CHAPTER 5

Nonparametric Maximum Likelihood

We now return to the nonparametric maximum likelihood problem that was introduced in Section 1.6 of Chapter 1, and do the necessary theory to prove the results given there.

The problem is to maximize the mixture likelihood

$$L(Q) = \prod_{i=1}^{n} L_i(Q) = \prod_{j=1}^{D} \left[\int L_j(\phi)\, dQ(\phi) \right]^{n_j}. \tag{5.1}$$

Here $L_j(\phi)$ is the *likelihood kernel*, generally the one-component likelihood for a single observation, say y_j, and n_j is the number of times y_j was observed. The likelihood kernel may well depend on other auxiliary parameters and covariates, which will be held fixed in this discussion. As far as the maximization problem is concerned, the only critical assumption is that L_j is a nonnegative function of ϕ and that the number D is minimal among all such product representations. That is, the terms have been grouped to the maximal extent. In the multinomial setting, this can substantially reduce the number of terms in the product.

5.1. The optimization framework. The basic results concerning the nonparametric maximum likelihood estimator $\hat{Q}$ have already been outlined in Section 1.6. These results can be derived by putting the problem of likelihood maximization into the formal setting of numerical optimization theory. That is, we view it as a problem of the form: maximize an *objective function* $l(\mathbf{p})$ over the elements $\mathbf{p}$ of a set $\mathbf{P}$. If this is done properly, then the results follow readily from standard optimization results.

5.1.1. *Reformulating the problem.* The key to putting this problem into this framework is to examine (5.1) and recognize that the maximum depends directly on the possible values of the *mixture likelihood vector*

$$\mathbf{L}(Q) = (L_1(Q), L_2(Q), \ldots, L_D(Q))'.$$

We change our perspective on this problem from maximizing the likelihood over all latent distributions Q into the problem of determining which of the eligible classes of mixture likelihood vectors $\mathbf{L}(Q)$ gives the largest value to the likelihood.

We break the formulation into three steps.

STEP 1. Construct the *feasible region* of $\mathscr{R}^D$. It will be the set of all possible fitted values of the likelihood vector:

$$\mathscr{M} = \{\mathbf{L}(Q) = (L_1(Q), \ldots, L_D(Q))' \colon Q \text{ a probability measure}\}.$$

We have already seen sets of this type in Chapter 2.

STEP 2. Define the appropriate *objective function*, here (using the log likelihood)

$$l(\mathbf{p}) := \sum_{j=1}^{K} n_j \ln(p_j),$$

which we wish to maximize over all $\mathbf{p} \in \mathscr{M}$.

Suppose we have found that element of $\mathscr{M}$, say $\hat{\mathbf{L}}$, that maximizes this objective function.

If we were to solve the mixture problem using only these two steps, then there is one more step to carry out:

STEP 3. Solve for the maximum likelihood estimators $\hat{Q}$ by solving from the known $\hat{\mathbf{L}}$ for the latent distribution $\hat{Q}$ via the D equations

$$\mathbf{L}(\hat{Q}) = \hat{\mathbf{L}}.$$

It is instructive to compare this formulation of the problem to the normal theory least squares problem. In the latter one minimizes the objective function $\sum(y_i - \hat{y}_i)^2$ over the feasible set $\mathscr{F}$ that consists of all vectors $\hat{\mathbf{y}}$ of possible fitted values under the model. Corresponding to Step 3, the regression parameters $\hat{\beta}$ can therefrom be determined by solving, from $\hat{\mathbf{y}}$,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}.$$

Much of the theory we derive here merely says that our optimization problem has many of the nice features of the linear regression problem. We examine the key structural features that give us this result.

5.1.2. *The feasible region.* The feasible region $\mathscr{M}$ has two key features that are of importance to us in the optimization problem. First, it is a convex set. As we shall see, this together with the concavity of our objective function, ensures that this is in a class of nice optimization problems.

The second key feature is that the convex feasible region can be expressed as the convex hull of a basic set Γ. In our case, if we define the *unicomponent likelihood vector* to be

$$\mathbf{L}_\phi = (L_1(\phi), \ldots, L_D(\phi))',$$

the mixture likelihood vector for the unicomponent model with parameter ϕ, then

$$\mathbf{L}(Q) = \int \mathbf{L}(\phi)\, dQ(\phi).$$

It follows that if we define the *unicomponent likelihood curve*

$$\Gamma = \mathbf{L}(\phi) \colon \phi \in \Omega\},$$

the desired mixture set is then

$$\mathcal{M} = \text{conv}(\Gamma).$$

The elements of Γ can be thought of as serving as the convex version of a basis, in that we can represent all eligible mixture vectors by convex combinations from this basic set.

We note the convex hull representation distinguishes this problem somewhat from the standard convex optimization problem, in which the convex region is expressed in terms of constraints that are satisfied by elements of the set.

We have already considered convex hull representations of a similar type in Chapter 2. In that chapter we considered the convex hulls of unicomponent density vectors $\mathbf{f}_\phi$, where

$$\sum_t f(t;\phi) = 1.$$

It follows that if the likelihood kernels are multinomial densities $f(t;\phi)$, and $n_t > 0$ for all t, then we can equate the likelihood vector $\mathbf{L}_\phi$ with the density vector $\mathbf{f}_\phi$. Otherwise, even in the multinomial model, the mixture likelihood vectors do not lie in the probability simplex, as we omit components for which $n_t = 0$.

When the likelihood is smoothly parameterized, then Γ is a curve. For example, suppose $f(x;\phi)$ is the Cauchy location density

$$\pi^{-1}[1 + (x - \phi)^2]^{-1}.$$

In Figure 5.1, we show the curve Γ for a pair of observations (y_1, y_2) that are separated by two units, such as $(-1, +1)$, so that the curve has the form

$$\Gamma = \{[1 + (1 - \phi)^2]^{-1}, [1 + (1 + \phi)^2]^{-1} : \phi \in \Omega\}.$$

(We have done the usual simplification of removing constant factors, here π^{-1}, from the likelihood.) The convex hull $\mathcal{M}$ of Γ includes the regions bounded by the dashed lines.

