CHAPTER 6. THE DUAL TO THE MAXIMUM LIKELIHOOD ESTIMATOR

KULLBACK-LEIBLER INFORMATION (ENTROPY)

Before turning to the dual of the maximum likelihood estimator we define the Kullback-Leibler information, and prove a few of its simple properties. The goal of this detour is to provide a natural probabilistic interpretation for this dual as the minimum entropy expectation parameter.

6.1 Definitions

Suppose F, G are two probability distributions with densities f, g relative to some dominating σ-finite measure ν. The *Kullback-Leibler information* of G at F is

$$(1) \qquad K(F, G) = E_F(\ln(f(x)/g(x)))$$

with the convention that $\infty \cdot 0 = 0$, $0/0 = 1$, and $y/0 = \infty$ for $y > 0$. K is als referred to as the *entropy* of G at F.

It can easily be verified that K(F, G) is independent of the choice of dominating measure ν. The existence of K will be established in Lemma 6.2 where it is shown that $0 \leq K \leq \infty$.

In exponential families it is convenient to write

$$(2) \qquad K(\theta_0, \theta_1) = K(P_{\theta_0}, P_{\theta_1}) , \qquad \theta_0, \theta_1 \in N \quad .$$

For $S \subset N$ let

$$(3) \qquad K(S, \theta_1) = \inf\{K(\theta_0, \theta_1): \theta_0 \in S\} \quad ,$$

etc.

$K(\cdot,\cdot)$ as defined in (2) has domain $N \times N$. It is convenient to also transfer this definition to the expectation parameter space. Accordingly, define $\tilde{K}(\xi_0, \xi_1)$ by

$$(4) \qquad \tilde{K}(\xi_0, \xi_1) = K(\theta(\xi_0), \theta(\xi_1))$$

for $(\xi_0, \xi_1) \in \xi(N^\circ) \times \xi(N^\circ)$. If the family is steep this definition is valid on $K^\circ \times K^\circ$.

It is also sometimes convenient to extend the definition of $\tilde{K}(\cdot, \xi_1)$ to all of R^K, by lower semicontinuity. Accordingly, for a minimal steep family, and for $\xi_0 \in \bar{K} - K^\circ$, $\xi_1 \in K^\circ$, define

$$(5) \qquad \tilde{K}(\xi_0, \xi_1) = \lim_{\varepsilon \downarrow 0} \inf\{\tilde{K}(\xi, \xi_1): \ \xi \in K^\circ, \ ||\xi - \xi_0|| < \varepsilon\} \quad .$$

For $\xi \notin \bar{K}$, $\xi_1 \in K^\circ$ define

$$(6) \qquad \tilde{K}(\xi, \xi_1) = \infty \quad .$$

It is to be emphasized that this is a formal, analytic extension of the definition. $\tilde{K}(\xi_0, \xi_1)$ for $\xi_0 \notin K^\circ$ does not necessarily have a probabilistic interpretation like (1). (Sections 6.18+ give a probabilistic interpretation of $\tilde{K}$, valid under some auxiliary conditions.)

K is often called the Kullback-Leibler "distance" from θ_0 to θ_1, but it is not a metric in the topological sense. In particular, it is -- in general -- not symmetric. There is, however, one very important special case where K is symmetric and $(K)^{\frac{1}{2}}$ is a metric: the normal location family, $\{P_\theta\} = \{\Phi_{\theta,\Sigma}: \theta \in R^k$, forms a standard exponential family with canonical statistic $\Sigma^{-1}x$ (see Example 1.14), and has

$$(7) \qquad K(\theta_0, \theta_1) = (\theta_1 - \theta_0)'\Sigma^{-1}(\theta_1 - \theta_0)/2 \quad .$$

The following proposition has already been mentioned above.

6.2 Proposition

For any two distributions K(F, G) exists and satisfies

$$0 \le K(F, G) \le \infty \qquad . \tag{1}$$

K(F, G) = 0 if and only if F = G.

Proof.

$$\begin{aligned} E_F(\ln(f(X)/g(X))) &= E_F(-\ln(g(X)/f(X))) \\ &\ge -\ln E_F(g(X)/f(X)) \\ &= -\ln 1 = 0 \end{aligned}$$

by Jensen's inequality, with equality if and only if $f = g$ a.e.(ν). ||

For exponential families K has an especially simple and appealing form.

6.3 Proposition

Let $\{p_\theta\}$ be a standard exponential family. If $\theta_0 \in N^\circ$, $\theta_1 \in N$ then

$$\begin{aligned} K(\theta_0, \theta_1) &= (\theta_0 - \theta_1) \cdot \xi(\theta_0) - (\psi(\theta_0) - \psi(\theta_1)) \\ &= \log (p_{\theta_0}(\xi(\theta_0))/p_{\theta_1}(\xi(\theta_0))) \qquad . \end{aligned} \tag{1}$$

(*Remark.* Suppose $\{p_\theta\}$ is steep and $\theta_0 \in N - N^\circ$, $\theta_1 \in N^\circ$. Then $K(\theta_0, \theta_1) = \infty = \lim_{\eta_i \to \theta_0} K(\eta, \theta_1)$ for $\{\eta_i\} \subset N^\circ$ by steepness. Since the only sensible interpretation for $(\theta_0 - \theta_1) \cdot \xi(\theta_0)$ is ∞ here, (1) may be considered valid for all $\theta_0 \in N$ for regular or steep families.)

Proof. Note that

$$\ln(p_{\theta_1}(x)/p_{\theta_0}(x)) = (\theta_1 - \theta_0) \cdot x - (\psi(\theta_1) - \psi(\theta_0))$$

and $E_{\theta_0}(X) = \xi(\theta_0)$. ||

6.4 Remark

The second part of 6.3(1) shows how the Kullback-Leibler information is related to maximum likelihood estimation. For $S \subset N$ let

$$K(\theta_0, S) = \inf\{K(\theta_0, \theta_1): \theta_1 \in S\} \quad . \tag{1}$$

Then, by 6.3(1), if $\theta_0 \in N^\circ$

$$K(\theta_0, S) = K(\theta_0, \theta) \tag{2}$$

for $\theta \in S$ if and only if $\theta \in \hat{\theta}_S(\xi(\theta_0))$.

In other words, for steep families, for $\Theta = S$, and for an observation $x \in K^\circ$ the maximum likelihood estimator is the closest point in S to $\theta(x)$ in the Kullback-Leibler sense. (For observations $x \in K - K^\circ$ such an interpretation requires an extension of the definition of K like that to be provided in Sections 6.18+.)

Note also that

$$K(\theta_0, \theta_1) = \ell(\theta_0, \xi(\theta_0)) - \ell(\theta_1, \xi(\theta_0)) \quad . \tag{3}$$

The fact that the quantity on the right is positive (for $\theta_0 \in N^\circ$, $\theta_1 \neq \theta_0$) has already been used in 5.8(3) and 5.12(3).

6.5 Theorem

Let $\{p_\theta\}$ be a standard exponential family. Then $K(\cdot,\cdot)$ is infinitely differentiable on $N^\circ \times N^\circ$. On N°

$$\nabla K(\theta_0, \cdot) = \xi(\cdot) - \xi(\theta_0) \tag{1}$$

$$D_2K(\theta_0, \cdot) = D_2\psi(\cdot) = \Sigma(\cdot) \ , \qquad \theta_0 \in N^\circ \tag{2}$$

If $\{p_\theta\}$ is minimal and steep then on K°

$$\nabla\tilde{K}(\cdot, \xi_1) = \theta(\cdot) - \theta(\xi_1) \tag{3}$$

$$(4)\qquad D_2\tilde{K}(\cdot,\ \xi_1) = \not{\Sigma}^{-1}(\theta(\cdot))\ , \qquad \xi_1 \in K^\circ \quad .$$

Consequently, given $\xi_1 \in K^\circ$ and $\varepsilon_1 > 0$ there is an $\varepsilon_2 > 0$ such that

$$(5)\qquad \tilde{K}(\xi,\ \xi_1) \geq \varepsilon_2||\xi - \xi_1|| \quad \text{whenever} \quad ||\xi - \xi_1|| > \varepsilon_1 \quad .$$

If $S \subset K^\circ$ is compact then a value $\varepsilon_2 > 0$ can be chosen so that (5) is valid uniformly for all $\xi_1 \in S$.

Proof. Formulae (1) - (3) are straightforward from 6.3(1). (Note also that (1), (2) are merely a restatement of 5.3(1), (2).) (4) follows from (3) by the inverse function theorem since $\theta(\cdot) = \xi^{-1}(\cdot)$ and $\nabla\xi(\cdot) = \Sigma(\cdot)$. Formula (5) follows from (3), (4) as did the analogous conclusion 5.3(3), and 5.3(5) of Lemma 5.3 follow from 5.3(1), (2). The asserted uniformity of (5) over $\xi_1 \in S$ is easy to check in that proof. ||

(Note: if p_θ is not minimal 6.5(3) is still valid and 6.5(4) is valid with $\not{\Sigma}^{-1}$ interpreted as a generalized inverse.)

