EXISTENCE AND DETERMINATION OF OPTIMAL ESTIMATORS RELATIVE TO CONVEX LOSS*)

M. M. RAO

(Received Sept. 22, 1964)

1. Introduction

Let X be a random variable on a probability space (Ω, Σ, μ) with values in an abstract set $\mathfrak{X}$, endowed with a σ-field. If the probability distribution of X in $\mathfrak{X}$ depends on a parameter θ, an element of a set ϑ, denoted by $F(x \mid \theta)$, then an important statistical problem is to estimate θ, or a real valued function g of θ, based on an observation of X in an 'optimal' way relative to a general convex loss function. Here the optimality has (at least) two interpretations leading to two types of problems as follows.

Let $W(\cdot)$ be a symmetric nonnegative convex function on the line vanishing at the origin and $W(x) \uparrow \infty$ as $x \uparrow \infty$. The first problem is this. If $\mathfrak{M}$ is the class of all unbiased estimators of $g(\theta)$ [i.e. random variables (r. v.'s) $T(X)$ such that $E_\theta(T(X))=g(\theta)$, $\theta \in \vartheta$], for which $E_\theta(W(T-g))$ exists, then the problem is to find (necessary and sufficient) conditions for the existence (and uniqueness) of estimators T^* in $\mathfrak{M}$ which are optimal at θ_0, relative to W, in the sense that $E_{\theta_0}[W(T^*-g)] \leqq E_{\theta_0}[W(T-g)]$, for all T in $\mathfrak{M}$. Here $E_\theta(f(X))=\int_{\mathfrak{X}} f(x)dF(x \mid \theta)$, and W is called the loss function. The second problem is to estimate the value θ of Θ, a r. v. on (Ω, Σ, μ) with values in ϑ which has a σ-field in it. More precisely, if the distribution of X depends on Θ, then the problem is to find an estimator T^* on $\mathfrak{X}$ of $g(\Theta)$, such that $E[W(T^*(X)-g(\Theta))] \leqq E[W(T(X)-g(\Theta))]$ for all T for which the right side exists. Here $E(f)=\int_{\mathfrak{X}\times\vartheta} f(x, \theta)d\nu(x, \theta)$, where $\nu(\cdot, \cdot)$ is the joint distribution of X and Θ. (The symbols E_θ and E will be used in this sense.) The solution T^* in this case is termed a Bayes estimator. (Optimality has other interpretations such as the minimax type, but only the above two cases will be considered here.)

*) This work was supported under contract No. DA-36-061-ORD-477 and NSF grant GP-1349.

The purpose of this paper is to consider both the above problems, (i. e., the Bayes and unbiased cases) and to show that similar (closely related) *methods and techniques* yield the solutions (cf. sections 2 and 3). Also an explicit construction of the optimal estimator in each case will be presented (cf. section 4). Finally, the technical relations between the results of the above two problems will be discussed.

At this point a comparison of this paper with earlier results is in order. The problems of the first (i.e. unbiased) type have been extensively treated in the literature if $W(x)=x^2$, and a general treatment for $W(x)=|x|^p$, $p>1$, is found in the important paper of Barankin, [1].

Then some of these results were extended in [6] if $W(\cdot)$ is convex and its complementary function $V(\cdot)$ (see (1) below) satisfies a growth condition ($V(2x)\leqq CV(x)$). The general case ($W(\cdot)$ is an arbitrary continuous convex function) was briefly outlined in ([7], section 5). Theorem 1 below is a completion of the latter. Thus $W(x)=|x|$, which was not treated before in these studies, is also covered now. The Bayes estimation for general convex functions is treated in [3]. However, if the parameter satisfies certain prescribed conditions, then the methods of [3] do not apply and that case is treated in theorem 2 below so that [3] and what follows complement each other. The construction of optimal estimators was given in [1] for the case, $W(x)=|x|^p$, $1<p<\infty$, and the corresponding result for (general) convex loss functions was not available before. This is treated here for a certain class. Further generalization is clearly possible in this case, but it depends on several unsolved problems of the Orlicz space theory. The present note subsumes almost all the earlier results in this direction. To aid reading, some explanatory comments are inserted at various places.