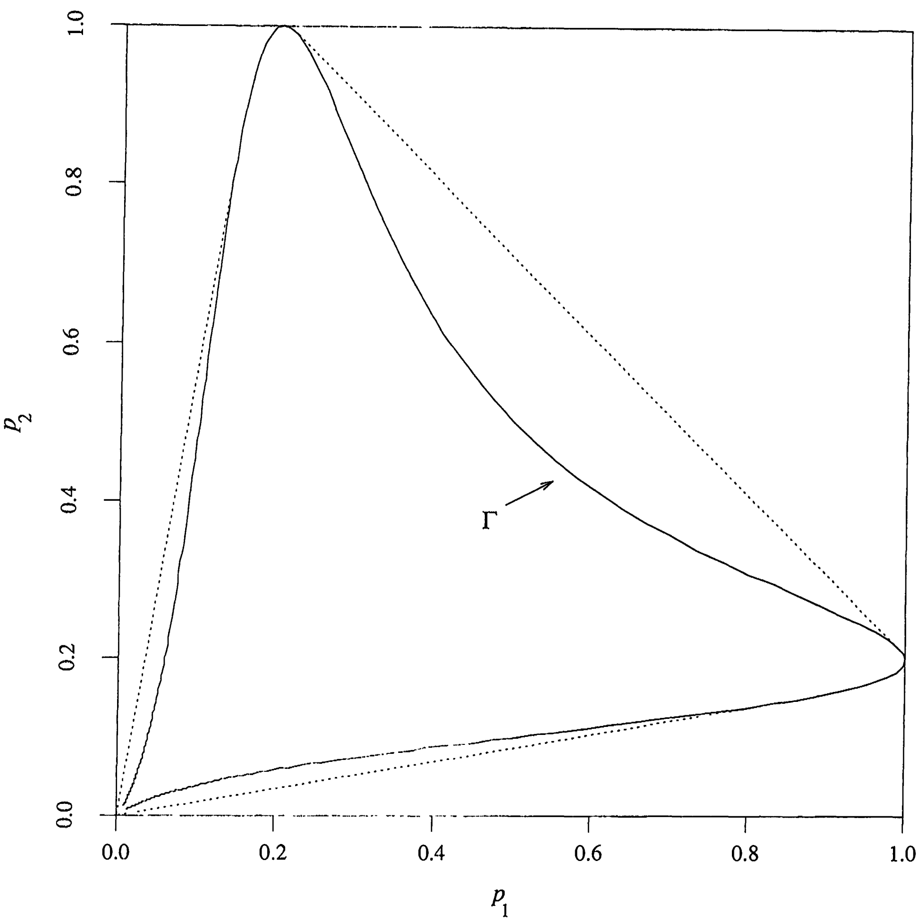

FIG. 5.1. *The unicomponent likelihood curve* Γ *for two Cauchy observations.*

It is not difficult to show that if the set Γ, is closed, then conv(Γ) is closed. This will be of some importance to the existence of the nonparametric MLE, so we offer some comments after stating the theorem.

5.1.3. *The objective function.* The second key feature of our reformulated optimization problem is the concavity of the objective function. Since we are working in a convex set with a convex hull representation, it is natural to examine the properties of the objective function in terms of its behavior along *paths*

$$\mathbf{p}_\varepsilon = [1-\varepsilon]\mathbf{p}_0 + \varepsilon\mathbf{p}_1$$

between pairs $(\mathbf{p}_0, \mathbf{p}_1)$ of elements of the convex set. Restricted to this path, the objective function can be viewed a function of the parameter ε.

We first determine the path derivative of the log likelihood objective function as we go along the path between any two points $\mathbf{p}_0$ and $\mathbf{p}_1$ in the positive orthant:

$$\begin{aligned} d_1(\mathbf{p}_0, \mathbf{p}_1) &:= \frac{d}{d\varepsilon} l([1-\varepsilon]\mathbf{p}_0 + \varepsilon\mathbf{p}_1)\Big|_{\varepsilon=0} \\ &= \sum_i \left[\frac{p_{2i}}{p_{1i}} - 1\right] n_i. \end{aligned} \tag{5.2}$$

The log likelihood objective function is strictly concave along any path:

$$\frac{d^2}{d\varepsilon^2} l([1-\varepsilon]\mathbf{p}_0 + \varepsilon\mathbf{p}_1) = -\sum \frac{(p_{1i} - p_{0i})^2}{p_{\varepsilon i}^2} < 0 \quad \text{if } \mathbf{p}_0 \neq \mathbf{p}_1.$$

It follows that for any $\mathbf{p}_0 \neq \mathbf{p}_1$, we have the *likelihood-gradient inequality*

$$l(\mathbf{p}_1) \leq l(\mathbf{p}_0) + d_1(\mathbf{p}_0, \mathbf{p}_1). \tag{5.3}$$

(This can be proved by creating a first order Taylor expansion in ε about $\varepsilon = 0$ of the likelihood along the path and using the second derivative property to show that the remainder is negative.) This inequality will suffice to prove our fundamental results about the mixture maximum likelihood estimator.

5.2. Basic theorems. We are now ready for the main results, here given more formally than in Chapter 1.

5.2.1. *Existence and support size.*

THEOREM 18. *Suppose that Γ is closed and bounded and that $\mathscr{M}$ contains at least one point with positive likelihood. Then there exists unique $\hat{\mathbf{L}} \in \partial\mathscr{M}$, the boundary of $\mathscr{M}$, such that $\hat{\mathbf{L}}$ maximizes $l(\mathbf{p})$ over $\mathscr{M}$.*

The statement is a slight correction of Lindsay (1983a), which failed to state that if *no point* in $\mathscr{M}$ has positive likelihood, then the uniqueness of the maximum must fail, because then all elements have likelihood zero. This theorem corresponds to Parts 1 and 4 of the mixture NPMLE theorem described in Chapter 1.

The proof is an elementary application of a fundamental result from optimization. We invite the interested reader to consult a general book on convex optimization, such as Roberts and Varberg (1973), to gain further perspective on the following description.