CONVEX DUALITY

6.6 Definition

Let ϕ: $R^k \to (-\infty,\infty]$ be convex. The *convex dual* of ϕ is the function d_ϕ: $R^k \to [-\infty,\ \infty]$ defined by

$$(1)\qquad d_\phi(x) = \sup\{\ell_\phi(\theta,\ x):\ \theta \in R^k\} \quad .$$

(Recall, $\ell_\phi(\theta,\ x) = \theta \cdot x - \phi(\theta)$.)

We will be interested in the situation when ϕ is regularly strictly convex and steep. (See Definition 5.2.) Then if $x \in R = \xi(N_\phi^\circ)$, $\ell(\cdot,\ x)$ is strictly concave on N_ϕ and $\nabla\ell(\cdot,\ x)_{|\theta(x)} = 0$. Thus

$$(2)\qquad d_\phi(x) = \ell_\phi(\theta(x),\ x) \quad \text{for} \quad x \in R = \xi(N_\phi^\circ) \quad .$$

(In such cases, and somewhat more generally, the pair $(d_\phi,\ R\ )$ is called the

Legendre transform of (ϕ, N_ϕ). It is easy to check from (2) and Theorem 6.5 that

$$(3) \qquad d_{d_\phi}(\theta) = \phi(\theta) \quad \text{for} \quad \theta \in N^\circ \quad .$$

It can be shown that (3) actually holds for all $\theta \in R^k$, but we do not need this fact in what follows.)

Suppose ψ is the cumulant generating function of a steep exponential family. Then

$$(4) \qquad d_\psi(x_0) = \tilde{K}(x_0, x_1) + \theta(x_1) \cdot x_0 - \psi(\theta(x_1)) \; , \qquad x_0 \in K^\circ \quad .$$

If the coordinate system and dominating measure are chosen so that $\psi(0) = 0 = \xi(0)$ then (4) becomes

$$(4') \qquad d_\psi(x_0) = \tilde{K}(x_0, 0) \qquad x \in K^\circ \quad .$$

This provides a probabilistic interpretation for $d(x)$ on K°. It will be seen later that $d(\cdot)$ is the maximal lower semicontinuous extension of $(d(x): \; x \in K^\circ)$ to all of R^k, and (4) is valid for all $x_0 \in R^k$.

Lemmas 6.7 and 6.8 and Theorem 6.9 present some important basic facts about convex duality. They are just the tip of a rich theory. We will not further develop this theory as an abstract unit; although other important features of the theory are implict in results we state elsewhere (e.g. Theorem 5.5). A unified presentation of the theory appears in Rockafeller (1970), and many elements of it are in Barndorff-Nielsen (1978, especially Chapters 5 and 9).

6.7 Lemma

The convex dual d is a lower semicontinuous convex function. Hence, N_d is convex. Suppose ϕ is regularly strictly convex. Then d is strictly convex and twice differentiable on R. On R

(1) $$\nabla d(x) = \theta(x) ,$$

and

(2) $$D_2 d(x) = (D_2\phi)^{-1} (\theta(x)) .$$

Proof. Since d is the supremum of linear functions it is lower semicontinuous and convex.

For $x \in R$, $d(x) = x \cdot \theta(x) - \psi(\theta(x))$. Hence (1), (2) hold, by the same computation that yielded 6.5(3), (4). d is strictly convex on R since $D_2 d$ is positive definite. (It is possible to also directly establish strict convexity without requiring that ϕ be twice differentiable.) ||

It is now convenient to consider

$$\ell_d(x, \theta) = x \cdot \theta - d(x) .$$

Under the conditions of Lemma 6.7 $\nabla d(x) = \theta(x)$ so that for $\theta \in N^\circ$ $\ell_d(\cdot, \theta)$ is uniquely maximized at the value x for which $\theta(x) = \theta$. This value is precisely $\xi(\theta)$. This interpretation is developed further below, especially in Definition 6.10.

The following equivalent expression for steepness is a fundamental building block in the proof of Theorem 6.9, and has other uses.

6.8 Lemma

Let ϕ be regularly strictly convex. Then ϕ is steep if and only if

(1) $$(\{\theta_i\} \subset N^\circ, \quad \theta_i \to \theta \in N - N^\circ)$$

implies

(2) $$||\nabla\phi(\theta_i)|| \to \infty .$$

Proof. Assume (1) implies (2). Let $\theta_0 \in N°$, $\theta_1 \in N - N°$, $\theta_\rho = \theta_0 + \rho(\theta_1 - \theta_0)$. Then

$$(3) \qquad -\ell_d(\xi(\theta_\rho), \theta_0) = d(\xi(\theta_\rho)) - \xi(\theta_\rho) \cdot \theta_0$$

$$= \xi(\theta_\rho) \cdot (\theta_\rho - \theta_0) - \phi(\theta_\rho) \quad .$$

d is strictly convex and twice differentiable on the open set R with $(D_2 d)$ nonsingular on R. Hence

$$(4) \qquad \lim_{||x|| \to \infty} \ell_d(x, \theta) = -\infty$$

for every $\theta \in \theta(R) = N°$ by Lemma 5.3(3). Since $||\xi(\theta_\rho)|| \to \infty$, by (2), we have

$$(5) \qquad \xi(\theta_\rho) \cdot (\theta_\rho - \theta_0) - \phi(\theta_\rho) = -\ell_d(\xi(\theta_\rho), \theta_0) \to \infty.$$

Since $\theta_1 \in N$, $\lim_{\rho \to 1} \phi(\theta_\rho) = \phi(\theta_1)$ is finite. This implies

$$(6) \qquad \xi(\theta_\rho) \cdot (\theta_1 - \theta_0) = \xi(\theta_\rho) \cdot (\theta_\rho - \theta_0)/\rho \to \infty \quad \text{as} \quad \rho \uparrow 1 \ .$$

By definition, ϕ is steep.

Conversely, suppose there is a sequence satisfying (1) for which (2) fails. The sequence can be chosen so that

$$\sup ||\nabla\phi(\theta_i)|| = B < \infty \quad .$$

This means that $\xi(\theta_i) = \nabla\phi(\theta_i)$, $i=1,\ldots$ is a bounded sequence, thus, without loss of generality, the original sequence $\{\theta_i\}$ can be assumed to have been chosen to satisfy $\xi(\theta_i) \to x^*$.

Hence, for any $\theta' \in R^k$

$$(7) \qquad \theta \cdot x^* - \phi(\theta) = \lim (\theta_i \cdot \xi(\theta_i) - \phi(\theta_i))$$

$$\geq \limsup (\theta' \cdot \xi(\theta_i) - \phi(\theta'))$$

$$= \theta' \cdot x^* - \phi(\theta') \quad .$$

It follows that

$$(8) \qquad d(x^*) = \theta \cdot x^* - \phi(\theta) < \infty \quad .$$

This means that $\theta \notin N^\circ$ satisfies $\theta \in \hat{\theta}(x^*)$. By Theorem 5.5 this is impossible if ϕ is steep. Hence ϕ is not steep. ||

Proof of Proposition 3.3. It is now easy to prove the converse assertion in Proposition 3.3, namely that a minimal exponential family satisfying

$$(9) \qquad E_\theta(||x||) = \infty \quad \text{for} \quad \theta \in N - N^\circ$$

is steep.

By Fatou's lemma if $\{\theta_i\}$ satisfies (1) then

$$\lim ||\nabla\psi(\theta_i)|| = \lim ||E_{\theta_i}(x)|| \geq \lim E_{\theta_i}(||x||) = \infty \quad .$$

Hence (2) is satisfied. Thus ψ is steep, which is the desired result. ||

6.9 Theorem

Assume ϕ is steep and regularly strictly convex. Then d_ϕ is also, and

$$(1) \qquad N^\circ_{d_\phi} = R_\phi = \xi(N^\circ) \quad .$$

Proof. Let $x_0 \in R$, $v \in R^k$. Let $\rho_v = \inf \{\rho > 0: \ x_0 + \rho v \notin R\}$. Note that $\rho_v > 0$ since R is open. Assume $\rho_v < \infty$ and let $x_1 = x_0 + \rho_v v$ and $x_\rho = x_0 + \rho(x_1 - x_0)$. Note that $x_1 \notin R$.

