag{14}$$

3. The probability of a hidden state, given some visible state, is naturally the same whether the visible neurons of the network in thermal equilibrium are clamped in that state by the external environment or arrive at that state by free running of the network, as shown by

$$\overset{-}{p}_{\beta|\alpha} = \overset{+}{p}_{\beta|\alpha} \tag{15}$$

In light of this relation we may rewrite Eq. (13) as

$$p^{-}_{\alpha\beta} = p^{+}_{\beta|\alpha}p^{-}_{\alpha} \tag{16}$$

Moreover, we may write

$$\begin{aligned}\left(\frac{p^{+}_{\alpha}}{p^{-}_{\alpha}}\right)p^{-}_{\alpha\beta} &= p^{+}_{\alpha}p^{+}_{\beta|\alpha} \\ &= p^{+}_{\alpha\beta} \end{aligned} \tag{17}$$

Accordingly, we may rewrite (3) as follows:

$$\begin{aligned}\Delta w_{ji} &= \frac{\varepsilon}{T}\left(\sum_{\alpha}\frac{p^{+}_{\alpha}}{p^{-}_{\alpha}}\sum_{\beta}p^{-}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta} - \sum_{\alpha}p^{+}_{\alpha}\ \sum_{\alpha}\ \sum_{\beta}p^{-}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta}\right) \\ &= \frac{\varepsilon}{T}\left(\sum_{\alpha}\ \sum_{\beta}p^{+}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta} - \sum_{\alpha}\sum_{\beta}p^{-}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta}\right)\end{aligned}$$

Define the following terms:

$$\begin{aligned}\eta \quad &= \quad \text{learning rate parameter} \\ &= \quad \frac{\varepsilon}{T}\end{aligned}$$

$$\begin{aligned}\rho^{+}_{ji} \quad &= \quad < x_{j|\alpha\beta}x_{i|\alpha\beta} >^{+} \\ &= \quad \sum_{\alpha}\sum_{\beta}p^{+}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta}\end{aligned}$$

$$\begin{aligned}\rho^{-}_{ji} \quad &= \quad < x_{j|\alpha\beta}x_{i|\alpha\beta} >^{-} \\ &= \quad \sum_{\alpha}\sum_{\beta}p^{-}_{\alpha\beta}x_{j|\alpha\beta}x_{i|\alpha\beta}\end{aligned}$$

We may then finally formulate the Boltzmann learning rule as

$$\Delta w_{ji} = \eta(\rho^{+}_{ji} - \rho^{-}_{ji})$$

Problem 11.12

(a) We start with the relative entropy:

$$D_{p^+||p^-} = \sum_{\alpha}\sum_{\gamma} p^+_{\alpha\gamma}\log\left(\frac{p^+_{\alpha\gamma}}{p^-_{\alpha\gamma}}\right) \tag{1}$$

From probability theory, we have

$$p^+_{\alpha\gamma} = p^+_{\gamma|\alpha}p^+_{\alpha} \tag{2}$$

$$p^-_{\alpha\gamma} = p^-_{\gamma|\alpha}p^-_{\alpha} = p^-_{\gamma|\alpha}p^+_{\alpha} \tag{3}$$

where, in the last line, we have made use of the fact that the input neurons are always clamped to the environment, which means that

$$p^-_{\alpha} = p^+_{\alpha}$$

Substituting (2) and (3) into (1):

$$D_{p^+||p^-} = \sum_{\alpha} p^+_{\alpha}\sum_{\gamma} p^+_{\gamma|\alpha}\log\left(\frac{p^+_{\gamma|\alpha}}{p^-_{\gamma|\alpha}}\right) \tag{4}$$

where the state α refers to the input neurons and γ refers to the output neurons.