The objective function l is strictly concave on the positive orthant. In particular, this means that the upper sets

$$\mathbf{U}_c = \{\mathbf{p}: l(\mathbf{p}) \geq c\}$$

are closed convex sets. Since Γ is closed and bounded, so is $\mathscr{M}$, and therefore the likelihood objective function l takes on some maximum value at some

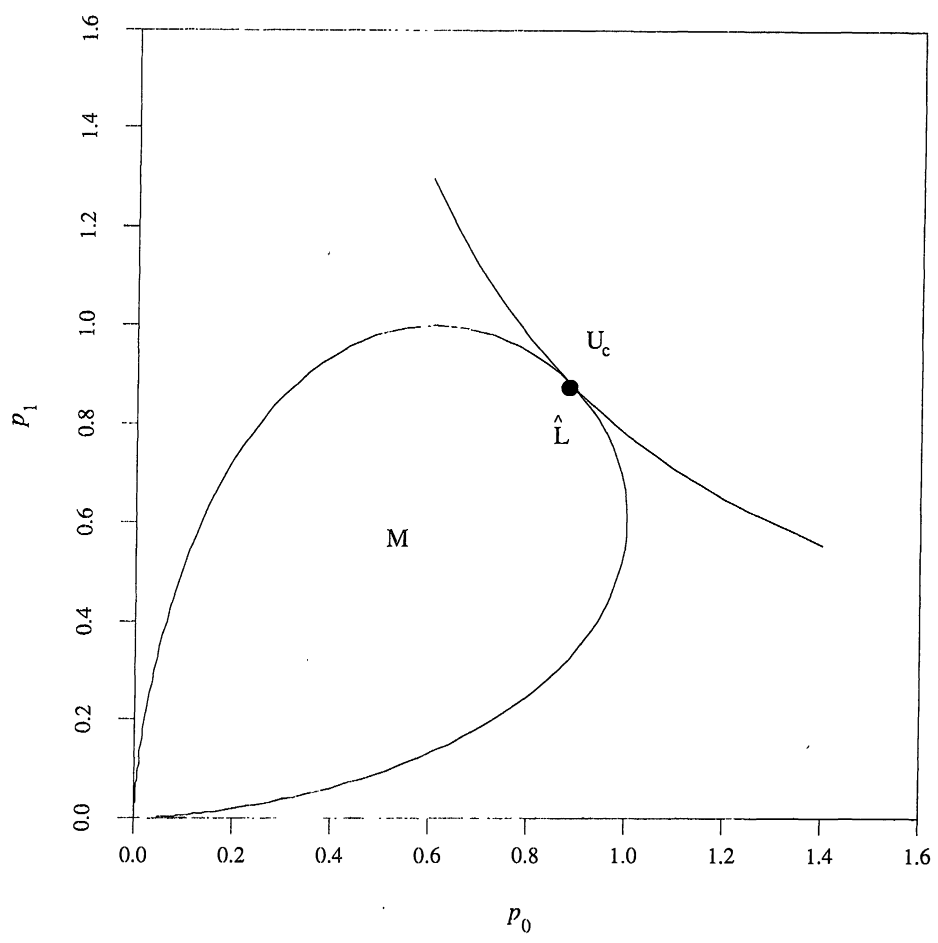

FIG. 5.2. *The geometry of the likelihood maximization problem, showing the unique solution.*

point with strictly positive likelihood. Geometrically, that uniqueness can be seen to correspond to the unique contact point between the upper set $\mathbf{U} = \{\mathbf{p}: l(\mathbf{p}) \geq l(\hat{\mathbf{L}})\}$ and $\mathcal{M}$. (See Figure 5.2.)

5.2.2. *Closed and bounded?* Before proceeding, we wish to address one question about the preceding result that sometimes arises in application. Boundedness of the curve Γ is essential, because if the likelihood vectors have unbounded components, then one can construct unbounded likelihoods. However, the requirement that the set Γ be closed is more of a technical requirement to make the theory simple. I have not found a case where this is a significant issue. To illustrate, we consider some examples.

- There will be cases, such as the mixture of Cauchy densities above, where the parameter ϕ has the range $(-\infty, +\infty)$. To ensure closure of Γ, we must include the left- and right-hand limits; in the Cauchy example, the likelihood vector $\mathbf{L}_\phi$ converges to $\mathbf{0}$, the origin in both directions. We can include this limit point in the set Γ without real consequence because it can never appear in the maximizing mixture. (Otherwise a contradiction arises, since one could eliminate it from the latent distribution and strictly increase the likelihood.)

- Consider next a distribution such as $\text{Bin}(n, p)$, whose boundary parameter values $p = 0$ and $p = 1$ correspond to true distributions. These are limit points of Γ, so that even if we were to set the parameter space as $(0, 1)$, we must necessarily include them in Γ. Since $p = 0$ and $p = 1$ correspond to $\phi = \pm\infty$ in the natural parameterization, we must allow for the possibility of putting mass at ∞ in our estimated latent distribution for the natural parameter unless we prespecify a finite closed range, say $[L, U]$ for ϕ. This comment applies to many contingency table models with log linear modeling, and is relevant in the Rasch model discussion of Lindsay, Clogg and Grego (1991).
- If the likelihood kernel is discontinuous in ϕ, then the set $\mathscr{M}$ may depend on the version of the density function that is used. For example, if f is uniform $(0, \phi)$, then there are two natural versions of the likelihood $L(\phi; x)$ at $\phi = x$, either $1/x$ or 0, depending on whether one chooses right or left continuity: In Figure 5.3 we have plotted such a unicomponent likelihood curve for the case when there are two observations, $x_1 = 1$ and $x_2 = 4$. To make Γ closed, we need to include all the possible limit points, which for $\phi = 4$ means including both $(0.25, 0)$ and $(0.25, 0.25)$. However, even though the closure of Γ appears then to contain two points $\mathbf{L}(\phi)$ corresponding to the same value of ϕ, only one of them is able to play a role in the maximum likelihood solution. This is because mixing using the point on Γ corresponding to using the value $1/x$, here $(0.25, 0.25)$, must necessarily create a strictly greater likelihood than the other value, here $(0.25, 0)$, and so will eliminate the other from being in the final mixture. [*Exercise.*] This remark can clearly be applied to any similar univariate parameter likelihood where the individual components each display a distinct finite set of discontinuities—while in theory we would need to include both right and left limits to apply the theorem, the maximum likelihood estimator will only use the limit point with the larger component values.

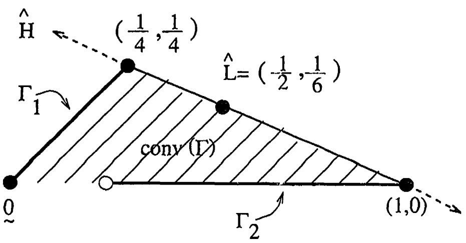

FIG. 5.3. *The uniform likelihood curve, together with its limit points.*

The moral of this story is that, for many problems, as long as one is careful about defining the parameter space to be a closed set and the likelihood vector components are defined to be maximal at discontinuities, there will exist a unique maximum likelihood estimator $\hat{\mathbf{L}}$, even if Γ is not technically closed.

5.2.3. *Gradient characterization.* We now turn to the second part of the mixture NPMLE theorem, the gradient characterization. Recall from Section 1.6 of Chapter 1 the gradient function

$$D_Q(\phi) = \sum_i n_i \left[\frac{L_i(\phi)}{L_i(Q)} - 1 \right].$$

As derived there, this is the path derivative of the log likelihood $\ln(L(Q))$ for the one parameter mixture

$$(1-\pi)Q + \pi\Delta_\phi,$$

evaluated at $\pi = 0$.