Suppose it were true that

$$\text{(2)} \qquad \liminf_{\rho \uparrow 1} ||\theta(x_\rho)|| < \infty \quad .$$

Then there would be a sequence $\rho_i \uparrow 1$ with $\theta(x_{\rho_i}) \to \theta^*$, say. $\theta^* \notin N^\circ$ since $x_1 \notin R = \xi(N^\circ)$. But then, since ϕ is steep, this would imply

$$||x_{\rho_i}|| = ||\xi(\theta(x_{\rho_i}))|| \to \infty$$

by Lemma 6.8, which is a contradiction since $x_{\rho_i} \to x_1$. Hence (2) is false; so that actually

$$\text{(3)} \qquad \lim_{\rho \uparrow 1} ||\theta(x_\rho)|| = \infty \ .$$

The argument in the first part of the proof of Lemma 6.8 applies to yield the dual to 6.8(6), namely

$$\text{(4)} \qquad \theta(x_\rho) \cdot (x_1 - x_0) \to \infty \quad \text{as} \quad \rho \uparrow 1 \quad .$$

(Technically, the lemma as stated cannot be directly quoted since we have not yet established that $R = N_d$ so that d is regularly strictly convex. But, d has the desired convexity and differentiability properties on $R \subset N_d$ by Lemma 6.7. It is then easy to check that the first part of Lemma 6.8 indeed applies since $\{x_{\rho_i}\} \subset R$ and yields (4) as the dual of 6.8 (6).)

d is therefore a convex function with

$$\text{(5)} \qquad \frac{d}{d\rho} d(x_0 + \rho(x_1 - x_0)) \to \infty \quad \text{as} \quad \rho \uparrow 1 \quad .$$

This implies that

$$\text{(6)} \qquad d(x_0 + \rho(x_1 - x_0)) = \infty \quad \text{for} \quad \rho > 1 \ .$$

Since the above argument applies for all $v \in R^k$, it yields that

$$\text{(7)} \qquad d(x) = \infty \quad \text{for} \quad x \notin \bar{R} \quad .$$

Thus $\bar{R} \supset N_d$. This yields (1) since, also, $R \subset N_d$ because $d(x) = \theta(x) \cdot x - \phi(\theta(x)) < \infty$ on R.

It now follows that d is regularly strictly convex since it has the desired smoothness properties, etc., on $R = N_d^\circ$ by Lemma 6.7. And, finally, d is steep since (5) applies to any $x_0 \in R$, $x_1 \in \bar{R} - R$. ||

Remark. Since d is convex, lower semicontinuous, and $d(x) = \infty$ for $x \notin \bar{R}$ it must be that $d(\cdot)$ on R^k is the maximal lower semicontinuous extension of $d(x)$: $x \in R$ $(= K^\circ)$ to all of R^k. That is, for $x_1 \in \bar{R} - R$

$$d(x_1) = \liminf_{\varepsilon \downarrow 0} \{d(x): x \in R, ||x - x_1|| < \varepsilon\} .$$

It follows that if $\{p_\theta\}$ is a steep exponential family. The relation 6.6(4) between $d(x_0)$ and $\tilde{K}(x_0, x_1)$ is valid for all $x_0 \in R^k$, $x_1 \in K^\circ$.

MINIMUM ENTROPY PARAMETER

The path has been prepared for the definition of the dual to maximum likelihood estimation, and for the basic existence and construction theorems.

6.10 Definition

Let d: $R^k \to (-\infty, \infty]$ be convex and lower semicontinuous. Let $S \subset R^k$. Define

$$(1) \qquad \tilde{\xi}_S(\theta) = \{\xi \in S: \ell_d(\xi, \theta) = \ell_d(S, \theta) = \inf\{\ell_d(x, \theta): x \in S\}\} .$$

Obviously $\tilde{\xi}_S$ is related to ℓ_d in the same fashion as $\hat{\theta}$, the maximum likelihood estimator for an exponential family, is related to the log likelihood function ℓ_ψ. (It would therefore seem logical to adopt the notation $\hat{\xi}_S$ rather than $\tilde{\xi}_S$. However for reasons of convenience and tradition we wish to reserve the notation $\hat{\xi}_S$ for the set of maximum likelihood estimates of expectation parameters. That is, $\hat{\xi}_S(x) = \xi(\hat{\theta}_S(x))$.)

The function $\tilde{\xi}_S$ has been given a variety of fairly inconvenient appelations. For example, values in $\tilde{\xi}_S(\theta)$ can be called *minimum entropy* (expectation) *parameters* relative to the set $S \subset K°$. Barndorff-Nielsen (1978) refers to values $\tilde{\theta}_S(x) = \theta(\tilde{\xi}_S(\theta(x)))$, $x \in K°$, as *maximum likelihood predictors*. (Note however that $\tilde{\xi}_S(\theta) \cap (K - K°) \neq \phi$ is possible even if $\{p_\theta\}$ is regular as long as S is not convex (see Theorem 6.13). Hence values in $\tilde{\xi}$ need not always be expectation parameters.)

Another interpretation is provided by the Kullback-Leibler information. Consider a steep minimal exponential family. If $\tilde{\xi} \in \tilde{\xi}_S(\theta) \cap K°$ then

$$\tilde{K}(\tilde{\xi}, \xi(\theta)) = \inf \{\tilde{K}(x, \xi(\theta)): x \in S \cap K°\} .$$

Thus, $\tilde{\theta} \in \theta(\tilde{\xi}_S(\theta_1))$ is a parameter in $\theta(S)$ whose Kullback-Leibler distance to θ_1 is a minimum over all parameters in $\theta(S)$.

Suppose $\{p_\theta\}$ is a minimal, steep standard exponential family. Then Theorem 6.9 establishes that d_ψ is steep and regularly strictly convex with $R = \xi(N°) = K°$. Consequently $\tilde{\xi}$ possesses the properties established for $\hat{\theta}$ in Chapter 5. The main properties are formally stated below; their proofs consist only of reference to the appropriate results in Chapter 5.

Convention. In the following statements $\{p_\theta\}$ is a minimal steep standard exponential family. Note that $R = \overset{\circ}{K} \subset N_d \subset K$.

6.11 Theorem

If $\theta \in N°$ then

(1) $$\tilde{\xi}_N(\theta) = \{\xi(\theta)\} \subset K° \quad .$$

If $\theta \in N - N°$ then $\tilde{\xi}_N(\theta)$ is empty.

Proof. This is the dual statement to Theorem 5.5. ||

Note that

(2) $$\theta(\tilde{\xi}_N(\theta(x))) = \hat{\theta}_N(x) , \quad \text{etc.}$$

In other words, for a full exponential family the maximum likelihood predictor is the same as the maximum likelihood estimator. However (2) does not extend to non-full families.

6.12 Theorem

Let $S \subset N_d$ be a non-empty, relatively closed subset of N_d. Suppose $\theta \in N^\circ$. Then $\tilde{\xi}(\theta)$ is non-empty.

Suppose $\theta \in N - N^\circ$ and there are values $\theta_i \in N^\circ$, $i=1,\ldots,I$ and constants $\beta_i < \infty$ such that

(1) $$S \subset \bigcup_{i=1}^{I} H^-(\theta - \theta_i, \beta_i) \quad .$$

Then $\tilde{\xi}(\theta)$ is non-empty.

For any $\tilde{\xi} \in \tilde{\xi}_S(\theta) \cap K^\circ$

(2) $$\theta - \theta(\tilde{\xi}) \in \nabla_S(\tilde{\xi}) \quad .$$

Proof. Invoke Theorem 5.7 and Theorem 5.12. ||

6.13 Theorem

Suppose $S \cap N_d$ is a relatively closed convex subset of N_d with $S \cap K^\circ$ non-empty. Then $\tilde{\xi}_S(\theta)$ is non-empty if and only if $\theta \in N^\circ$ or $\theta \in \bar{N} - N^\circ$ and

(1) $$S \subset H^-(\theta - \theta_1, \beta_1)$$

for some $\theta_1 \in N^\circ$, $\beta_1 \in R$.

If $\tilde{\xi}_S(\theta)$ is non-empty then it consists of the unique point $\tilde{\xi} \in S \cap K^\circ$ satisfying

(2) $$(\theta - \theta(\tilde{\xi})) \cdot (\tilde{\xi} - \xi) \geq 0 \quad \forall \quad \xi \in S \quad .$$

Proof. Invoke Theorem 5.8. ||

6.14 Construction

Theorems 6.12(2) and 6.13 have a geometrical interpretation which looks exactly like that of their counterparts in Chapter 5. For example, suppose $S = H \cap K$ with H the hyperplane $H(a, \alpha)$, and $H \cap K°$ is non-empty. Then in order to find $\tilde{\xi}_S(\theta)$ one need only search for the unique point $\xi^* \in H$ for which $\theta - \theta(\xi^*) = \rho a$ for some $\rho \in R$. The process can be pictured from two different perspectives. Both of these are shown in Figure 6.14(1).

(i) One may proceed from $\xi(\theta)$ along the curve $\{\xi(\theta + \rho a): \rho \in R\}$ until the unique point at which $\xi(\theta + \rho a) \in H$.

(ii) Alternatively one may map $S \cap K°$ back into Θ as $\theta(S \cap K°)$ and then proceed along the line $\{\theta + \rho a: \rho \in R\}$ until the unique point at which $\theta + \rho a \in \theta(S \cap K°)$.