(b) With $p_{\gamma|\alpha}$ denoting the conditional probability of finding the output neurons in state γ, given that the input neurons are in state α, we may express the probability distribution of the output states as

$$p^-_{\gamma} = \sum_{\alpha} p^-_{\gamma|\alpha}p^-_{\alpha}$$

The conditional $p^-_{\gamma|\alpha}$ is determined by the synaptic weights of the network in accordance with the formula

$$p^-_{\gamma|\alpha} = \frac{1}{Z_{1\alpha}}\sum_{\beta}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) \tag{5}$$

where

$$E_{\gamma\beta\alpha} = \frac{1}{2}\sum_{j}\sum_{i} w_{ji}[s_j s_i]_{\gamma\beta\alpha} \tag{6}$$

The parameter $Z_{1\alpha}$ is the partition function:

$$\frac{1}{Z_{1\alpha}} = \sum_{\beta}\sum_{\gamma}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) \tag{7}$$

The function of the Boltzmann machine is to find the synaptic weights for which the conditional probability $p^{-}_{\gamma|\alpha}$ approaches the desired value $p^{+}_{\gamma|\alpha}$.

Applying the gradient method to the relative entropy of (1):

$$\Delta w_{ji} = -\varepsilon \ \frac{\partial D_{p^{+}||p^{-}}}{\partial w_{ji}} \tag{8}$$

Using (4) in (8) and recognizing that $p^{+}_{\gamma|\alpha}$ is determined by the environment (i.e., it is independent of the network), we get

$$\Delta w_{ji} = \varepsilon \ \sum_{\alpha} p^{+}_{\alpha} \sum_{\gamma} \frac{p^{+}_{\gamma|\alpha}}{p^{-}_{\gamma|\alpha}} \ \frac{p^{-}_{\gamma|\alpha}}{\partial w_{ji}} \tag{9}$$

To evaluate the partial derivative $\partial p^{-}_{\gamma|\alpha}/\partial w_{ji}$ we use (5) to (7):

$$\begin{aligned}\frac{p^{-}_{\gamma|\alpha}}{\partial w_{ji}} &= \frac{1}{Z_{1\alpha}}\sum_{\beta}\left(-\frac{1}{T}\right)\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right)\frac{E_{\gamma\beta\alpha}}{\partial w_{ji}} \\ &\quad -\frac{1}{Z^{2}_{1\alpha}} \ \frac{\partial Z_{1\alpha}}{\partial w_{ji}}\sum_{\beta}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) \\ &= \frac{1}{Z_{1\alpha}} \ \sum_{\beta}\frac{1}{T}[s_j s_i]_{\gamma\beta\alpha}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) \\ &\quad +\frac{1}{Z^{2}_{1\alpha}} \ \sum_{\beta}\sum_{\gamma}[s_j s_i]_{\gamma\beta\alpha}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right)\end{aligned} \tag{10}$$

Next, we recognize the following pair of relations:

$$\frac{1}{Z_{1\alpha}}\exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) = p^{-}_{\gamma\beta 1\alpha} \tag{11}$$

$$\frac{1}{Z_{1\alpha}} \sum_{\beta} \sum_{\gamma} [s_j s_i]_{\gamma\beta\alpha} \exp\left(-\frac{E_{\gamma\beta\alpha}}{T}\right) = <s_j s_i>^{-}_{\alpha} \;\; p^{-}_{\gamma|\alpha} \tag{12}$$

where the term $<s_j s_i>^{-}_{\alpha}$ is the averaged correlation of the states s_j and s_i with the input neurons clamped to state α and the network in a free-running condition. Substituting (11) and (12) in (10):

$$\frac{\partial p^{-}_{\gamma|\alpha}}{\partial w_{ji}} = \frac{1}{T}\left(\sum_{\beta} [s_j s_i]_{\gamma\beta\alpha} p^{-}_{\gamma\beta\alpha} - <s_j s_i>^{-}_{\alpha} \;\; p^{-}_{\gamma|\alpha}\right) \tag{13}$$

Next, substituting (13) into (9):

$$\begin{aligned}\Delta w_{ji} = \frac{\varepsilon}{T}\Bigg\{ & \sum_{\alpha} p^{+}_{\alpha} \sum_{\alpha} \sum_{\beta} [s_j s_i]_{\gamma\beta\alpha} p^{-}_{\gamma\beta 1\alpha} \left(\frac{p^{+}_{\gamma|\alpha}}{p^{-}_{\gamma|\alpha}}\right) \\ & - \sum_{\alpha} p^{+}_{\alpha} <s_j s_i>^{-}_{\alpha} \;\; \sum_{\gamma} p^{+}_{\gamma|\alpha} \Bigg\}\end{aligned} \tag{14}$$