We can relate this to our formal optimization theory as follows. From this we see that

$$d_1(\mathbf{L}(Q), \mathbf{L}(\phi)) = D_Q(\phi).$$

That is, the gradient function equals the first derivative of the objective function l along the path from the current fitted vector toward a basic vector. (That the derivatives match up in this way is a consequence of the linear structure of the mixture model). Put into this form, we are then in the setting for classical optimization theory. We can record the basic result in terms of a theorem:

THEOREM 19. *The following three statements are equivalent:*

1. $\hat{Q}$ *maximizes* $l(Q)$.
2. $\hat{Q}$ *minimizes* $\sup_\phi D_Q(\phi)$.
3. $\sup_\phi\{D_{\hat{Q}}(\phi)\} = 0$.

PROOF. If we let $\hat{\mathbf{L}}$ play the role of $\mathbf{p}_0$ in the likelihood-gradient inequality, we see that for $\hat{Q}$ to maximize the likelihood, it suffices that

$$d_1(\hat{\mathbf{L}}, L(Q)) \le 0,$$

for all Q. Therefore, it is sufficient that

$$d_1(\hat{\mathbf{L}}, L(\phi)) \le 0,$$

for all ϕ; hence statement 3 implies 1. Item 1 implies 3 because if the gradient is anywhere positive, we can necessarily increase the likelihood along that path. Finally, $\hat{Q}$ must minimize the sup gradient at the value 0, because if there were Q_1 with supremum less than zero, we could use the likelihood-gradient inequality (5.3) to show Q_1 has greater likelihood than $\hat{Q}$. □

5.2.4. *Properties of the support set.* This completes the proof of the second part of the theorem. Now, the third part:

THEOREM 20. *The support of any maximum likelihood estimator $\hat{Q}$ lies in the set*

$$\{\phi: D_{\hat{Q}}(\phi) = 0\}.$$

PROOF. Consider the one parameter family of mixtures

$$Q_\varepsilon := (1 - \varepsilon)\hat{Q} + \varepsilon\Delta_\phi.$$

If ϕ is a support point of $\hat{Q}$, then Q_ε continues to be a true probability measure for some negative values of ε. It follows that the maximum value is taken on an interior point of the allowable range of ε. This implies that the derivative of the likelihood along the one parameter path equals zero at this point; however, this derivative is just the gradient function $D_{\hat{Q}}(\phi)$. □

These results have very simple geometric interpretations. For any candidate mixture likelihood vector $\mathbf{L}$ in $\mathscr{M}$, the gradient function determines a hyperplane

$$\mathscr{H} := \{\mathbf{p}: d_1(\mathbf{L}, \mathbf{p}) = 0\}$$

that contains the point $\mathbf{L}$. If $\mathbf{L}$ is indeed $\hat{\mathbf{L}}$, the maximum likelihood point, then this hyperplane $\mathscr{H}$ is a support hyperplane to the set $\mathscr{M}$ and separates that convex set from the convex upper set of the log likelihood objective function $\mathbf{U} = \{\mathbf{p}: l(\mathbf{p}) \geq l(\hat{\mathbf{L}})\}$. Lying exactly in the support hyperplane are all the support vectors $\mathbf{L}_{\phi_i}$. See Figure 5.4. This interpretation allows us to apply Carathéodory's theorem (Section 2.3.4) to characterize the existence of a discrete latent distribution that maximizes the likelihood:

THEOREM 21. *The solution $\hat{\mathbf{L}}$ can be represented as $\mathbf{L}(\hat{Q})$, where $\hat{Q}$ has no more than D points of support.*

We note that if we are in the setting of the multinomial exponential family of Chapter 2, we can employ the superior bounds on the mixture representations that were given there, with a bound of roughly $D/2$. (See the discussion of index in Chapter 2 for more precise descriptions.) However, we emphatically note that it was *absolutely critical* that the mixture likelihood vectors lie in the probability simplex for this reduction to take place. For example, in a normal mixture model with fixed σ^2, one can construct sets of data for which the bound D is attained simply by spreading the observations widely apart.

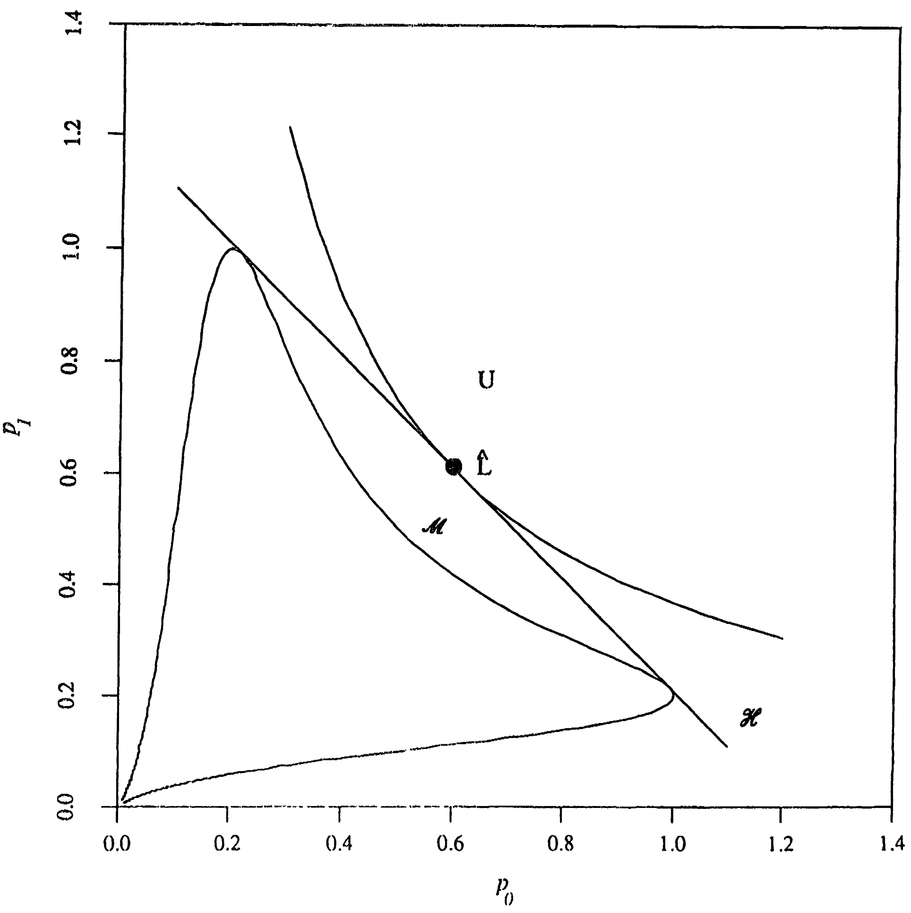

FIG. 5.4. *The support hyperplane of the maximum likelihood point.*

5.3. Further implications of the theorems.

5.3.1. *Duality theorem.* We specify two versions of the dual problem, with the understanding that they are equivalent, but the first one arises naturally in the context of the problem, whereas the second is in the form of a classical optimization problem:

DUAL 1. Minimize $l(\mathbf{p})$ subject to the constraints $\mathbf{p} \geq \mathbf{0}$, and $d_1(\mathbf{p}, \mathbf{L}(\phi)) \leq 0$, for all $\phi \in \Omega$.