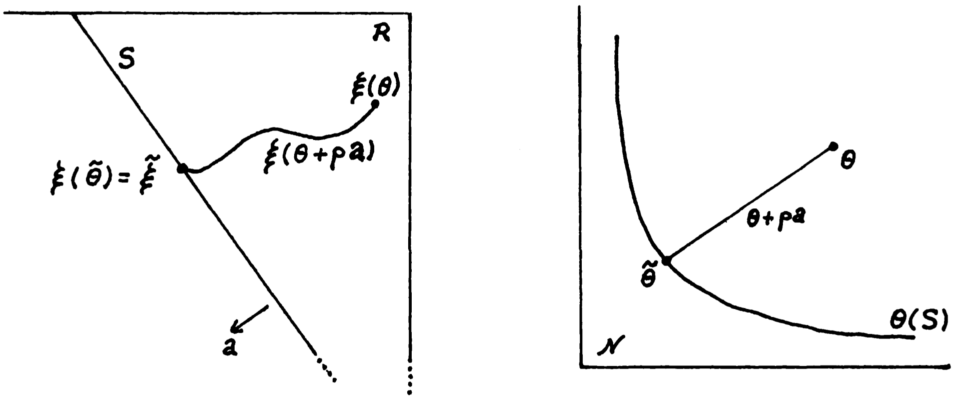

Figure 6.14(1): Construction of $\tilde{\xi}_S(\theta)$ when $S = H(a, \alpha) \cap K$

There is an important statistical difference between the situation pictured here and the dual situation displayed in 5.9.

In Construction 5.9 $\Theta = H \cap N$ and the problem considered was to find $\hat{\theta}_\Theta$. In that case one could proceed via the geometrical dual to Figure 6.14(1). See Figures 5.9(1) and 5.9(2). However, one could also reduce by sufficiency to a minimal exponential family with parameter space Θ. $\hat{\theta}_\Theta$ could then be found by applying Theorem 5.5 to this minimal family. A corresponding statistical interpretation is not available for the dual problem of finding $\tilde{\xi}_{H \cap K}$.

Furthermore, if $\Theta = H \cap N$ and $S = \xi(\Theta)$ the maximum likelihood predictor relative to S cannot legally be found by first reducing by sufficiency. This very undesirable property of a statistical estimator is displayed in the following example.

6.15 Example

Consider the Hardy-Weinberg problem discussed earlier in Examples 1.8 and 5.10. Let $S = \xi(\Theta)$ and consider the problem of finding $\tilde{\xi}_S$. Rather than provide a general formula for $\tilde{\xi}$ (a messy exercise) we discuss a special case, and some implications.

Suppose $N = 18$ and $x = (3,6,9)$. We have already seen that $\hat{p} = \frac{2x_1 + x_2}{2N} = \frac{1}{3}$. Thus $\hat{\xi}(x) = 18(\frac{1}{9}, \frac{4}{9}, \frac{4}{9}) = (2,8,8)$, and

$$(1) \qquad \theta(\hat{\xi}(x)) = \hat{\theta}(x) = \{\rho(1,1,1) + (\ln 1, \ln 4, \ln 4)\}$$

$$= \{\beta_1(1,1,1) - (\ln 2)(2,1,0) + (0, \ln 2, 0)\} \subset \Theta$$

Note also that

$$(2) \qquad \theta(x) = \{\rho(1,1,1) + (\ln 1, \ln 2, \ln 3)\} \ .$$

Of course $\theta(x) \cap \Theta = \phi$.

Since $\xi(p) = (p^2, 2pq, q^2) = (p^2, 2p(1-p), (1-p)^2)$ the tangent space to $S = \{\xi(p)\colon \ 0 < p < 1\}$ can be found by taking $\frac{d}{dp}\,\xi(P)$. Evaluated at $\hat{p} = \frac{1}{3}$ this tangent space, T, is spanned by the vector

$$\tau = (2\hat{p}, \ 2 - 4\hat{p}, \ -2 + 2\hat{p})$$

$$= (\tfrac{2}{3}, \ \tfrac{2}{3}, -\tfrac{4}{3}) \ .$$

By definition $\nabla_S(\hat{\xi}) = \{v\colon \ v \cdot \tau = 0\}$.

Now, from (1) and (2)

$$\theta(x) - \theta(\hat{\xi}) = \{\rho'(1,1,1) + (0, \ln 2 - \ln 4, \ln 3 - \ln 4): \rho' \in R\} .$$

Thus

(3) $$(\theta(x) - \theta(\hat{\xi})) \cdot \tau = (2/3) \ln (1/2) - (4/3)\ln (3/4) \neq 0 .$$

The implication of (3) is that $\theta(x) - \theta(\tilde{\xi}) \notin \nabla_S(\hat{\xi})$. It follows from Theorem 6.12(2) that

(4) $$\tilde{\theta}(x) \cap \hat{\theta}(x) = \phi ,$$

or, in other words,

(4') $$\tilde{\xi}(x) \neq \hat{\xi}(x) .$$

Finally, suppose instead that the sample point is $x^* = (2,8,8)$. Note that $x^* = \hat{\xi}(x)$ with $x = (3,6,9)$, as above. In this case $\hat{\xi}(x^*) = x^*$ and hence

(5') $$\tilde{\xi}(x^*) = \hat{\xi}(x^*) = x^*$$

and

(5) $$\tilde{\theta}(x^*) = \hat{\theta}(x^*) = \theta(x^*) .$$

Recall from the discussion in Example 5.10 that,over the domain $K°$, $\hat{\xi}(x)$ coincides with the minimal sufficient statistic. Thus, from (4) and (5) (or (4') and (5')) it can be seen that here the "estimator" $\tilde{\theta}(x) = \theta(\tilde{\xi}(\theta(x)))$ *is not a function of the minimal sufficient statistic.* This is a very undesirable property for a statistical estimator. Indeed, we emphasize, the primary statistical use of $\tilde{\theta}$ does not lie in its use as a statistical estimator, but rather in its use in the theory of large deviations. See, for example, 7.5 and Exercises 7.5.1 - 7.5.6.

ENTROPY

6.16 Discussion

In statistical mechanics and elsewhere the term entropy appears and has a definition whose connection with the quantity $K(\theta_0, \theta_1)$ for exponential families is not at first obvious. See Ellis (1984a; 1984b).

Let F be a probability distribution on R^k. Let $x \in R^k$ and define the *entropy of x under* F as

(1) $$\mathcal{E}_F(x) = \inf \{K(G, F)\colon \; E_G(X) = x\} \quad .$$

There is, as yet, no exponential family apparent in this definition. However, there is indeed an intimate connection between $\mathcal{E}$ and $\tilde{K}$, as revealed in the following theorem. The theorem is proved only for the case where F satisfies certain mild assumptions and $x \in \mathcal{K}_F^{\circ}$ or $x \notin \mathcal{K}_F$. We leave it to the reader to develop the appropriate results when F does not satisfy these assumptions. The situation where $x \in \mathcal{K} - \mathcal{K}^{\circ}$ can sometimes be treated using the methods at the end of this chapter.

6.17 Theorem

Suppose the exponential family generated by F is a steep minimal family with $0 \in \operatorname{int} \mathcal{N}$. Let $\xi_0 = \xi(0) = E_F(X)$. Let $\tilde{K}$ denote the usual Kullback-Leibler function, 6.1(4), for this exponential family. Then

(1) $$\mathcal{E}_F(y) = \tilde{K}(y, \xi_0)$$

if $y \in \mathcal{K}^{\circ}$. If $y \notin \mathcal{K}$

(2) $$\infty = \mathcal{E}_F(y) = K(y, \xi_0) \quad .$$

Proof. Suppose $y \in \mathcal{K}^{\circ}$, it is obviously true that

(3) $$\mathcal{E}_F(y) \le K(y, \xi_0)$$

since the distribution $G(dx) = p_{\theta(y)}(x)F(dx) = P_{\theta(y)}(dx)$ satisfies $E_G(X) = y$

and $K(G, F) = \tilde{K}(y, \xi_0)$. Suppose $K(G, F) < \infty$ and

$$E_G(X) = y = E_{P_{\theta(y)}}(X) \quad . \tag{4}$$

It must be that $G \propto\propto F$, for otherwise $K(G, F) = \infty$. Let $g = \frac{dG}{dF}$, and $p = p_{\theta(y)}$. Then

$$\begin{aligned} K(G, F) - K(P_{\theta(y)}, F) &= \int[g(x) \ln g(x) - p(x) \ln p(x)] F(dx) \\ &= \int g(x)(\ln g(x) - \ln p(x))F(dx) \\ &\qquad + \int(g(x) - p(x))(\ln p(x))F(dx) \\ &= K(G, P_{\theta(y)}) \geq 0 \end{aligned} \tag{5}$$

since $\int(g(x) - p(x))(\ln p(x))F(dx) = \int(g(x) - p(x))(\theta \cdot x - \psi(\theta))F(dx) = 0$ by (4). It follows from (3) and (5) that (1) holds. (Also, note that $G = F_{\theta(y)}$ is the unique distribution satisfying (4) and yielding $K(G, F) = \mathcal{E}_F(y)$.)

If $y \notin \mathcal{K}$ then $E_G(X) = y$ implies $G \prec\prec F$ and hence $K(G, F) = \infty = \tilde{K}(y, \xi_0)$. ||

AGGREGATE EXPONENTIAL FAMILIES

If $\{p_\theta\}$ is a full canonical exponential family and $x \in \partial\mathcal{K}$ then $\hat{\theta}(x) = \phi$. (See Theorem 5.5.) If $\nu(\partial\mathcal{K}) > 0$ then this means that with positive probability the maximum likelihood estimator fails to exist. This occurs most commonly when ν has countable support. In most such cases the family of distributions $\{p_\theta: \theta \in \mathcal{N}\}$ can be augmented in a natural way so that the maximum likelihood estimator is always defined over this new, larger family of distributions. The augmented family will be called an aggregate exponential family.