We now recognize that

$$\sum_{\alpha} p^{+}_{\alpha} = 1 \qquad \text{for all } \alpha \tag{15}$$

$$\begin{aligned}\sum_{\gamma} \sum_{\beta} [s_j s_i]_{\gamma\beta\alpha} p^{-}_{\gamma\beta|\alpha} \left(\frac{p^{+}_{\gamma|\alpha}}{p^{-}_{\gamma|\alpha}}\right) &= \sum_{\gamma} p^{+}_{\gamma|\alpha} \sum_{\beta} [s_j s_i]_{\gamma\beta\alpha} \left(\frac{p^{+}_{\gamma\beta|\alpha}}{p^{-}_{\gamma|\alpha}}\right) \\ &= \sum_{\gamma} p^{+}_{\gamma|\alpha} <s_j s_i>_{\gamma\alpha} \\ &= <s_j s_i>^{+}_{\alpha}\end{aligned} \tag{16}$$

Accordingly, substituting (15) and (16) into (14):

$$\begin{aligned}\Delta w_{ji} &= \frac{\varepsilon}{T} \sum_{\alpha} p^{+}_{\alpha} (<s_j s_i>^{+}_{\alpha} - <s_j s_i>^{-}_{\alpha}) \\ &= \eta \sum_{\alpha} p^{+}_{\alpha} (\rho^{+}_{ji|\alpha} - \rho^{-}_{ji|\alpha})\end{aligned}$$

where $\eta = \varepsilon / T$; and $\rho^{+}_{ji|\alpha}$ and $\rho^{-}_{ji|\alpha}$ are the averaged correlations in the clamped and free-running conditions, given that the input neurons are in state α.

Problem 11.15

Consider the expected distortion (energy)

$$E = \sum_{\mathbf{x}} \sum_{j} P(\mathbf{x} \in C_j) d(\mathbf{x}, \mathbf{y}_j) \qquad (1)$$

where $d(\mathbf{x}, \mathbf{y}_j)$ is the distortion measure for representing the data point $\mathbf{x}$ by the vector $\mathbf{y}_j$, and $P(\mathbf{x} \in C_j)$ is the probability that $\mathbf{x}$ belongs to the cluster of points represented by $\mathbf{y}_j$. To determine the association probabilities at a given expected distortion, we maximize the entropy subject to the constraint of (1). For a fixed $Y = \{\mathbf{y}_j\}$, we assume that the association probabilities of different data points are independent. We may thus express the entropy as

$$H = -\sum_{\mathbf{x}} \sum_{j} P(\mathbf{x} \in C_j) \log P(\mathbf{x} \in C_j) \qquad (2)$$

The probability distribution that maximizes the entropy under the expectation constraint is the Gibbs distribution:

$$P(\mathbf{x} \in C_j) = \frac{1}{Z_{\mathbf{x}}} \exp\left(-\frac{1}{T} d(\mathbf{x}, \mathbf{y}_j)\right) \qquad (3)$$

where

$$Z_{\mathbf{x}} = \sum_{j} \exp\left(-\frac{1}{T} d(\mathbf{x}, \mathbf{y}_j)\right)$$

is the partition function. The inverse temperature $B = 1/T$ is the Lagrange multiplier defined by the value of E in (1).

Problem 11.6

(a) The free energy is

$$F = D - TH \qquad (1)$$

where D is the expected distortion, T is the temperature, and H is the conditional entropy. The expected distortion is defined by

$$D = \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})d(\mathbf{x}, \mathbf{y}) \tag{2}$$

The conditional entropy if defined by

$$H(\mathbf{Y}|\mathbf{X}) = -\sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})\log P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) \tag{3}$$

The minimizing $P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$ is itself defined by the Gibbs distribution:

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = \frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right) \tag{4}$$

where

$$Z_{\mathbf{x}} = \sum_{y}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right) \tag{5}$$

is the partition function. Substituting (2) to (5) into (1), we get

$$\begin{aligned} F^* &= \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})\frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right)d(\mathbf{x}, \mathbf{y}) \\ &\quad + T\sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})\frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right)\left(-\log Z_{\mathbf{x}} - \frac{1}{T}d(\mathbf{x}, \mathbf{y})\right) \\ &= T\sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})\frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right)(- \log Z_{\mathbf{x}}) \end{aligned}$$