DUAL 2. Maximize $l(\mathbf{w})$ subject to the constraints $\mathbf{w} \geq \mathbf{0}$ and $\sum w_i L_i(\phi) \leq 0$, for all $\phi \in \Omega$.

THEOREM 22. *If the mixture MLE solution is $\hat{\mathbf{L}}$, then $\mathbf{p} = \hat{\mathbf{L}}$ solves Dual 1 and $\hat{w}_i = n_i/\hat{L}_i$ solves Dual 2.*

PROOF. First, $\hat{\mathbf{L}}$ satisfies the constraints of Dual 1 by our gradient characterization theorem, so it is a feasible solution. Next, the fact that $\hat{\mathbf{L}}$ is in the mixture set and the definition of the constraint set imply that for any other feasible point $\mathbf{p}$, $d_1(\mathbf{p}, \hat{\mathbf{L}}) \leq 0$. The likelihood-gradient inequality (5.3) therefore implies that

$$l(\hat{\mathbf{L}}) \leq l(\mathbf{p}) + d(\mathbf{p}, \hat{\mathbf{L}}) \leq l(\mathbf{p}),$$

for any other feasible $\mathbf{p}$, as was to be shown. Dual 2 can then be solved simply by a change in variables in Dual 1, namely, $w_i := n_i/p_i$. □

We now note that Dual 2 has the form of a classic optimization problem: minimize a convex objective function $(-l)$ over a region described by linear inequality constraints. The solution to this can therefore be found by using standard optimization programs, with the possible limitation that there are *infinitely* many linear constraints whenever there are infinitely many ϕ, a subject we must deal with later.

5.3.2. *Gradient bounds on the likelihood.* The result that follows shows that if we compute the log likelihood at a candidate estimator Q_c, getting $l(\mathbf{L}_{Q_c})$, then we can determine from the gradient function $D_{Q_c}(\phi)$ not only if we have the solution, but also both upper and lower bounds on the maximum value of the likelihood.

THEOREM 23. *Let Q_c be the current mixing distribution in an iterative algorithm designed to find the maximum likelihood estimator. Define $\delta = \sup_\phi D_{Q_c}(\phi)$. Then*

$$A(\delta) \leq l(\hat{\mathbf{L}}) - l(L_{Q_c}) \leq B(\delta) \leq \delta,$$

where $B(\delta) := n \ln(1 + \delta/n)$, $A(\delta) := B(\delta) - n^* \ln[1 + \delta/n^*]$ and $n^* = n - \min_k\{n_i\}$.

PROOF. For the lower bound, see the not-so-simple argument in Lindsay (1983b). For the upper bound, we can use Dual 2. The point

$$\left[\frac{n}{n+\delta}\right]\mathbf{L}_{Q_c}$$

meets the linear constraints of the optimization problem, so it is a feasible solution. It follows that

$$l\left(\left[\frac{n}{n+\delta}\right]\mathbf{L}_{Q_c}\right) \leq l(\hat{\mathbf{w}}),$$

which gives the desired upper bound. □

In any algorithm, it is useful to have a way to determine how much more computation might be needed to converge to the solution. The preceding result shows that the maximum of the gradient function can be used for this. We will consider convergence criteria for algorithms further in the next chapter.

5.3.3. *Link to m-component methods.* The preceding results enable us to contrast the properties of the *global* maximum likelihood estimator of the latent distribution $\hat{Q}$, and estimators formed by maximizing the likelihood, or solving the likelihood equations, when the latent distribution is restricted to having a fixed number of support points. In this section, we extend the results of Section 3.3.

Let $\hat{Q}_m$ be a latent distribution, such as described in Chapter 3, that maximizes the m-point mixture likelihood. Earlier we derived part of the relationship between the gradient function and the EM algorithm that is often used to find $\hat{Q}_m$. That is, we noted that the EM algorithm for the weights can be written as

$$\hat{\pi}_{j,\text{new}} = \hat{\pi}_{j,\text{old}}[1 + n^{-1}D_Q(\phi)],$$

so that the weights increase or decrease according to the sign and magnitude of the gradient function. In addition, it is easy to check that the new support points move left or right from the old ones in agreement with the direction of the greater gradient. [That is, if $D'_Q(\phi)$ is positive at a support point ϕ, then the EM algorithm puts the new point to the right.]

Next, the following basic results are from Lindsay (1981):

THEOREM 24. *Suppose ϕ^* is a support point of $\hat{Q}_m$. If the gradient function is twice differentiable in ϕ at ϕ^*, then:*

1. $D_{\hat{Q}_m}(\phi^*) = 0.$
2. $D'_{\hat{Q}_m}(\phi^*) = 0.$
3. $D''_{\hat{Q}_m}(\phi^*) \leq \sum n_k[(L_k(\phi^*)/(L_k(\tilde{Q}_m))]^2.$

PROOF. [*Exercise.*] These arise in a straightforward way through manipulation of the likelihood equations and the formula for the gradient. □

The global MLE is, of course, an m-point MLE as well, for some m, so that it satisfies all three of the above properties. However, it satisfies a stronger property than statement 3 because each support point is at a local maximum of the gradient and so

$$D''_{\hat{Q}}(\phi^*) \leq 0.$$

5.3.4. *Moment and support point properties.* We have already mentioned some properties of the nonparametric MLE in Section 5.2.4. We now mention a few further results that can be found in Lindsay (1981). One of the more important results is relevant in the search of the parameter space Ω for potential support points.