Aggregate exponential families can also be satisfactorily defined in a few special cases where ν does not have countable support, but $\nu(\partial K) > 0$ nevertheless. However, such situations are rare in applications and the general theory involves difficulties not present in the countable case; hence we do not treat such situations below. For similar reasons of convenience we avoid non-regular exponential families.

Special cases of the theory are extremely familiar -- for example the aggregate family of binomial distributions, which is just $\mathcal{B}(n, p)$, $0 \le p \le 1$. The general theory for the case where ν has finite support appears in Barndorff-Nielsen (1978, p.154-158), along with some observations about generalizations.

6.18 Definitions

Let ν be a measure concentrated on the countable subset $X = \{x_1, x_2, \ldots\} \subset R^k$. Thus

$$(1) \qquad \nu(\{x_i\}) > 0 \quad i=1,2,\ldots , \quad \nu(X^c) = 0 \quad .$$

Consider the closed convex set $K = K_\nu$. The *faces* of K_ν are the non-empty sets of the form

$$(2) \qquad F = K \cap H(v, \alpha) \quad \text{where} \quad K \subset \bar{H}^-(v, \alpha) \quad .$$

By convention the set K is itself a face of K (corresponding to $v = 0$, $\alpha = 0$).

A face, F, is itself a closed convex subset, which has dimension s, $0 \le s \le k$. (Only the face $F = K$ can have dimension k.) The *relative interior* of F, denoted ri(F) is the interior of F considered as a subset of R^s. An analytic characterization of ri(F) is that $x \in \text{ri}(F)$ if $x \in F$ and if for every hyperplane $H \in R^k$ such that $x \in H$ but $F \not\subset H$ then both $F \cap H^+ \neq \phi$, and $F \cap H^- \neq \phi$.

Let F be a face of K. If $\nu(F) > 0$ then the restriction of ν to F, $\nu_{|F}$ is uniquely defined and non-zero. We use the notation $K_{|F} = K_{\nu_{|F}}$. Note that while it is usually true that $K_{|F} = F$ this need not always be the case.

See Exercise 6.18.1.

The first main theorem involves the following structural assumption on X:

(3) For every $\xi \in X$ there is a face F of K such that $K_{|F} = F$ and $\xi \in \mathrm{ri}(F)$.

If X is finite then (3) is clearly satisfied. Another important case where (3) is satisfied is when $X = \{0,1,\ldots\}^k$, as for example when $X_1,\ldots,X_k$ are independent Poisson or independent negative binomial variables. Assumption 6.22(1) provides an easily verified structural condition which implies (3).

6.19 Definition (Aggregate family)

Let X and ν be as in 6.18. Let $\{p_\theta\}$ be the canonical exponential family of densities generated by ν. Assume the family is regular. As shown in Chapter 3 this family can be reparametrized by the expectation parameter $\xi = \xi(\theta)$. Let

$$q_{\xi(\theta)}(x) = p_\theta(x) \qquad \theta \in N \quad . \tag{1}$$

Then, $\{q_\xi: \ \xi \in K^\circ\} = \{p_\theta: \ \theta \in N\}$.

Now, for each face, F, of K with $\nu(F) > 0$ let $\psi_{|F} = \psi_{\nu_{|F}}$ and define the family of densities

$$p_{\theta|F}(x) = \begin{matrix} \exp(\theta \cdot x - \psi_{|F}(\theta)) & x \in F \\ 0 & x \notin F \end{matrix}$$

relative to the measure ν. This is an exponential family relative to the measure $\nu_{|F}$. Assume this family is regular. Let $\xi_{|F}$ denote its expectation parameter, and let

$$q_{\xi(\theta)|F}(x) = p_{\theta|F}(x) \quad . \tag{2}$$

Thus ξ ranges over the set ri $K_{|F}$ as θ ranges over $N_{|F} = N_{\nu_{|F}}$. Note that the family $\{p_{\theta| F}\colon\ \theta \in \mathcal{N}_{|F}\}$ is not minimal. Hence the map $\theta \to \xi_{|.F}(\theta)$ is not 1 - 1. However, $q_{\xi_1|F} = q_{\xi_2|F}$ if and only if $\xi_1 = \xi_2$, by virtue of Theorems 1.9 and 3.6.

Let

(3) $$\mathcal{F} = \{x\colon\ \exists \text{ face } F \text{ of } \mathcal{K} \ni \nu_{|F} \neq 0 \text{ and } x \in \operatorname{ri}(F)\} .$$

Lemma 6.20, below, establishes that for each $\xi \in \mathcal{F}$ there is a unique F such that $\xi \in \operatorname{ri}(F)$ and a unique density $q_{\xi|F}$ corresponding to the pair ξ, F. This density has

(4) $$E_{q_{\xi|F}}(X) = \xi .$$

We denote this density as q_ξ. The *aggregate family* of densities generated by ν with parameter space $\mathcal{F}$ is the family

(5) $$\{q_\xi\colon\ \xi \in \mathcal{F}\} .$$

Note that

(6) $$P_\xi(X) = 1 \quad \forall\ \xi \in \mathcal{F} .$$

6.20 Lemma

Make the assumptions in 6.18 and 6.19. Then for each $\xi \in \mathcal{F}$ there is a unique F such that $\xi \in \operatorname{ri}(F)$. The density $q_\xi = q_{\xi|F}$ satisfies 6.19(4). It is, in fact, the unique density of the form $q_{\xi'|F'}$ having expectation ξ.

Proof. Suppose $\xi \in \operatorname{ri}(F)$ and also $\xi \in F' = H(v', \alpha') \cap \mathcal{K}$ where $\mathcal{K} \subset \bar{H}^-(v', \alpha')$. Then either (i) $F \subset H(v', \alpha')$ or (ii) $F \cap H^+(v', \alpha') \neq \phi$ and $F \cap H^-(v', \alpha') \neq \phi$. In case (ii) $H(v', \alpha')$ is not a supporting hyperplane, a contradiction. Hence (i) holds, and so $F' \supset F$. Reversing the roles of F, F' in the above now shows that $\xi \in \operatorname{ri}(F)$ and $\xi \in \operatorname{ri}(F')$ implies $F = F'$.

By Theorem 3.6, $\{E_{q_{\xi(\theta)|F}}(x)\colon\ \theta \in \mathcal{N}_{\nu|F}\} = \operatorname{ri}(\mathcal{K}_{|F}) = \operatorname{ri}(F)$ by 6.18(3) since $\nu_{|F}$ generates a regular family. Thus $q_{\xi|F}$ satisfying 6.19(4) exists.

For every $\xi \in X$ the preceding shows that $\xi = E_{q_\xi}(X) \in \text{ri}(F)$ where F is the unique face of K with $\xi \in \text{ri}(F)$. Hence $\xi = E_{q_{\xi|F}}(X) = E_{q_{\xi'|F'}}(X)$ implies $F = F'$, and thus, as previously noted, implies $q_\xi = q_{\xi'}$. ||

Assumption 6.18(3) guarantees that $F \supset X$. If the conclusion of 6.18(3) holds for all $\xi \in$ conhull X then $F =$ conhull X. Otherwise it may occur that $F \subsetneqq$ conhull X. Exercise 6.20.1 sketches an example. If Assumption 6.22(1) is satisfied then

$$(1) \qquad F = \text{conhull } X = K \ .$$

Here is the first main theorem providing the extension of Theorem 5.5.

6.21 Theorem

Make the assumptions in 6.18 and 6.19. Then for $x \in F \supset X$ the maximum likelihood estimator, $\hat{\xi}(x)$, is uniquely determined by the trivial equation

$$(1) \qquad \hat{\xi}(x) = x \ .$$

Proof. Let $x \in \text{ri}(F)$ for some face $F = H(v, \alpha) \cap K$ of K. If $\xi' \in \text{ri}(F')$ and $x \notin F'$ then $q_{\xi'}(x) = 0$.

Now suppose $\xi' \in \text{ri}(F')$, $x \in F'$, but $F' \neq F$. It follows (as in Lemma 6.20) that $F' \supset F$. The argument now takes place in F'. Hence we can assume for convenience, and without loss of generality, that $F' = R^k \cap K$ and $\xi' \in K^\circ$. We may further assume that $x = 0$, $K \subset \bar{H}^-(e_1, 0)$, and $0 \in \text{ri}(F)$ with $F = H(e_1, 0) \cap K$. Then, $\xi' = \xi(\theta')$ for some $\theta' \in N^\circ \subset R^k$. Let $\theta_\rho = \theta' + \rho e_1$, $\rho > 0$. Then

$$(2) \qquad q_{\xi(\theta_\rho)}(0) = \exp(-\psi(\theta_\rho))$$

and

(3)
$$e^{\psi(\theta_\rho)} = \int_{x_1<0} e^{\theta'\cdot x+\rho x_1}\nu(dx) + \int_{x_1=0} e^{\theta'\cdot x}\,\nu(dx)$$

$$\downarrow \int_{x_1=0} e^{\theta'\cdot x}\,\nu(dx) = \psi_{|F}(\theta')$$

by the monotone convergence theorem and the definition of $\psi_{|F}$. It follows from (2) and (3) that

(4)
$$q_{\xi'}(0) < q_{\xi(\theta_\rho)}(0) < q_{\xi''|F}(0) , \quad 0 < \rho < \infty \quad ,$$

where ξ'' is the unique point in ri(F) defined by $\xi'' = \xi_{|F}(\theta')$.