2. Optimal unbiased estimation

Let $\{P_\theta, \theta\in\Theta\}$ be a family of probability distributions on $\mathfrak{X}$ of the r. v. X. Here X may be a vector or a sequence r. v. Suppose that P_θ is dominated by a fixed (σ-finite) measure λ_0 with densities $p_\theta(x)(=p(x\,|\,\theta))$. If $W(\cdot)$ is a symmetric convex function on the line such that $W(0)=0$, $W(t)\uparrow\infty$ as $t\uparrow\infty$, then it is called a Young's function and there exists a convex function $V(\cdot)$ with similar properties and satisfying the (W. H. Young's) inequality (cf., [8], p. 25, and [5])

$$ab\leqq W(a)+V(b) \tag{1}$$

for all real numbers a, b with equality if and only if either $a=V'(b)$ or $b=W'(a)$. [Here W', V' are derivatives of W, V which exist a. e. (Lebesgue).] Then W, V are termed complementary Young's functions and it can be seen that they satisfy also

$$(1') \qquad |ab| \geqq \min(W(a), V(b)).$$

If L^W is the space of (equivalence classes of) real functions on $\mathfrak{X}$ such that $f \in L^W$ if $\|f\|_W < \infty$ where

$$(2) \qquad \|f\|_W = \sup_h \int_{\mathfrak{X}} |fh| d\lambda, \quad \text{with} \quad \int_{\mathfrak{X}} V(h) d\lambda \leqq 1,$$

and where $d\lambda = p(x \mid \theta_0) d\lambda_0$, then it is known that $\|\cdot\|_W$ is a norm and with it L^W is a complete normed linear (or B-) space. Sometimes L^W is termed an Orlicz space. (Here and elsewhere $f \in L^W$ means f is any member of the equivalence class to which it belongs.) Moreover it contains all the r.v.'s $T(X)$ such that $\int_{\mathfrak{X}} W(T(x)) d\lambda < \infty$. (See [7], sec. 1. Only special cases of that section are used here. See also [5].) The space L^V is defined similarly. Assume that $D_0 = \frac{p_\theta(\cdot)}{p_{\theta_0}(\cdot)} \in L^V$ for θ in ϑ, so that $\int_{\mathfrak{X}} TD_0 d\lambda$ exists (by Hölder inequality) and it follows that

$$(3) \qquad g(\theta) = \int_{\mathfrak{X}} T(x) p(x \mid \theta) d\lambda_0 = E_\theta(T(X)), \quad \theta \in \vartheta.$$

There is another (equivalent) norm $N_W(\cdot)$ on L^W given by

$$(4) \qquad N_W(f) = \inf \left\{ k > 0, \int_{\mathfrak{X}} W\left(\frac{f}{k}\right) d\lambda \leqq 1 \right\}, \quad f \in L^W,$$

and the norms (2) and (4) are connected by

$$(5) \qquad N_W(f) \leqq \|f\|_W \leqq 2N_W(f).$$

Thus L^W is a B-space under either norm. Unless the contrary is stated, in what follows the norm in L^W is taken to be (4) and that in L^V is (2). Due to (5), this is just a convenience and a result of [7] can be directly used. (See also [A].)

With the above preliminaries the problem can be recast and the solution given as follows. It is more convenient, for the L^W-theory, to consider the optimality of an (unbiased) estimator T of $g(\theta)$ at θ_0 as the minimum value of $N_W(T - g(\theta_0))$ instead of that of $E_{\theta_0}[W(T - g(\theta_0))]$. (This is actually more general since the existence of the latter expectation implies the existence of the norm but not conversely. Moreover, this consideration corresponds naturally to the case $W(x) = |x|^p$, $p > 1$, of [1].) Let $D_i(X, \theta)$, $i = 0, 1, 2, \cdots$, be such that (D_0 is defined above (3)) (i) $\{D_i\} \subset L^V$ and (ii) $\int_{\mathfrak{X}} TD_i d\lambda = \alpha_i(\theta)$, $\theta \in \vartheta$, where T is any unbiased estimator of $g(\theta)$, $T \in L^W$, and $\alpha_i(\theta)$, for each θ, are known constants. Such