PROPOSITION 25. *Suppose that the parameter ϕ is real-valued and that for every i, the likelihood kernel $L_i(\phi)$ is unimodal in ϕ, with unique mode at $\tilde{\phi}_i$. Then all the support points of $\hat{Q}$ lie in the interval*

$$[\min_i \tilde{\phi}_i, \max_i \tilde{\phi}_i].$$

PROOF. [*Exercise.*] Show that the gradient function is increasing whenever $\phi \leq \min_i \tilde{\phi}_i$ and decreasing for $\phi \geq \max_i \tilde{\phi}_i$. Since the latent support points are in the set of maxima to the gradient, this proves the result. □

The next result relates to the dispersion score v_2 of Neyman and Scott that was introduced in Chapter 4. We recall from Section 4.1 that if the one-component solution is also the NPMLE, then it is necessarily true that

$$\sum v_2(\hat{\phi}, x_i) \leq 0,$$

because this score is also $D_1''(\hat{\phi})$ and we would otherwise have a local failure of the gradient inequality. We can extend this result to any of the support points of $\hat{Q}$, say ϕ^*, by recalling the second derivative inequality, $D_{\hat{Q}}''(\phi^*) \leq 0$. With some further manipulation, we can relate this again to the dispersion score in that it is equivalent to requiring that

$$\sum E[v_2(\Phi, x_i)h(\Phi) \mid X = x_i;\ \hat{Q}] \leq 0,$$

for every nonnegative function $h(\phi)$. In the one parameter exponential family, (2.2), one can use the last equation to show that the sample variance is always smaller than the estimated variance under the model:

$$n^{-1} \sum (x_i - \bar{x})^2 \leq \mathrm{Var}(X; \hat{Q}).$$

One can strengthen this dispersion result—that the fitted model is biased in the direction of overdispersion—even further by using the full force of the gradient inequality. The following is a challenging exercise: Let $\xi_1, \ldots, \xi_m$ be the support set for $\hat{Q}$ in a one parameter exponential family, in the natural parameter. For every value of t such that $t+\xi_i$ is in the parameter space for all i, the following inequality of sample and model moment generating functions holds:

$$n^{-1} \sum e^{tx_i} \leq \int e^{tx}\, dF(x; \hat{Q}).$$

This result is not in Lindsay (1981), but can be proved by using the gradient inequality and the fact that the exponential family density has exponential form. Differentiating twice with respect to t leads to the variance inequality above.

5.4. Applications. At this point we offer some simple examples that may help elucidate these mathematical structures and methods.

5.4.1. *A binomial mixture.* First, we have some plots that illustrate how one can use the gradient function. In Figure 5.5 we have shown a sequence of plots for a particular sequential optimization scheme. The data are the sibship data of Chapter 2, and we use the binomial mixture model discussed there. In the algorithm that was used, the nonparametric mixture estimator was found by first finding the best *one* support point model, then the best *two* and so on, until we find that we can no longer increase the likelihood by adding support points; that is, the gradient inequality is satisfied. One can use the EM algorithm with a fixed number of support points, iterate until convergence and then check the gradient function. If it is greater than zero at some point ϕ, then one can add that support point to the rest in a way that increases the likelihood and then return to the EM, but with one larger support size.

The first plot shows the best two support point model. In accordance with the above discussion, we see that the two support points show up as zeros and

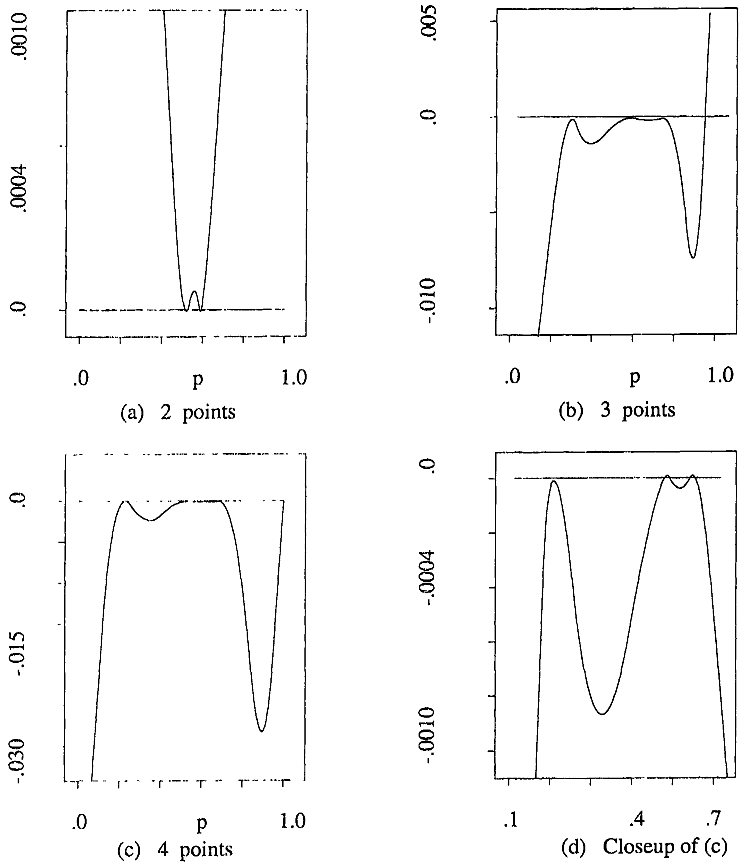

FIG. 5.5. *The sequence of gradient plots for the sibship data.*

the derivative of the gradient is zero there. However, the gradient inequality is certainly violated, and indeed the support points are local minima to the gradient. Next, the three point fit also violates the gradient inequality and we can see that the likelihood can be increased by adding mass near $p = 1$. When this is done, we achieve the gradient inequality, as can be seen in the lower two frames.

5.4.2. *Empirical CDF.* In Figure 5.6 we show the rather simple geometry that arises for the nonparametric distribution function problem of Section 1.7.1. In this case, the likelihood kernel was $L_i(\phi) = \mathscr{I}[\phi = x_i]$. Thus for a sample of size two, the likelihood "curve" Γ consists of three points only: $(1, 0)$ at $\phi = x_1$, $(0, 1)$ at $\phi = x_2$ and $(0, 0)$ for any other value of ϕ. As noted earlier, the nonparametric maximum likelihood estimator corresponds to the sample proportions at the two observed values, a point on the simplex determined by the convex hull of these three points.

We can use this problem to illustrate the dual problem approach. Here the constraints of the dual problem become

$$\sum w_i L_i(\phi) \leq n \qquad \forall\phi \Leftrightarrow w_i \leq n.$$

The problem is to maximize $\sum n_i \ln(w_i)$ subject to these constraints. The solution is obvious: Set w_i equal to its maximal value n. This in turn implies that the solution to the primal problem is $\hat{L}_i = n_i/n$. Thus turning to the dual problem simplifies the optimization problem to a triviality.