Finally, if $\xi''' \in$ ri(F) then applying Theorem 5.5 to the measure $\nu_{|F}$ yields

(5)
$$q_{\xi'''|F}(0) \leq q_{0|F}(0)$$

with equality only if $\xi''' = 0$. Combining (4), (5), and the first comment in the proof yields

(6)
$$\hat{\xi}(0) = 0 .$$

This verifies (1) when $\xi = 0$, and completes the proof. ||

Remark. As noted in the remark preceding the theorem it is usually true that $F \supset$ conhull X. Assume so and assume the hypotheses of the theorem. Let $X_1,\ldots,X_n$ be i.i.d. random variables with density q_ξ, $\xi \in F$. As usual, let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$. Then $\bar{X}_n \in$ conhull $X \subset F$ with probability one. The family of distributions of the sufficient statistic $\bar{X}_n$ is then also an aggregate family fitting the specifications of the theorem. Hence the maximum likelihood estimator of $\xi \in F$ based on $X_1,\ldots,X_n$ satisfies the trivial equation

(6)
$$\hat{\xi}(X_1,\ldots,X_n) = \bar{X}_n .$$

The preceding theorem yields the existence of maximum likelihood estimates when the parameter space is F. In order to guarantee existence of these estimates when the parameter space is a proper closed subset of K it suffices to establish continuity in ξ of $q_\xi(x)$, $x \in X$. This continuity is useful for other purposes as well. Somewhat unfortunately, the assumptions of Theorem 6.21 do not imply that $q_\xi(x)$ is continuous in ξ (see Exercises 6.23.5-6) and the following theorems demand stronger assumptions. Sufficient assumptions are described below.

There is a further, aesthetic, reason for wanting to know that $q_\xi(x)$ is continuous in ξ. The definition given in 6.19 of the aggregate family $\{q_\xi(x)\colon \ \theta \in F\}$ is structurally natural. But there is also an analytically natural definition for the family of distributions generated from $\{p_\theta\colon \ \theta \in N\}$ -- namely, the set of all probability distributions on X which are limits of sequences of distributions in $\{p_\theta\}$. These two definitions coincide when $q_\xi(x)$ is continuous in ξ.

6.22 Assumptions

K is called a polyhedral convex set if it can be written as the intersection of a finite number of half spaces (see Rockafellar (1970)). Assume that K is a polyhedral convex set and that for every one of the finite number of faces, F, of K

$$(1) \qquad F = K_{|F} \quad .$$

As previously noted in 6.20(1), this implies $F = K = \text{conhull } X$.

For any convex set $S \in R^k$ define the centered span of S to be the subspace spanned by vectors of the form $x - y$, $x,y \in X$. Denote this subspace by csp S. Note that if $x_0 \in \text{ri } S$ then

$$(2) \qquad \text{csp } S = \text{span } \{x - x_0\colon \ x \in S\} \quad .$$

Assume that for every face F of K

$$(3) \qquad \text{Proj}_{\text{csp } F} \, N = \text{Proj}_{\text{csp } F}(N_{|F}) \quad .$$

Note that if X is finite then (1) is satisfied, and (3) is trivially satisfied since $N_{|F} = R^k$ for all faces F (including $F = K$). If ν is a product measure then (1) and (3) are again satisfied. See Exercise 6.22.2.

6.23 Theorem

Make the assumptions in 6.18, 6.19, and 6.22. Then for every $x \in K$, $q_\xi(x)$ is continuous for $\xi \in K$.

Proof. The proof involves an induction on the dimension, k. For k = 1 the result is nearly obvious. Suppose $\xi_0 \in \partial K$. Without loss of generality assume $K \subset (-\infty, \xi_0]$. Then $\xi_i \to \xi_0$ with $\xi_i \neq \xi_0$, i=1,... implies $\xi_i = \xi(\theta_i)$, $\theta_i \to N$, and $\theta_i \to \infty$. It follows that $q_{\xi_i}(\xi_0) = p_{\theta_i}(\xi_0) \to \nu(\{\xi_0\})^{-1} = q_{\xi_0}(\xi_0)$, and for $x \neq \xi_0$, $q_{\xi_i}(x) \to 0 = q_{\xi_0}(x)$.

For arbitrary k, including k=1, if $\xi_0 \in K^\circ$ then $q_\xi(x) = p_{\theta(\xi)}(x)$ is continous on a neighborhood of ξ_0. This completes the proof for k = 1.

We now turn to the case $k \geq 2$. We need to prove continuity of q_ξ at $\xi_0 \in \partial K$. Let $\xi_i \to \xi_0$. We need consider only the case where $\{\xi_i\} \subset F$ with F some face of K, since K has only a finite number of faces. If this F is a proper face of K then $q_{\xi_i} \to q_{\xi_0}$ by the induction hypothesis. Hence we need consider only the case where each $\xi_i = \xi(\theta_i)$, $\theta_i \in N$.

There is a unique face F_0 of K such that $\xi_0 \in \text{ri } F_0 = \text{ri } K_{|F_0}$. Without loss of generality assume $\xi_0 = 0$, $K \subset \bar{H}^-(e_1, 0)$, $-\sigma e_1 \in K^\circ$ for some $\sigma > 0$, $F_0 = H(e_1, 0) \cap K$ and csp $F_0 = \{w \in R^k\colon\ w = (0, \omega),\ \omega \in R^s\}$, $(0 \leq s \leq k-1)$. Let S = csp F_0. For $w \in R^k$ write $w' = (w'_{(1)}, w'_{(2)})$ with $w_{(2)} \in R^s$. Further, assume $0 \in N_{|F_0}$, $\psi_{|F_0}(0) = 0$, $\xi_{|F_0}(0) = 0$. Note that $\psi_{|F_0}(\theta)$ is a function of $\theta_{(2)}$, and so we will write $\psi_{|F_0}(\theta_{(2)})$, where convenient.

We have already assumed $0 \in N_{|F_0}$. Hence $\{\theta \in S\colon\ ||\theta|| \leq \delta_0\} \subset N_{|F_0}$ for some $\delta_0 > 0$. It then follows from 6.22(3) that for each such θ there is a $\sigma(\theta)$, say, such that $\theta + \sigma e_1 \in N$, $\theta \geq \sigma(\theta)$. Since $\{\theta \in S\colon\ ||\theta|| \leq \delta_0\}$ is compact, with $\{\theta \in S: ||\theta|| \le \delta_0, \ \theta + \sigma e_1 \in N\}$ as a relatively open subset, there must, further, exists a $\sigma_0 \ge 0$ such that $\theta + \sigma e_1 \in N$ for all $\sigma \ge \sigma_0$, $\theta \in S$, $||\theta|| \le \delta_0$.

For $\delta \le \delta_0$, $\sigma \ge \sigma_0$ define

$$(1) \qquad Q = Q(\sigma, \delta) = \{\theta \in R^k: ||\theta_{(2)}|| \le \delta,$$

$$\theta_{(1)} \cdot x_{(1)} \le -\sigma||x_{(1)}|| \quad \forall \ x \in K\} \qquad .$$

Note that $\theta_{(1)} \cdot x_{(1)} - \sigma_0 e_1 \cdot x_{(1)} \le (-\sigma + \sigma_0)||x_{(1)}|| \le 0, \ \forall x \ \in K$.

Hence for $\theta \in Q$

$$\lambda(\theta) \le \lambda(\sigma_0 \, e_1) < \infty$$

as in 6.21(4). It follows that $Q \subset N$.

Now assume for convenience, and without loss of generality, that $\sigma_0 = 0$. Then for $\theta \in Q$

$$(2) \qquad \lambda(\theta) = \int e^{\theta \cdot x} \nu(dx) \le \int e^{-\sigma||x_{(1)}||+\theta_{(2)} \cdot x_{(2)}} \nu(dx)$$

$$\to \int e^{\theta_{(2)} \cdot x_{(2)}} \nu_{|F_0}(dx)$$

as $\sigma \to \infty$, uniformly for $\theta_{(2)} \le \delta_0$. In particular

$$(3) \qquad \sup \{|\psi(\theta)|: \ \theta \in Q(\sigma, \delta)\} \to \psi_{|F_0}(0) = 0$$

as $\sigma \to \infty$, $\delta \to 0$. It follows that

$$(4) \qquad \sup \{|p_\theta(x) - q_0(x)|: \ \theta \in Q(\sigma, \delta)\} \to 0 \quad \text{as} \quad \sigma \to \infty, \ \delta \to 0$$

for each $x \in K$. [For $x \in F_0$ the convergence in (4) is uniform over compact subsets of F_0; however if $x \notin F_0$ then as $\sigma \to \infty$, $\delta \to 0$, $p_\theta(x) = e^{\theta \cdot x - \psi(\theta)} \sim e^{\theta \cdot x} \to 0 = q_0(x)$, but the convergence is not uniform over arbitrary compact subsets of K. (It is uniform over bounded subsets of X if $e_1 \cdot x < -\varepsilon < 0$ for all $x \in X - F_0$.)]

