The second equality is valid if $\Gamma_{\mathbf{XX}}$ is nonsingular. See Exercise 4.17. Comparing (4.46) with (4.21), we make the following observation.

Remark 4.36. If we assume the linear data model (4.16), where $\mathbf{X}$ and $\mathbf{N}$ are independent random vectors with zero means, then the form of the minimum variance linear estimator (4.46) is the same as that of the MAP estimator (4.21). Note that the derivation of the MAP estimator required the additional assumption that $\mathbf{X}$ and $\mathbf{N}$ are Gaussian random vectors.

4.5 The EM Algorithm

Let $\mathbf{Y}$ be a random vector with a parameter-dependent probability distribution. The EM algorithm is an iterative procedure that, given a realization of $\mathbf{Y}$, yields a sequence of approximations to a maximum likelihood estimator for the parameter. The algorithm relies on an auxiliary random vector $\mathbf{X}$ that corresponds to hidden or missing data. $\mathbf{X}$ and $\mathbf{Y}$ together make up the complete data. A very general development can be found in [83]. For simplicity, we suppose that $\mathbf{X}$ and $\mathbf{Y}$ are discrete and let them have a joint probability mass function $p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}; \theta)$, where θ denotes the parameter of interest. Then the conditional probability mass function for $\mathbf{X}$ given $\mathbf{Y}$ is

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta) = \frac{p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}; \theta)}{p_{\mathbf{Y}}(\mathbf{y}; \theta)}, \tag{4.47}$$

where the denominator gives the marginal probability mass function of Y,

$$p_{\mathbf{Y}}(\mathbf{y}; \theta) = \sum_{\mathbf{x}} p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}; \theta). \tag{4.48}$$

By $\sum_{\mathbf{x}}$ we mean the sum over components $\mathbf{x}$ for which $P\{\mathbf{X} = \mathbf{x}\} > 0$. The log likelihood function for $\mathbf{Y}$ given observed data $\mathbf{y}$ takes the form

$$\begin{aligned} l_{\mathbf{Y}}(\theta; \mathbf{y}) &= \log p_{\mathbf{Y}}(\mathbf{y}; \theta) \\ &= \log p_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}; \theta) - \log p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta) \\ &= l_{(\mathbf{X},\mathbf{Y})}(\theta; \mathbf{x}, \mathbf{y}) - l_{\mathbf{X}|\mathbf{Y}}(\theta; \mathbf{x}|\mathbf{y}). \end{aligned}$$

Then for any fixed parameter θ_ν,

$$\begin{aligned} l_{\mathbf{Y}}(\theta; \mathbf{y}) &= l_{\mathbf{Y}}(\theta; \mathbf{y}) \sum_{\mathbf{x}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta_\nu) \quad \text{by (4.47)–(4.48)} \\ &= \sum_{\mathbf{x}} l_{\mathbf{Y}}(\theta; \mathbf{y})\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta_\nu) \\ &= \sum_{\mathbf{x}} l_{(\mathbf{X},\mathbf{Y})}(\theta; \mathbf{x}, \mathbf{y})\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta_\nu) - \sum_{\mathbf{x}} l_{\mathbf{X}|\mathbf{Y}}(\theta; \mathbf{x}|\mathbf{y})\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}; \theta_\nu) \\ &\stackrel{\text{def}}{=} Q(\theta|\mathbf{y}; \theta_\nu) - H(\theta|\mathbf{y}; \theta_\nu). \end{aligned}$$

Proposition 4.37. *If*

$$Q(\theta_{\nu+1}|\mathbf{y}; \theta_\nu) \geq Q(\theta_\nu|\mathbf{y}; \theta_\nu),$$

then

$$l_{\mathbf{Y}}(\theta_{\nu+1}; \mathbf{y}) \geq l_{\mathbf{Y}}(\theta_\nu; \mathbf{y}).$$

Proof. For any parameter $\theta_{\nu+1}$,

$$l_{\mathbf{Y}}(\theta_{\nu+1};\mathbf{y}) - l_{\mathbf{Y}}(\theta_\nu;\mathbf{y}) = [Q(\theta_{\nu+1}|\mathbf{y};\theta_\nu) - Q(\theta_\nu|\mathbf{y};\theta_\nu)] + [H(\theta_\nu|\mathbf{y};\theta_\nu) - H(\theta_{\nu+1}|\mathbf{y};\theta_\nu)].$$