elements D_i exist. In fact if ϑ is the line and $p(x \mid \theta)$ are differentiable, then $D_i=[p(\cdot \mid \theta)]^{-1}\partial^i p/\partial\theta^i$ are such functions whenever $\int_{\mathfrak{X}} V(D_i)d\lambda<\infty$, and the standard regularity conditions on $p(x \mid \theta)$ for interchange of the derivative and integral hold. (In this case $\int_{\mathfrak{X}} D_i d\lambda=0$, $i\geqq 1$, also holds.) In [1], $D_i=p(\cdot \mid \theta_i)[p(\cdot \mid \theta_0)]^{-1}$. It is not difficult to construct other types of functions even without assuming the above regularity conditions. Also without loss of generality $g(\theta_0)$ may be taken as zero, as otherwise $\tilde{T}=T-g(\theta_0)$ will suffice in the following. The complete solution to the problem of unbiased estimators that are optimal (at θ_0) is given by

THEOREM 1. *Let W and V be convex (Young's) functions and let L^W and L^V be the associated Orlicz spaces (relative to $\mathfrak{X}$ and λ) with norms (2) and (4). If $\mathfrak{M}\subset L^W$ is the set of all unbiased estimators T of $g(\theta)$, and $\{D_i\}\subset L^V$ such that $\int_{\mathfrak{X}} TD_i d\lambda=\alpha_i(\theta)$, then the following conclusions hold.*

(a) $\mathfrak{M}$ is nonempty only if (i) there exists a K such that for any finite set $\{D_{i_1}, \cdots, D_{i_n}\}$ of $\{D_i\}$ and any scalars $\{a_1, \cdots, a_n\}$, (and fixed $\theta\in\vartheta$),

$$(6)\qquad \left|\sum_{j=1}^{n} a_j\alpha_{i_j}(\theta)\right| \leqq K\left\|\sum_{j=1}^{n} a_j D_{i_j}\right\|_V,$$

and (ii) $\{D_i\}\subset M^V$, where M^V is the closed subspace of L^V spanned by the λ-simple (or just bounded) functions. On the other hand, if $\mathfrak{M}$ is nonempty (i) always holds whether or not (ii) is true, with $K=N_W(T)$, T in $\mathfrak{M}$. $\left[\text{So } M^V=\left\{f: \int_\Omega V(\alpha f)\,d\lambda<\infty,\ f\in L^V,\ \alpha>0 \text{ arbitrary}\right\}.\right]$

(b) If $T\in\mathfrak{M}$ then $N_W(T)\geqq K_0=\inf\{K>0$ satisfying (6)$\}$.

(c) If $T_0\in\mathfrak{M}$, $N_W(T_0)=K_0$ and V, W are continuous, then it is essentially unique so that T_0 is the unique unbiased estimator of $g(\theta)$ which is optimal at θ_0. There may be nonuniqueness if either W or V is discontinuous.

(d) If $\{D_i\}\subset M^V$ is fundamental (i.e. spans M^V), then $\mathfrak{M}$ contains at most one element.

Remark. If $W(x)=|x|^p$, $1<p\leqq\infty$, this was proved in [1] and if $V(2x)\leqq CV(x)$ (which includes the former) it was proved in ([6], theorem 4). Also it is known that a Young's function $W(\cdot)$ must either be continuous or, if it is discontinuous, must be of the form $W(x)=0$, for $|x|<x_0(<\infty)$, and $=\infty$ for $|x|>x_0$. In the second case, (since the measure λ is finite) $L^W=L^\infty$ (and $L^V=L^1$). Thus the theorem in this case was already treated by Barankin ([1], p. 485) so that only the case

that $W(\cdot)$ is continuous is new. It will be treated here. Note that $W(x)=|x|$ is included now. The main point of this result is that $W(\cdot)$ is allowed to grow exponentially fast. In the case $M^V=L^V$, then a(ii) is automatic and the statement of (a) can be stated as (6) being necessary and sufficient. The present statement is more general since $M^V \subset L^V$ can be proper. [In [6], the growth condition should be on ϕ. Through an oversight φ, ϕ were interchanged for this condition.] It is clear that the optimal estimator at θ_0 in general depends on θ_0 since λ depends on θ_0. (Note that $g(\theta_0)=0$, by convension.)