It remains to show that for given $\sigma \geq \sigma_0$, $\delta \leq \delta_0$ there is an $\alpha > 0$ such that $||\xi|| < \alpha$, $\xi \in K^\circ$, implies $\theta(\xi) \in Q(\sigma, \delta)$. Once this has been done it follows from (4), and the induction hypothesis, that $q_\xi(x)$ is continuous in $\xi \in K$ for each $x \in K$.

For convenience we show below only that there is an $\alpha > 0$ such that $||\xi|| < \alpha$ implies $\theta(\xi) \in Q(0, \delta)$. The proof for arbitrary $\alpha > 0$, in place of $\sigma = 0$, requires only minor alterations of the constants appearing in the proof. In the following α, ε are generic positive constants whose numerical value may decrease as the proof progresses. Since $0 \in N_{|F_0}$ there is an $\alpha > 0$ such that $||\theta_{(2)}|| > \delta$ implies $\psi_{|F_0}(\theta_{(2)}) \geq 2\beta||\theta_{(2)}||$. Let $C \subset X$ be a finite subset of X such that $C \cap F_0 \neq \phi$ and $F \cap C \neq \phi$ for every face F of K which properly contains F_0. The existence of C is guaranteed by 6.22(1).

Suppose $||\theta_{(2)}|| > \delta$ and $\theta_{(1)} \cdot x_{(1)} > 0$ for some $x \in K$. Then $\max\{\theta_{(1)} \cdot x_{(1)}\colon\ x \in C\} > 0$. If $||\xi|| < \alpha$ and α is sufficiently small then $\xi_{(1)}$ is in the convex hull of $\{x_{(1)}\colon\ x \in C\} \cup \{0\}$. Hence there is an $\eta \in R$ such that

(5) $$\theta_{(1)} \cdot \xi_{(1)} \leq \eta\alpha \max\{\theta_{(1)} \cdot x_{(1)} \cdot x \in C\}$$

for all $||\xi|| < \alpha$. Let $\rho = \max\{||x_{(2)}||\colon\ x \in C\}$, $\nu_0 = \min\{\nu(\{x\})\colon\ x \in C\}$. Then

$$\begin{aligned} \ell(\theta, \xi) &= \theta \cdot \xi - \psi(\theta) \\ &= \theta_{(2)} \cdot \xi_{(2)} - \beta||\theta_{(2)}|| + \theta_{(1)} \cdot \xi_{(1)} - \ln(e^{-\beta||\theta_{(2)}||}\lambda(\theta)) . \end{aligned}$$

Now,

(6) $$\begin{aligned} \lambda(\theta) &\geq \lambda_{|F_0}(\theta_{(2)}) + \nu_0 \exp(\theta_{(1)} \cdot x_{(1)} + \theta_{(2)} \cdot x_{(2)}) \\ &\geq \exp(2\beta||\theta_{(2)}||) + \nu_0 \exp(\theta_{(1)} \cdot x_{(1)} - \rho||\theta_{(2)}||) . \end{aligned}$$

For notational simplicity let $t = \theta_{(1)} \cdot x_{(1)} > 0$. Then for $\alpha \leq \beta/2$

$$(7)\quad \ell(\theta, \xi) \leq \theta_{(2)} \cdot \xi_{(2)} - \beta||\theta_{(2)}|| + \eta\alpha t - \ln (e^{\beta||\theta(2)||} + \nu_0 \exp (t - \rho||\theta_{(2)}|| - \beta||\theta_{(2)}||)$$

$$\leq -\varepsilon + \eta\alpha t - (\beta||\theta_{(2)}|| \ V(t - (\rho + \beta)||\theta_{(2)}|| + \ln \nu_0))$$

$$\leq -\varepsilon$$

for $\alpha > 0$ sufficiently small, since

$$\beta||\theta_{(2)}|| \ V(t - (\rho + \beta)||\theta_{(2)}|| - a\delta) \geq \frac{\beta t}{\rho + 2\beta + a}$$

for $||\theta_{(2)}|| > \delta$, $a \geq 0$.

If $||\theta_{(2)}|| > \delta$ but $\theta_{(1)} \cdot x_{(1)} \leq 0$ for all $x \in K$ then $\theta_{(1)} \cdot \xi_{(1)} \leq 0$ and

$$(8)\quad \ell(\theta, \xi) \leq \theta_{(2)} \cdot \xi_{(2)} - \psi_{IF_0}(\theta_{(2)}) + \theta_{(1)} \cdot \xi_{(1)} - \ln\left(\frac{\lambda(\theta)}{\lambda_{IF_0}(\theta)}\right)$$

$$\leq \theta_{(2)} \cdot \xi_{(2)} - \psi_{F_0}(\theta_{(2)}) \leq -\varepsilon \ .$$

If $||\theta_{(2)}|| \leq \delta_1$ but $\theta_{(1)} \cdot x_{(1)} > 0$ for some $x \in K$ then $\theta_{(1)} \cdot x_{(1)} > 0$ for some $x \in C$; and

$$(9)\quad \ell(\theta, \xi) \leq \theta_{(2)} \cdot \xi_{(2)} - \psi_{IF_0}(\theta_{(2)}) + \eta\alpha\theta_{(1)} \cdot x_{(1)}$$

$$- \ln\left(1 + \frac{\nu_0 e^{\theta(1) \cdot x(1)}}{\psi_{IF_0}(\theta_{(2)})}\right)$$

$$\leq -\varepsilon < 0$$

for $\alpha > 0$ and some $\varepsilon > 0$ sufficiently small, since $\psi_{IF_0}(\theta_{(2)}) \geq 0$ but $\sup \{\psi_{IF_0}(\theta_{(2)}): \ ||\theta_{(2)}|| \leq \delta_1\} < \infty$. If $||\xi|| < \alpha$ and $\theta \notin Q$ one of (7), (8), or (9) apply so that

$$\ell(\theta, \xi) \le -\varepsilon < 0 \; . \tag{10}$$

On the other hand, there is a $\sigma > 0$ sufficiently large so that by (2) or (3),

$$\begin{aligned} \ell(\sigma e_1, \xi) &= \sigma e_1 \cdot \xi - \psi(\sigma e_1) \ge \sigma e_1 \cdot \xi - \varepsilon/3 \\ &\ge -2\varepsilon/3 \end{aligned} \tag{11}$$

for $||\xi|| < \alpha \le \frac{\varepsilon}{3\sigma}$. It follows from (10) and (11) that if $||\xi|| < \alpha$, $\xi \in K°$, then if $\theta \notin Q$

$$\ell(\theta, \xi) \le -\varepsilon < -2\varepsilon/3 \le \ell(\theta(\xi), \xi).$$

Hence $\theta \neq \theta(\xi)$. It follows that $\theta(\xi) \in Q$.

We have thus proved that given σ, δ there is an $\alpha > 0$ such that $||\xi|| < \alpha$, $\xi \in K°$, implies $\theta(\xi) \in Q(\sigma, \delta)$. As previously noted, this completes the proof of the theorem. ||

EXERCISES

6.6.1

Assume ϕ is regularly strictly convex. Verify 6.6(3).

6.7.1

For ϕ regularly strictly convex, when does $d_\phi = \phi$?

6.9.1

Generalize Theorem 3.9 to apply to steep, regularly convex functions ϕ [i.e.; write $\phi = \begin{pmatrix} \phi_{(1)} \\ \phi_{(2)} \end{pmatrix}$ and consider the map $\theta \to \begin{pmatrix} \xi_{(1)}(\theta) \\ \phi_{(2)}(\theta) \end{pmatrix}$. Show this map is 1 - 1 and continuous on $N°$ with range $\xi_{(1)}(N°) \times \phi_{(2)}(N°) = K_{(1)} \times \phi_{(2)}(N°)$].

6.18.1

(i) Show that $K_{|F} \neq F$ in the following example:

$X = (1, -1) \cup \{((i^2 - 1)^{\frac{1}{2}}/i,\ 1/i);\ i=1,2,\dots\}$, $F = K \cap H((1, 0), 1)$.

(ii) Construct an example of the same phenomenon in R^3 where X is a discrete set (i.e. X has no accumulation points in R^3). [Construct X so that the set X in (i) is its projection on the space spanned by the first two coordinate axes.]

6.19.1

Show that the following three families are aggregate exponential families:

(i) Binomial (n, p), $0 \le p \le 1$

(ii) Poisson (λ), $\lambda \ge 0$

(iii) Multinomial (N, $\underset{\sim}{p}$), $0 \le p_i$, $\sum_{i=1}^{k} p_i = 1$.