It suffices to show that the second bracketed term on the right-hand side is nonnegative. Note that

$$\begin{aligned} H(\theta_\nu|\mathbf{y};\theta_\nu) - H(\theta_{\nu+1}|\mathbf{y};\theta_\nu) &= -\sum_{\mathbf{x}} [l_{\mathbf{X}|\mathbf{Y}}(\theta_{\nu+1};\mathbf{x}|\mathbf{y}) - l_{\mathbf{X}|\mathbf{Y}}(\theta_\nu;\mathbf{x}|\mathbf{y})]\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_\nu) \\ &= -\sum_{\mathbf{x}} \log\left(\frac{p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_{\nu+1})}{p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_\nu)}\right)\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_\nu) \\ &\geq -\log\sum_{\mathbf{x}} p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_\nu) \\ &= 0. \end{aligned}$$

The last inequality follows from the convexity of the negative log (see Exercise 4.18), while the equality that follows it comes from (4.47)–(4.48). □

This proposition motivates the following iterative procedure.

Algorithm 4.5.1. The EM Algorithm.
To maximize the log likelihood function $l_{\mathbf{Y}}(\theta;\mathbf{y})$, given a realization $\mathbf{y}$ of a random vector $\mathbf{Y}$ and an initial guess θ_0 for the parameter θ,

for $\nu = 0, 1, \ldots,$ repeat,

1. Compute $Q(\theta|\mathbf{y};\theta_\nu)$, the conditional expectation of the log likelihood function for the complete data, given the observed $\mathbf{y}$ and the MLE approximation θ_ν. In the discrete case, this takes the form

$$Q(\theta|\mathbf{y};\theta_\nu) = \sum_{\mathbf{x}} l_{(\mathbf{X},\mathbf{Y})}(\theta;\mathbf{x},\mathbf{y})\ p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y};\theta_\nu). \tag{4.49}$$

2. Compute a maximizer $\theta_{\nu+1}$ of $Q(\theta|\mathbf{y};\theta_\nu)$.

Step 1 is called the E-step, and step 2 is called the M-step. That the sequence $\{\theta_\nu\}$ actually converges to a maximum likelihood estimator can be confirmed under fairly mild conditions. See [83] and the references therein.

4.5.1 An Illustrative Example

The following development is taken from Vardi and Lee [111]. In applications like nonnegative image reconstruction (see Chapter 9) it is important to find a nonnegative solution $\mathbf{f}$ to a linear system

$$K\mathbf{f} = \mathbf{g},$$

where the $m \times n$ coefficient matrix K and the vector $\mathbf{g} \in \mathbb{R}^n$ have nonnegative components. Of course such a solution need not be attained, so we seek to minimize some measure of the discrepancy between the data $\mathbf{g}$ and the model $K\mathbf{f}$. Here we minimize the Kullback–Leibler information divergence (see (2.52)):

$$\rho_{KL}(\mathbf{g}, K\mathbf{f}) = \sum_{i=1}^{m} g_i(\log g_i - \log[K\mathbf{f}]_i), \tag{4.50}$$

subject to

(4.51)

By a suitable

(4.52)

together with

(4.53)

(4.54)

Then we have

(4.55)

(4.56)

(4.57)

(4.58)

These conditi
Minimi

(4.59)

subject to the
random varia
joint probabi

(4.60)

Here $\mathbf{f} \in \mathbb{R}$
that $p_{(X,Y;)}(j$
probability n

(4.61)

Suppose we
number, ther

(4.62)

is an integer
take r indep

subject to

$$f_j \geq 0, \qquad j = 1, \ldots, n. \tag{4.51}$$

By a suitable rescaling (see Exercise 4.19) one may assume

$$\sum_{j=1}^{n} f_j = 1, \tag{4.52}$$

together with the following conditions on the entries of K:

$$k_{ij} \geq 0, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{4.53}$$

$$\sum_{i=1}^{m} k_{ij} = 1, \qquad j = 1, \ldots, n. \tag{4.54}$$

Then we have

$$g_i \geq 0, \qquad i = 1, \ldots, m, \tag{4.55}$$

$$\sum_{i=1}^{m} g_i = 1, \tag{4.56}$$

$$[K\mathbf{f}]_i \geq 0, \tag{4.57}$$

$$\sum_{i=1}^{m} [K\mathbf{f}]_i = 1. \tag{4.58}$$

These conditions guarantee that $(\mathbf{g}, K\mathbf{f})$ lies in the domain of ρ_{KL}.

