we want to estimate may be dropped; this includes factors which depend on the data but not on $\boldsymbol{\alpha}$.

If the solution to Eq. (17.6) is at a maximum, $\partial \ln \mathscr{L}/\partial \alpha_n$ will have negative slope in its vicinity. In many practical problems, one often uses nonlinear algorithms for finding the maximum, and must be alert to various possibilities for error: (a) Eq. (17.6) may yield a minimum, therefore one must check the second derivative; (b) there may be more than one maximum—one must try to find the global maximum; (c) the global maximum may lie at a boundary of the physical region, in which case Eq. (17.6) will not find it.

If an unbiased, efficient estimator exists, this method will find it. If $\partial \ln \mathscr{L}/\partial \alpha_n$ is linear in the vicinity of the root, an efficient estimator is guaranteed; other efficient cases are discussed in the literature. For large data samples, the central limit theorem will usually assure this condition in some significant neighborhood of zero; hence the estimator is usually efficient in that case, provided certain conditions are met (*e.g.*, that the solution does not lie on a boundary). In this case, in the neighborhood of the maximum $\ln \mathscr{L}$ is a downward-curving paraboloid and $\mathscr{L}$ is proportional to a multivariate Gaussian.

The results of two or more experiments may be combined by forming the product of the $\mathscr{L}$'s, or the sum of the $\ln \mathscr{L}$'s.

Under a one-to-one change of parameters from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\alpha})$, the maximum likelihood estimate is simply $\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})$, given the solution $\widehat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$. That is, the maximum likelihood solution for $\boldsymbol{\beta}$ is found by simple substitution of $\widehat{\boldsymbol{\alpha}}$ into the transformation equation. It is possible that the new solution $\widehat{\boldsymbol{\beta}}$ will be a biased solution for the true value of $\boldsymbol{\beta}$ even if $\widehat{\boldsymbol{\alpha}}$ is not biased, and vice-versa. In the asymptotic limit (of large amounts of data) both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ will (usually) converge to unbiased solutions, but at different rates.

Except in special cases like the least-squares method, the value of the likelihood function at the solution does not necessarily tell us whether the final fit was a sensible description of the data or not. In special cases such as the one discussed in Sec. 17.3.3, one can define a quantity which approaches the $\chi^2$-distribution in the limit of a large number of counts in the experiment, but in general some other strategy must be used. For example, data generated by Monte Carlo simulations of the experiment can be analyzed by the same method. If the experimental likelihood is lower than that of some agreed-upon fraction of these results, one should question the appropriateness of the p.d.f. At the same time one can check for bias in the solution.

#### 17.3.2. *Error estimates*:

The covariance matrix $V$ may be estimated from

$$V_{nm} = \left( E \left[ - \frac{\partial^2 \ln \mathscr{L}}{\partial \alpha_n \, \partial \alpha_m} \bigg|_{\widehat{\alpha}} \right] \right)^{-1} . \tag{17.7}$$

If $\partial \ln \mathscr{L}/\partial \alpha_n$ is linear, the "expectation" operation in Eq. (17.7) has no effect because the second derivative of $\ln \mathscr{L}$ is constant. Otherwise, it may be approximated by taking the average of the quantity in square brackets over a range of $\alpha_n$ and $\alpha_m$ near the solution. For complex cases it may be more practical to evaluate $s$-standard-deviation errors from the contour

$$\ln \mathscr{L}(\boldsymbol{\alpha}) = \ln \mathscr{L}_{\max} - s^2/2 \ , \tag{17.8}$$

where $\ln \mathscr{L}_{\max}$ is the value of $\ln \mathscr{L}$ at the solution point (compare with Eq. (17.27), below). The extreme limits of this contour parallel to the $\alpha_n$ axis give an approximate $s$-standard-deviation confidence interval in $\alpha_n$. These intervals may not be symmetric and they may even consist of two or more disjoint intervals. This procedure gives one-standard-deviation errors in $\alpha_n$ equal to $\sqrt{V_{nn}}$ (not summed) of Eq. (17.7) if the estimator is efficient. If it is not efficient, the level of confidence implied by the value of $s$ is only approximate.