PROOF. (a) First consider the direct case. If $\mathfrak{M}$ is nonempty then (b) always holds whether or not a(ii) is assumed since, by Hölder inequality and the definition of $\alpha_i(\theta)$,

$$|\sum_{j=1}^{n} a_j \alpha_{i_j}(\theta)| = |\sum_{j=1}^{n} \int_{\mathfrak{X}} TD_{i_j} d\lambda| \leqq N_W(T) \|\sum_{j=1}^{n} a_j D_{i_j}\|_V .$$

If $K=N_W(T)$, this is (6). Note that K is a constant independent of θ $(\neq\theta_0)$.

Conversely suppose the hypothesis of (a) holds. Since $M^V \subset L^V$ is a closed subspace it is a B-space and by hypothesis $D_i \in M^V$. It then follows from (6) and a theorem of Hahn (cf., [4], p. 86) that there exists a continuous linear functional F on M^V such that $F(D_i)=\alpha_i(\theta)$, for $i=0, 1, \cdots$, and that $\|F\| \leqq K$. If moreover $K_0=\inf\{K>0$ satisfying (6)$\}$, then the F may be chosen such that $\|F\|=K_0$. Now by ([7], theorem 4) this functional can be represented as

$$F(D_i)=\int_{\mathfrak{X}} TD_i d\lambda \tag{7}$$

for a unique T in L^W where $N_W(T)=\|F\|$. This T has also the property by (7), taking $i=0$,

$$g(\theta)=\int_{\mathfrak{X}} TD_0 d\lambda=\int_{\mathfrak{X}} Tp(x \mid \theta) d\lambda_0 ,$$

so that it is an unbiased estimator of $g(\theta)$ and $N_W(T)=K_0$. It follows that $T \in \mathfrak{M}$ and that T is optimal at θ_0. It is remarked that (7) need not hold (so there need not exist a T in L^W) if $\{D_i\} \not\subset M^V$, (cf., [7], theorem 4 and remark 2 after it). This completes the proof of (a).

(b) By the very definition of K_0, it follows that $N_W(T) \geqq K_0$ for every $T \in \mathfrak{M}$.

(c) If $T_0 \in \mathfrak{M}$, $N_W(T_0)=K_0$, and W, V are continuous then T_0 is unique. For, if $T_1 (\in \mathfrak{M})$ has the property $N_W(T_1)=K_0$ then by definition (or convexity) of $\mathfrak{M}$, $1/2(T_0+T_1) \in \mathfrak{M}$ and $N_W(1/2(T_0+T_1)) \geqq K_0$. Excluding the (true and) trivial case $K_0=0$, one has by the triangle inequality,

(8) $$K_0 \leqq 1/2 N_W(T_0+T_1) \leqq 1/2[N_W(T_0)+N_W(T_1)] = K_0 .$$

From the continuity of W and V it follows that $T_0 = aT_1$ for some $a > 0$. Since $K_0 = N_W(T_0) = aN_W(T_1) = aK_0$, it results that $a=1$. If $W(\cdot)$ is discontinuous, however, $L^W = L^\infty$, the norm is equivalent to the *sup*-norm so that (8) does not imply $T_0 = aT_1$. If V is discontinuous $L^V = L^\infty$ and so $L^W = L^1$, and in this case also (8) does not imply $T_0 = aT_1$ if T_0 is positive or negative a.e. If these two cases are excluded, then $T_0 = aT_0$ holds, a.e. Thus in general in these cases (W or V is discontinuous) there may be nonuniqueness. This completes the proof of (c).

(d) If $\{D_i\} \subset M^V$ is fundamental, then it is well-known that a continuous linear function defined on such a set is unique on M^V. This means that there is at most one unbiased estimator T in $\mathfrak{M}$ that is optimal at θ_0. This completes the proof of (d) and with it the theorem.