This result simplifies as follows by virtue of the definition given in (5) for the partition function:

$$F^* = -T\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x})\log Z_{\mathbf{x}} \tag{6}$$

(b) Differentiating the minimum free energy F* of (6) with respect to **y**:

$$\frac{\partial F^*}{\partial \mathbf{y}} = -T\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x})\frac{1}{Z_{\mathbf{x}}}\ \frac{\partial Z_{\mathbf{x}}}{\partial \mathbf{y}} \tag{7}$$

Using the definition of $Z_{\mathbf{x}}$ given in (5), we write:

$$\frac{\partial Z_{\mathbf{x}}}{\partial \mathbf{y}} = -\frac{1}{T}\sum_{\mathbf{y}}\exp\left(-\frac{d(\mathbf{x}, \mathbf{y})}{T}\right)\frac{\partial d(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \tag{8}$$

Hence, we may rewrite (7) as

$$\begin{aligned}\frac{\partial F^*}{\partial \mathbf{y}} &= \sum_{\mathbf{x}}\sum_{\mathbf{y}} P(\mathbf{X} = \mathbf{x})\frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d(\mathbf{x},\mathbf{y})}{T}\right)\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \\ &= \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x})P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = x)\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} \qquad (9)\end{aligned}$$

where use has been made of (4). Noting that

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})P(\mathbf{X} = \mathbf{x})$$

we may then state that the condition for minimizing the Lagrangian with respect to **y** is

$$\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} = \mathbf{0} \qquad \text{for all } \mathbf{y} \qquad (10)$$

Normalizing this result with respect to $P(\mathbf{X} = \mathbf{x})$ we get the minimizing condition:

$$\sum_{\mathbf{x}} P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} = \mathbf{0} \qquad \text{for all } \mathbf{y} \qquad (11)$$

(c) Consider the squared Euclidean distortion

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2 = (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})$$

for which we have

$$\begin{aligned}\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} &= \frac{\partial}{\partial \mathbf{y}}(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \\ &= -2((\mathbf{x}-\mathbf{y})) \qquad (12)\end{aligned}$$

For this particular measure we find it more convenient to normalize (10) with respect to the probability

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x},\mathbf{Y} = \mathbf{y})$$

We may then write the minimizing condition with respect to **y** as

$$\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})\frac{\partial d(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}} = 0 \qquad (13)$$

Using (12) in (13) and solving for **y**, we get the desired minimizing solution

$$\mathbf{y} = \frac{\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})\mathbf{x}}{\sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})} \tag{14}$$

which is recognized as the formula for a centroid.

Problem 11.17

The advantage of deterministic annealing over maximum likelihood is that it does not make any assumption on the underlying probability distribution of the data.

Problem 11.18

(a) Let

$$\varphi_k(\mathbf{x}) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} = \mathbf{t}_k\|^2\right), \qquad k = 1, 2, ..., K$$

where $\mathbf{t}_k$ is the center or prototype vector of the kth radial basis function and K is the number of such functions (i.e., hidden units). Define the normalized radial basis function

$$P_k(\mathbf{x}) = \frac{\varphi_k(\mathbf{x})}{\sum_k \varphi_k(\mathbf{x})}$$

The average squared cost over the training set is

$$d = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{F}(\mathbf{x}_i)\|^2 \tag{1}$$

where $\mathbf{F}(\mathbf{x}_i)$ is the output vector of the RBF network in response to the input $\mathbf{x}_i$. The Gibbs distribution for $P(\mathbf{x} \in R)$ is

$$P(\mathbf{x} \in R) = \frac{1}{Z_{\mathbf{x}}}\exp\left(-\frac{d}{T}\right) \tag{2}$$

where d is defined in (1) and

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}_i} \exp\left(-\frac{d}{T}\right) \tag{3}$$

(b) The Lagrangian for minimizing the average misclassification cost is

$F = d - TH$

where the average squared cost d is defined in (1), and the entropy H is defined by

$$H = -\sum_{\mathbf{x}} \sum_{j} p(j|\mathbf{x}) \log(j|\mathbf{x})$$

where $p(j|\mathbf{x})$ is the probability of associating class j at the output of the RBF network with the input **x**.