#### 17.3.3. *Application to Poisson-distributed data*:

In the case of Poisson-distributed data in a counting experiment, the unbinned maximum likelihood method (where the index $i$ in Eq. (17.5) labels events) is preferred if the total number of events is very small. If there are enough events to justify binning them in a histogram, then one may alternatively maximize the likelihood function for the contents of the bins (so $i$ labels bins). This is equivalent to minimizing [7]

$$\chi^2 = \sum_i \left[ 2(N_i^{\rm th} - N_i^{\rm obs}) + 2N_i^{\rm obs} \ln(N_i^{\rm obs}/N_i^{\rm th}) \right] . \tag{17.9}$$

where $N_i^{\rm obs}$ and $N_i^{\rm th}$ are the observed and theoretical (from $f$) contents of the $i$th bin. In bins where $N_i^{\rm obs} = 0$, the second term is zero. This function asymptotically behaves like a classical $\chi^2$ for purposes of point estimation, interval estimation, *and goodness-of-fit.* It also guarantees that the area under the fitted function $f$ is equal to the sum of the histogram contents (as long as the overall normalization of $f$ is effectively left unconstrained during the fit), which is not the case for $\chi^2$ statistics based on a least-squares procedure with traditional weights.

## 17.4. Method of least squares

The *method of least squares* can be derived from the maximum likelihood theorem. We suppose a set of $N$ measurements at points $x_i$. The $i$th measurement $y_i$ is assumed to be chosen from a Gaussian distribution with mean $F(x_i;\boldsymbol{\alpha})$ and variance $\sigma_i^2$. Then

$$\chi^2 = -2 \ln \mathscr{L} + \text{constant} = \sum_1^N \frac{[y_i - F(x_i;\boldsymbol{\alpha})]^2}{\sigma_i^2} . \tag{17.10}$$

Finding the set of parameters $\boldsymbol{\alpha}$ which maximizes $\mathscr{L}$ is the same as finding the set which minimizes $\chi^2$.

In many practical cases one further restricts the problem to the situation in which $F(x_i;\boldsymbol{\alpha})$ is a linear function of the $\alpha_m$'s,

$$F(x_i;\boldsymbol{\alpha}) = \sum_n \alpha_n \, f_n(x) \ , \tag{17.11}$$

where the $f_n$ are $k$ linearly independent functions (*e.g.*, 1, $x$, $x^2$, ..., or Legendre polynomials) which are single-valued over the allowed range of $x$. We require $k \le N$, and at least $k$ of the $x_i$ must be distinct. We wish to estimate the linear coefficients $\alpha_n$. Later we will discuss the nonlinear case.

If the point errors $\epsilon_i = y_i - F(x_i;\boldsymbol{\alpha})$ are Gaussian, then the minimum $\chi^2$ will be distributed as a $\chi^2$ random variable with $n = N - k$ degrees of freedom. We can then evaluate the goodness-of-fit (confidence level) from Figs. 16.1 or 16.3, as per the earlier discussion. The confidence level expresses the probability that a *worse* fit would be obtained in a large number of similar experiments under the assumptions that: (a) the model $y = \sum \alpha_n \, f_n$ is correct and (b) the errors $\epsilon_i$ are Gaussian and unbiased with variance $\sigma_i^2$. If this probability is larger than an agreed-upon value (0.001, 0.01, or 0.05 are common choices), the data are *consistent* with the assumptions; otherwise we may want to find improved assumptions. As for the converse, most people do not regard a model as being truly *inconsistent* unless the probability is as low as that corresponding to four or five standard deviations for a Gaussian ($6 \times 10^{-3}$ or $6 \times 10^{-5}$; see Sec. 17.5.1). If the $\epsilon_i$ are not Gaussian, the method of least squares still gives an answer, but the goodness-of-fit test would have to be done using the correct distribution of the random variable which is still called "$\chi^2$."

Finding the minimum of $\chi^2$ in the linear case is straightforward:

$$\begin{aligned} -\frac{1}{2}\frac{\partial \chi^2}{\partial \alpha_m} &= \sum_i f_m(x_i) \left( \frac{y_i - \sum_n \alpha_n \, f_n(x_i)}{\sigma_i^2} \right) \\ &= \sum_i \frac{y_i \, f_m(x_i)}{\sigma_i^2} - \sum_n \alpha_n \sum_i \frac{f_n(x_i) \, f_m(x_i)}{\sigma_i^2} . \end{aligned} \tag{17.12}$$