Note. If it is *assumed* that $dP_\theta(x) = p_\theta(x) d\lambda_0$ where $p_\theta(\cdot)$ is a measurable function of x, then λ_0 need not be σ-finite, as in ([1], p. 477). However, if $\{P_\theta, \theta \in \vartheta\}$ is a family of probability measures that are only dominated by λ_0, an arbitrary measure, then also $\mathrm{p}_\theta(\cdot)$ exists but, in general, will not be measurable. Even then, however, $p_\theta(\cdot)$ will be "quasi-measurable" (i. e., $p_\theta(\cdot)$ will be equivalent to a measurable function on every set of finite λ_0-measure consistently), and the same will be true of T in the converse part of (a) of the theorem. All the integrals are defined for such functions also, and thus the result holds in this generality with quasi-measurability instead of measurability. (cf., [7], sec. 4 and references there.)

In the following section the problem will be considered from the second (or Bayesian) point of view.

3. Bayes estimation

For the Bayesian case the problem takes on a different viewpoint. The r. v. X has a distribution depending on a parameter Θ, which is also a r. v. on Ω to ϑ. For a given prior (or marginal) distribution of Θ, if $F(\cdot \mid x)$ is the posterior (or conditional) distribution of Θ when X has the value x then the problem is to find a Bayes estimate $T^*(x)$ of $g(\theta)$ (based on x), i. e., to find $T^*(x)$ such that $E^B[W(g(\Theta) - T^*(x))] \leqq E^B[W(g(\Theta) - T(x))]$ for all estimators T. Here E^B is the conditional expectation relative to X (or the σ-field B of X). Equivalently to find a T^* such that $E[W(T^* - g(\Theta))] \leqq E[W(T - g(\Theta))]$ as mentioned in the introduction. This problem was solved in [3], and the result will be quoted now for comparison: Assuming that the conditional distribution $F(\cdot \mid x)$ exists, that $W'(t)$ is defined for all $t \neq 0$, and that the following integrals exist, the Bayes estimator $T^*(\cdot)$ is shown to exist and is given as a

solution of the pair of integral inequalities:

$$\int_{\{g(\theta)\geq T(x)\}} W'(g(\theta)-T(x))dF(\theta|x)\geqq \int_{\{g(\theta)<T(x)\}} W'(T(x)-g(\theta))dF(\theta\,|\,x)$$

$$\int_{\{g(\theta)>T(x)\}} W'(g(\theta)-T(x))dF(\theta\,|\,x)\leqq \int_{\{g(\theta)\leq T(x)\}} W'(T(x)-g(\theta))dF(\theta\,|\,x)$$

for almost all x. These inequalities reduce to an equation if $W'(0)=0$.

In what follows, as in section 2, the minimization problem will be relative to $N_W(T-g(\Theta))$ instead of $E[W(T-g(\Theta))]$. This is more general since the existence of the latter expectation implies the existence of the norm but not conversely. The resulting T^* will be termed a Bayes estimator of $g(\Theta)$, relative to $W(\cdot)$. If $W(\cdot)$ satisfies a growth condition [e.g., $M^W=L^W$] then it may be seen that T^* also minimizes the expectation. The general case will be considered below, referring the resulting T^* as the Bayes solution.

There is an associated problem of interest and it cannot be deduced from the above. That will now be treated so that it complements the former. Let $\{D_i\}\subset L^V$ such that $E^B(D_i)=0$, $i=1, 2, \cdots$. The symbols L^W and L^V have the same meaning as before except that the functions are defined on $\mathfrak{X}\times\vartheta$ and the λ is now replaced by ν, the joint distribution of Θ and X, (i. e., D_i: $\mathfrak{X}\times\vartheta\to$real line). If it is a priori known that $E(g(\Theta)D_i(X, \Theta))=\alpha_i$, for all i, then the problem is to find a Bayes estimator $T^*(X)$ of $g(\Theta)$ relative to the general convex loss function $W(\cdot)$. A complete solution is given by

THEOREM 2. *Let W and V be convex (Young's) functions and let L^W and L^V be the associated Orlicz spaces (relative to $\mathfrak{X}\times\vartheta$ and ν), with norms (2) and (4). If $\{D_i\}\subset L^V$ is a family of r. v.'s such that $g(\Theta)\in L^W$, $E^B(D_i)=0$, and $E(g(\Theta)D_i(X, \Theta))=\alpha_i$, $i=1, 2, \cdots$, then the following conclusions hold:*