Minimizing (4.50) under the above conditions is equivalent to maximizing

$$J(\mathbf{f}) = \sum_{i=1}^{m} g_i \log[Kf]_i \tag{4.59}$$

subject to the constraints (4.51) and (4.52). To apply the EM algorithm, we first construct random variables X and Y, with support $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively, and with joint probability mass function

$$P\{X = j, Y = i\} = p_{(X,Y)}(j, i; \mathbf{f}) = k_{ij} f_j. \tag{4.60}$$

Here $\mathbf{f} \in \mathbb{R}^n$ is the parameter to be estimated. The conditions (4.51)–(4.54) guarantee that $p_{(X,Y)}(j, i|\mathbf{f})$ is indeed a probability mass function. See Exercise 4.20. The marginal probability mass function for Y is then

$$p_Y(i; \mathbf{f}) = \sum_{j=1}^{n} p_{(X,Y)}(j, i; \mathbf{f}) = \sum_{j=1}^{n} k_{ij} f_j = [K\mathbf{f}]_i. \tag{4.61}$$

Suppose we are given observed data $\mathbf{g} \in \mathbb{R}^m$ satisfying (4.55)–(4.56). If each g_i is a rational number, there exists a positive integer r such that

$$N_i = r g_i \tag{4.62}$$

is an integer for each i. The assumption that each g_i is rational can be relaxed [111]. Now take r independent, identically distributed copies of Y to obtain a random vector $\mathbf{Y}$, and let

$\mathbf{y} = (y_1, \ldots, y_r)$ be a realization for which N_i gives the number of indices k with $y_k = i$. Then the log likelihood function for Y, given data $\mathbf{y}$ and parameter $\mathbf{f}$, is

$$\begin{aligned} l_Y(\mathbf{f}; \mathbf{y}) &= \sum_{k=1}^{r} \log p_Y(y_k; \mathbf{f}) \\ &= \sum_{k=1}^{r} \left(\sum_{i=1}^{m} \delta_{y_k,i} \right) \log p_Y(y_k; \mathbf{f}) \\ &= \sum_{i=1}^{m} \sum_{k=1}^{r} \delta_{y_k,i} \log p_Y(y_k; \mathbf{f}) \\ &= \sum_{i=1}^{m} N_i \log p_Y(y_i; \mathbf{f}) \\ &= r \sum_{i=1}^{m} g_i \log([K\mathbf{f}]_i) \qquad \text{by (4.61) and (4.62).} \end{aligned}$$

This establishes the connection between maximum likelihood estimation and nonnegatively constrained linear equations; see (4.59).

In a similar manner, we can construct r copies of X to obtain a random vector $\mathbf{X}$ for which the pairs (X_k, Y_k) are independent and distributed according to (4.60). Here $\mathbf{X}$ is the hidden data vector, and $\mathbf{X}$ and $\mathbf{Y}$ together make up the complete data. Take a realization $(\mathbf{x}, \mathbf{y})$ with pairs (x_k, y_k), $k = 1, \ldots, r$, for which $N_{ij}(\mathbf{y}, \mathbf{x})$ denotes the number of indices k such that $y_k = i$ and $x_k = j$. Note that for each i,

$$\sum_{j=1}^{n} N_{ij}(\mathbf{x}, \mathbf{y}) = N_i = r g_i, \tag{4.63}$$

the number of occurrences of $y_k = i$. The log likelihood function for the complete data is given by

$$l_{\mathbf{X},\mathbf{Y}}(\mathbf{f}; \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \sum_{i=1}^{m} N_{ij}(\mathbf{x}, \mathbf{y}) \, [\log k_{ij} + \log f_j].$$

From (4.47)–(4.48) and (4.60),

$$p_{X|Y}(j|i; \mathbf{f}_\nu) = \frac{k_{ij} f_j^\nu}{\sum_{l=1}^{n} k_{il} f_l^\nu} \stackrel{\text{def}}{=} \hat{p}_{ij}^\nu, \tag{4.64}$$

where f_j^ν denotes the jth component of $\mathbf{f}_\nu$. Then by (4.49),

$$\begin{aligned} Q(\mathbf{f}|\mathbf{y}; \mathbf{f}_\nu) &= \sum_{\mathbf{x}} \left(\sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij}(\mathbf{x}, \mathbf{y}) \, [\log k_{ij} + \log f_j^\nu] \right) \hat{p}_{ij}^\nu \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} r g_i \, [\log k_{ij} + \log f_j^\nu] \, \hat{p}_{ij}^\nu. \end{aligned} \tag{4.65}$$

The second equality follows from (4.63). This completes the E-step of the algorithm.