5.4.3. *Known component distributions.* We return briefly to the case of the known component densities (introduced in Section 1.3.1). Nonparametric maximum likelihood in this case is just a special case of what we have described, in

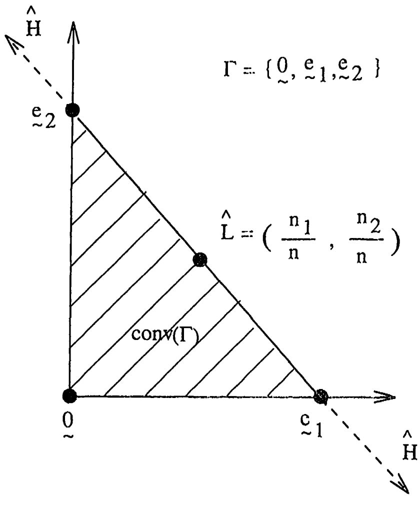
$\hat{H}$
$\Gamma = \{\underset{\sim}{0}, \underset{\sim}{e}_1, \underset{\sim}{e}_2\}$
$\underset{\sim}{e}_2$
$\hat{L} = (\frac{n_1}{n}, \frac{n_2}{n})$
conv(Γ)
$\underset{\sim}{0}$
$\underset{\sim}{e}_1$
$\hat{H}$

FIG. 5.6. *The likelihood curve for the nonparametric distribution problem.*

which the latent variable ϕ is just the index of the known component density. Roeder, Devlin and Lindsay (1989) give a relatively complete description of the implications of the theorem in this setting. In addition, see Vardi and Lee (1993), where one can find references to literature in areas other than mixture models in which pieces of the nonparametric maximum likelihood theory have been developed.

5.4.4. *The multinomial case.* We recall the multinomial geometry discussed in Section 2.3. One nice feature of the multinomial problem is that one can picture the behavior of the nonparametric maximum likelihood estimator $\hat{\mathbf{L}}$ as a function of the sample frequencies vector, denoted by $\mathbf{d}$ in Chapter 4. For example, it is clear that if $\mathbf{d}$ falls in the set of eligible mixture density vectors, then $\hat{\mathbf{L}} = \mathbf{d}$ (since $\mathbf{d}$ always globally maximizes the likelihood objective function l over the entire probability simplex and by assumption is also a mixture vector). This property was used in Lindsay, Clogg and Grego (1991) to show that the nonparametric mixture approach was equivalent to a conditional approach for estimating certain auxiliary regression parameters.

If $\mathbf{d}$ is not in the set of mixture vectors, then one can partition the probability simplex into regions corresponding to the number of components in the maximum likelihood solution. The reader is invited to consider how this might be done using the examples of Chapter 2.

5.5. Uniqueness and support size results. The final problem we address is the uniqueness of the estimator of the latent distribution, saved for last due to its technical difficulty. We will here just describe the basic issues and refer the reader desirous of more details to other sources.

Thus, to date, we have shown the uniqueness of the estimated mixture likelihood vector $\hat{\mathbf{L}}$. Can we infer from this the uniqueness of the latent distribution estimator $\hat{Q}$? That is, can we infer from $\hat{\mathbf{L}}$ the latent distribution itself by solving for $\hat{Q}$ in

$$\hat{\mathbf{L}} = \int \mathbf{L}(\phi)\, dQ(\phi).$$

From a geometric point of view, this may seem unlikely because the interior of the convex hull of a curve, such as $\mathscr{M} = \operatorname{conv}(\Gamma)$, generally has infinitely many representations in terms of elements of its generating set Γ. (We saw this in Chapter 2.) What saves the day for us is that the solution $\hat{\mathbf{L}}$ is on the boundary of $\mathscr{M}$. We describe a strategy for proving uniqueness that works for many important examples.

5.5.1. *The strategy.* We first note that the gradient function at the maximum likelihood estimator is completely determined by $\hat{\mathbf{L}}$ and is not related to the choice of $\hat{Q}$. Thus from $\hat{\mathbf{L}}$ we can unambiguously determine the full set of points $\{\phi_1, \ldots, \phi_r\}$ that satisfy

$$D_{\hat{Q}}(\phi) = 0 \tag{5.4}$$

and so are candidates to be in the support set of $\hat{Q}$.

TASK 1. Show that (5.4) has at most $r \leq D$ solutions, subject to $D_{\hat{Q}}(\phi) \leq 0$; that is, the solutions must also be local maxima. The proof of this property must depend on the gradient function having a polynomial type structure that bounds the number of solutions to such equations. [Although this is related to the ideas of Chebyshev systems (Chapter 2), those ideas are not strong enough to work here.]

If Task 1 is completed, the set of possible support points is fixed at $\{\phi_1, \ldots, \phi_r\}$ and we proceed to:

TASK 2. We must now see if the weights are uniquely determined by the equations

$$\sum \pi_j \mathbf{L}(\phi_j) = \hat{\mathbf{L}}.$$

This is just a set of linear equations in π_j, so it suffices to show that every set of D unicomponent likelihood vectors $\{\mathbf{L}(\phi_j)\}$ is linearly independent. This is simpler than Task 1, in that we can apply the ideas of Chebyshev systems directly to obtain this result, as discussed in Chapter 2.

5.5.2. *A geometric approach to Task 1.* The first proofs of uniqueness used special properties of the likelihood kernel involved to complete Task 1 [for the Poisson model, see Simar (1976); for the exponential, see Jewell (1982)]. In essence, in these cases, the gradient could be written as an *exponential polynomial*, and certain long known bounds on the number of zeros to these polynomials could be used [Pólya and Szegö (1925)]. However, this was a piecemeal approach to the problem, when clearly there were more general truths at work.

In Lindsay (1983a), the problem was attacked from a geometric point of view for a unicomponent exponential family and it was shown that the number of zeros to the gradient function could be bounded by considering the geometric structure of the unicomponent likelihood curve. In effect, there was a way to consider the complexity of the curve Γ that gave a bound on the number of support points; the bound depended on the number of zeros to a certain polynomial, but was always less than D.