6.19.2

Suppose the distribution of $X^{(i)}$ form an aggregate exponential family $\{q_\xi^{(i)}\}$, i=1,2, and $X^{(1)}$, $X^{(2)}$ are independent. Show that the distributions of $(X^{(1)}, X^{(2)})$ form a ($k_1 + k_2$ parameter) aggregate exponential family.

6.20.1

Construct an example in which 6.18(3) holds but $F \neq$ conhull X. [Let X' be the set in 6.18.1(i) and define $X \in R^3$ by

$X = \{x\colon (x_1, x_2) \in X',\ \ x_3 = \pm(1 - x_2)\} \cup (1,0,1) \cup (1,0,-1)$.]

6.21.1

Let X be the set defined in 6.20.1 with the additional point (1,0,0). Show

(i) 6.18(3) fails at $x = (1,0,0)$.

(ii) The maximum likelihood estimate for the aggregate family $\{q_\xi\colon\ \xi \in F\}$ fails to exist (i.e. is the empty set) when $X = (1,0,0)$, which occurs with positive probability.

(iii) The failure in (ii) can be rectified in a natural way by letting $G = \text{conhull}\ \{(1,0,-1), (1,0,1)\}$ and adding the densities $q_{\xi(\theta)|G} = p_{\theta|G}$ to the family $\{q_\xi\colon \xi \in F\}$.

(iv) Addition of the densities $q_{\xi|G}$ is "natural" in the sense that for each $\xi \in G$ there is a sequence $\theta_i \in N°$ such that $q_{\xi|G}(x) = \lim_{i\to\infty} p_{\theta_i}(x)$. [This sequence cannot be chosen to be of the form $\theta_i = \theta' + iv$ for fixed $v \in R^k$, $\theta' \in N°$ as was the case in the proof of Theorem 6.21.]

6.21.2

Let ν be linear measure on the perimeter ∂S, of the unit square, S. This measure does not have a countable supporting set. Nevertheless, describe its "natural aggregate family", having parameter space S and satisfying the conclusion of Theorem 6.21 for each $x \in S$.

6.21.3

(i) Let ν be uniform measure on the perimeter S, say, of the unit circle S. Thus, $\{p_\theta\}$ is the family of Von-Mises distributions (Example 3.8). Show there can be no possible way of constructing a family of densities $\{q_\xi\}$ which contains $\{p_\theta\}$ such that the maximum likelihood estimate for $\{q_\xi\}$ exists with probability one. [$\lim_{||\theta||\to\infty} p_\theta(x) = \infty$ for each $x \in \partial S$.]

(ii) Note that if $\bar{X}_n$ is the sample mean from a sample of size n, $n \geq 2$, having the above distribution, then the maximum likelihood estimate does exist with probability one.

(iii) Construct a measure ν for which $\{p_\theta\}$ is a regular exponential family but there does not exist an n for which it is possible to construct an "aggregate family" of densities $\{q_\xi\}$, containing the densities of $\bar{X}_n$ under θ, such that the maximum likelihood estimator exists with probability one. [There exists such a measure ν having $K_\nu = \{x \in R^3\colon\ x_2^2 + x_3^2 \leq x_1^2,\ 0 \leq x\ \leq 1\}$, and $\nu(\{0\}) > 0$.]

6.22.2

Show that 6.22(1) (including the polyhedral nature of K) implies 6.20(1). [The polyhedrality of K guarantees that for every $x \in \partial K$ there is a face F of K such that $x \in$ ri F.]

6.22.2

Prove that 6.22(1) and 6.22(3) are satisfied whenever ν is a product measure on a countable set $X = \prod_{j=1}^{k} X_j$, $X_j \in R$. [The faces $F = H(v, \alpha) \cap X$ of X are determined uniquely by $(\text{sgn } v_1,\ldots,\text{sgn } v_k)$.]

6.22.3

(i) Prove that

(1) $$N_{|F} = \text{Proj}_{\text{csp } F}(N_{|F}) \times (\text{csp } F)^\perp \quad , \qquad \text{and}$$

(2) $$\text{Proj}_{\text{csp } F}(N) \subset \text{Pr}_{\text{csp } F}(N_{|F}) \quad .$$

(ii) Give an example in which $X = \{0,1,\ldots\}^2$,

$$F = \{(0,\ 0),\ (1,\ 0),\ldots\}, \qquad N = (-\infty, 0)^2 \quad ,$$

(3) $$\text{Proj}_{\text{csp } F}(N) = (-\infty,\ 0) \times 0 \neq R \times 0 = \text{Proj}_{\text{csp } F}(N_{|F}) \quad ,$$

and

$$\xi_{|F}((0,\ 0)) = (1,\ 0) \in X\ . \tag{4}$$

(Thus 6.22(3) is not valid here.)

(ii) In the example (ii) show that $q_\xi((x_1,\ 0))$, $x_1 = 0, 1, \ldots$, is not continuous at $\xi = (\xi_1,\ 0)$, $\xi_1 > 1$. [If θ_i is chosen so that $\theta_{i1} \uparrow 0$ somewhat slowly and $\theta_{i2} \to -\infty$ then $\xi(\theta_i) \to (\xi_1,\ 0)$ but $q_{\xi(\theta_i)}(x) \to q_{(1,0)}(x)$.]

6.23.1

Prove versions of Theorems 5.7, 5.8 and 5.12 valid for aggregate exponential families. [Make the assumptions in Theorem 6.23.]

6.23.2

Show that $q_\xi(x)$ is not jointly continuous in $(\xi,\ x)$ at any point with $\xi = x \in \partial K$.

6.23.3

Are the analogs to Theorems 6.12 and 6.13 valid for aggregate exponential families under the assumptions of Theorem 6.23?

6.23.4

Suppose $X = (0,\ 0) \cup \{x \in R^2\colon\ x_i = 1, \ldots,\ i = 1,2\}$. Note that Assumption 6.22(1) is not satisfied. Show that, nonetheless, $q_\xi(x)$ is continuous at every $\xi \in$ conhull $X = F$. (If one *defines* $q_\xi(x) = q_0(x)$ for $\xi \in K$ - conhull X then it is even true that $q_{\cdot}(x)$ is continuous on K.)

6.23.5

Let $X = \{((i^2 - 1)^{\frac{1}{2}}/i,\ 1/i)\colon\ i = 1, \ldots\} \cup (1,\ 0)$. For $x = ((i^2 - 1)^{\frac{1}{2}}/i,\ 1/i) \in X$ let $\nu(\{x\}) = 1/2^i$, and let $\nu(\{0\}) = 1$. Note that 6.22(1) is not satisfied. Show that $q_\xi((1,0))$ is not continuous at $\xi = (1,0)$ [$q_{(1,0)}((1,0)) = 1$. Let $0 < c < 1$. For ℓ sufficiently large let $\theta_\ell = \rho_\ell x_\ell$ with ρ_ℓ chosen so that $p_{\theta_\ell}((1,0)) = c$ ($\{\rho_\ell\}$ is a swiftly increasing sequence.) Then $\xi(\theta_\ell) \to (1, 0)$ but $q_{\xi(\theta_\ell)}((1, 0)) \equiv c \neq 1$.] (In this example $q_\xi(0)$ is, however, upper semicontinuous; so that, for example, the conclusion of Theorem 6.23 remains valid. Exercise 6.23.4 shows this need not be the case.)

6.23.6

For $x = x^{(ij)} = ((i^2 - 1)^{\frac{1}{2}}/i\ ,\ 1/i,\ j)$, $i=1,\ldots,$ $j = \pm 1$, let $\nu(\{x^{(ij)}\}) = (4 + 3j)/2^i$. For $x = x^{(j)} = (1, 0, j)$, $j=-1, 0, +1$ let $\nu(\{x\}) = 2 - |j|$. Otherwise $\nu(\{x\}) = 0$.

Construct $\{\theta_\ell\}$ in a manner similar to 6.23.5 with $(\theta_\ell)_3 = 0$ so that $P_{\theta_\ell}(\{x^{(j)}:\ j=0, \pm 1\}) \uparrow 1/3$ and $(\xi(\theta_\ell))_1 \to 1$. Verify that $\xi(\theta_\ell) \to (1, 0, 1/2)$ and $P_{\theta_\ell}(\{x^{(-1)}\}) = p_{\theta_\ell}(x^{(-1)}) \uparrow 1/12$, but $q_{(1,0,\frac{1}{2})}(x^{(-1)}) = P_{(1,0,\frac{1}{2})}(x^{(-1)}) = (1/4)^2 < 1/12$. Hence $q_\xi(x^{(-1)})$ is not continuous at $\xi = (1, 0, 1/2)$ or even upper semicontinuous. If $E \subset K$ is the closed set $\{\xi(\theta_\ell): \ell=1,\ldots\} \cup (1, 0, 1/2)$ then the maximum likelihood estimator over the family $\{q_\xi: \xi \in E\}$ fails to exist at the possible observation $x^{(-1)}$.