Exercises

To imp
(4.51)–(4.52

(4.66)

Exercise

4.1. Show
Provi
4.2. Show
C exi
4.3. Supp
X_i ha
the m
estim
4.4. For th
diag(
4.5. Show
the in
4.6. Cons
fair c
4.7. Verify
distri
4.8. Prove
4.9. Show
4.10. Show
(4.21
using
transf
4.11. From
4.12. Prove
4.13. If A i
$U^T U$
(4.28
4.14. Verify
4.15. Prove
4.16. Verify
4.17. Show
4.18. Show

Show

To implement the M-step, we maximize Q with respect to $\mathbf{f}$ subject to the constraints (4.51)–(4.52). This yields (see Exercise 4.21) the vector $\mathbf{f}_{v+1}$ with components

$$f_j^{v+1} = f_j^v \sum_{i=1}^{m} k_{ij} \left(\frac{g_i}{\sum_{l=1}^{n} k_{il} f_l^v} \right), \qquad j = 1, \ldots, n. \tag{4.66}$$

Exercises

4.1. Show that the covariance matrix C (see (4.3)) is symmetric and positive semidefinite. Provide an example showing that C need not be positive definite.

4.2. Show that if a random vector has independent components and the covariance matrix C exists, then C is diagonal.

4.3. Suppose that for each $i = 1, \ldots, n$, d_i is a realization of a Gaussian random variable X_i having mean μ and variance $\sigma^2 > 0$. Show that if the X_i's are independent, then the maximum likelihood estimator for μ is $\sum_{i=1}^{n} d_i/n$ and the maximum likelihood estimator for σ^2 is $\sum_{i=1}^{n}(d_i - \mu)^2/n$.

4.4. For the Poisson random vector in Example 4.14, show that $E(\mathbf{X}) = \boldsymbol{\lambda}$ and $\operatorname{cov}(\mathbf{X}) = \operatorname{diag}(\lambda_i, \ldots, \lambda_n)$.

4.5. Show that the negative Poisson log likelihood function in (4.8) is strictly convex on the interior of the nonnegative orthant $\mathbb{R}^n_+$ and that it has $\mathbf{d}$ as its unique minimizer.

4.6. Construct tables analogous to those in Example 4.21 for the toss of three independent, fair coins.

4.7. Verify equation (4.11) under the assumption that X and Y are independent, jointly distributed, discrete random variables.

4.8. Prove Theorem 4.24.

4.9. Show that the expression (4.21) gives the MAP estimator in Example 4.26.

4.10. Show that under the assumptions of Example 4.26, the right-hand side of equation (4.21) gives the conditional expectation, $E(\mathbf{X}|\mathbf{Z} = \mathbf{z})$. This can be most easily done using characteristic functions. These are essentially the expected values of the Fourier transforms of random variables.

4.11. From (4.21), derive expression (4.22) with $\alpha = (\sigma_{\mathbf{N}}/\sigma_{\mathbf{X}})^2$.

4.12. Prove Proposition 4.29.

4.13. If A is symmetric, it has an orthogonal eigendecomposition $A = U \operatorname{diag}(\lambda_i) U^T$ with $U^T U = U U^T = I$. Use this fact, along with $a_{ii} = \mathbf{e}_i^T A \mathbf{e}_i$, to prove the equality (4.28).

4.14. Verify equation (4.38).

4.15. Prove Proposition 4.35.

4.16. Verify equations (4.43)–(4.45).

4.17. Show that $\Gamma_{\mathbf{XX}} K^T [K \Gamma_{\mathbf{XX}} K^T + C_{\mathbf{N}}]^{-1} = [K^T C_{\mathbf{N}}^{-1} K + \Gamma_{\mathbf{XX}}^{-1}]^{-1} K^T C_{\mathbf{N}}^{-1}$.

4.18. Show that if J is convex, $\sum_i w_i = 1$, and $w_i \geq 0$, then

$$J\left(\sum_i z_i w_i \right) \leq \sum_i J(z_i)\, w_i .$$

Show that the inequality is strict if the z_i's are not all equal and each $w_i > 0$.