These results are easy to visualize for the case of two observations. In Figure 5.7 we have shown how, in the case of the normal mixture, with scale parameter 1, the shape of the likelihood curve $\mathbf{L}(\phi)$ depends very strongly on the distance between the two observations. If the observations are 1 unit apart, say $x_1 = -0.5$, $x_2 = 0.5$, we get the "balloon" shape of $\boldsymbol{\Gamma}_1$, enclosing a convex region. It is clear that the NPML estimator has one support point. If the observations are 3 units apart, shown as $\boldsymbol{\Gamma}_3$, the curve has a substantial indentation and the corresponding latent distribution estimator has two support points. The boundary case occurs when the observations are exactly 2 units apart, shown as $\boldsymbol{\Gamma}_2$. In the two-dimensional case, these results can be obtained by analyzing the sign of the curvature of the likelihood curve. Unfortunately, these results were very difficult to obtain in higher dimensions, and hard to generalize outside the exponential family.

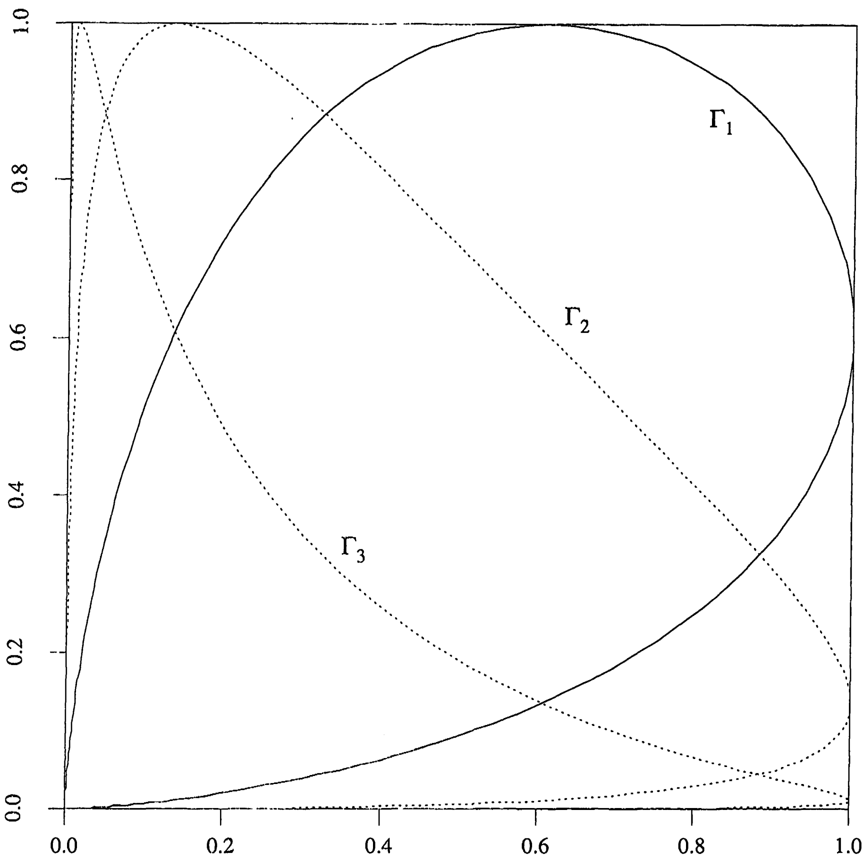

FIG. 5.7. *Three different unicomponent likelihood curves for the normal model.*

5.5.3. *A gradient function representation.* Therefore, we examine a great simplification of the proof. Since it relies on a powerful result from the theory of totally positive kernels, we merely illustrate the general method of attack here without pretending to be complete. For more, see Lindsay and Roeder (1993).

The key to the result is deriving a representation of the gradient function of the form

$$n^{-1}D_Q(\phi) = \int \left[\frac{f(x;\phi)}{f(x;Q)}\right] [h_1(x) - h_0(x)]\, d\lambda(x), \tag{5.5}$$

which we now construct. We start with the easy-to-prove representation

$$n^{-1}D_Q(\phi) = \int \left[\frac{f(x;\phi)}{f(x;Q)}\right] d[\hat{F}(x) - F_Q(x)]$$

Next, define the positive measure λ by the relationship

$$d\lambda(x) = d[\hat{F} + F_Q](x).$$

(If we were to divide by 2, we would have a probability measure representing the mixture of the empirical CDF and the mixture distribution under Q.) The key here is that λ is a measure that dominates (in the measure-theoretic sense) both the empirical CDF $\hat{F}$ and the model distribution F_Q, regardless of whether the latter is discrete or continuous.

The Radon–Nikodyn theorem therefore implies that we can write density functions h_1 and h_0, with respect to the measure $d\lambda(x)$, for $\hat{F}$ and F_Q, and that the representation (5.5) holds. We can calculate the densities explicitly: Suppose that $\hat{F}$ has mass $F(\{y_k\})$ at a set of observed data points $y_1, \dots, y_K$. (In our case, most likely $K = D$, the number of distinct factors in the likelihood.) Then we can write the Radon–Nikodyn derivatives as

$$h_1(x) = \frac{d\hat{F}(x)}{d\lambda(x)} = \begin{cases} \dfrac{\hat{F}(\{x\})}{\hat{F}(\{x\}) + F_Q(\{x\})}, & x \in \{y_1, \dots, y_K\}, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$h_0(x) = \frac{dF_Q(x)}{d\lambda(x)} = \begin{cases} \dfrac{F_Q(\{x\})}{\hat{F}(\{x\}) + F_Q(\{x\})}, & x \in \{y_1, \dots, y_K\}, \\ 1 & \text{otherwise.} \end{cases}$$

Returning to the gradient representation (5.5), we next recognize that it has the form

$$C(\phi) = \int A(x;\phi)B(x)\, d\lambda(x).$$

We next apply some powerful results from Karlin (1968). If $A(x;\phi)$ is a *strictly totally positive kernel*, λ is a positive measure and the function $B(x)$ is nonzero and has no more than M sign changes (relative to the measure λ), then the *variation diminishing property* of the totally positive kernel implies that $C(\phi)$ has no more than M sign changes unless it is identically zero.

Applying this to our case, we see that the difference in densities $B(x) = h_1(x) - h_0(x)$ can have at most $2K$ sign changes, one to each side of a data point. (It is negative between observations, but possibly positive at each observation.) It follows that if the family of densities is based on a totally positive kernel, then either the gradient will be identically zero or it will have at most K local maxima.

Lindsay and Roeder (1993) used this fact to show that in the exponential family, regardless of whether it is discrete or continuous, then either the latent estimator $\hat{Q}$ is unique *or* the gradient function is identically zero at the maximum likelihood solution and the latent estimator is nonunique. The latter can happen only with discrete f for which some mixtures are nonidentifiable, as in Chapter 2.