(a) A Bayes estimator T of $g(\Theta)$ exists only if (i) there is a constant K such that for any finite set $\{D_{i_1}, \cdots, D_{i_n}\}$ of $\{D_i\}$ and any scalars $\{a_1, \cdots, a_n\}$

$$|\sum_{j=1}^{n} a_j\alpha_{i_j}|\leqq K\|\sum_{j=1}^{n} a_jD_{i_j}\|_V\,, \tag{9}$$

(ii) $\{D_i\}\subset M^V$, the closed subspace of L^V spanned by the ν-simple (or bounded) functions, and that for almost all x, $\{D_i(x, \cdot)\}$ is fundamental in the subspace of θ-functions of M^V, relative to $F(\cdot\,|\,x)$. [This last condition is satisfied if $\{D_i\}$ is fundamental in M^V.] On the other hand, if there is a Bayes estimator T of g then (i) always holds whether or not

(ii) is true. $\left[\text{So } M^V=\left\{f: \int_\Omega V(\alpha f)d\nu<\infty,\ f\in L^V,\ \alpha>0 \text{ arbitrary}\right\}.\right]$

(b) If $\{D_i\}\subset M^V$ *is fundamental there exists at most one Bayes estimator of* $g(\Theta)$.

(c) If T *is a Bayes estimator of* $g(\Theta)$ *such that* $N_W(T-g(\Theta))=K_0$ *where* $K_0=\inf\{K>0, K \text{ satisfying } (9)\}$ *and* $W(\cdot)$ *and* $V(\cdot)$ *are continuous then* T *is essentially unique.*

Remark. The various expectations here are relative to the joint distribution of Θ and X, and that the conditional distribution $F(\cdot\,|\,x)$ is assumed to exist. Also compare the conditions here and in theorem 1.

PROOF. (a) The direct part is again easy. If T is a Bayes estimator, let $\bar{T}(X,\Theta)=g(\Theta)-T(X)$ so that $\alpha_j=E(g(\Theta)D_j)=E[E^B(\bar{T}D_j)]=E[\bar{T}D_j]$. From this it follows by Hölder inequality (let $\mathfrak{X}_0=\mathfrak{X}\times\vartheta$)

$$|\sum_{j=1}^n a_j\alpha_{i_j}|=\left|\int_{\mathfrak{X}_0}(\sum_{j=1}^n a_j\bar{T}D_{i_j})d\nu(x,\theta)\right|$$

$$\leqq N_W(\bar{T})\,\|\sum_{j=1}^n a_jD_{i_j}\|_V\,.$$

Setting $K=N_W(\bar{T})$, this becomes (9) without the hypothesis, a(ii).

For the converse all the hypothesis of (a) will be needed. As before, (9) implies the existence of a bounded linear functional G on M^V such that $G(D_i)=\alpha_i$, for all i, and $\|G\|\leqq K$. If $K_0=\inf\{K>0, K \text{ satisfying } (9)\}$, then G can be chosen such that $\|G\|=K_0$. [The same conclusion holds if $\{D_i\}$ in M^V is fundamental.] By ([7], theorem 4) there is a unique $\bar{T}$ in L^W such that

$$\alpha_i=G(D_i)=\int_{\mathfrak{X}_0}\bar{T}(x,\theta)D_i(x,\theta)d\nu(x,\theta)\,,\quad i\geqq 1\,. \tag{10}$$

Since also by hypothesis one has

$$\alpha_i=\int_{\mathfrak{X}_0}g(\theta)D_i(x,\theta)d\nu(x,\theta)\,,\qquad i\geqq 1\,, \tag{11}$$

it follows that

$$\begin{aligned}0&=\int_{\mathfrak{X}_0}[g(\theta)-\bar{T}(x,\theta)]D_i(x,\theta)d\nu,\ \text{all } i\,.\\ &=\int_{\mathfrak{X}}\int_{\vartheta}[g(\theta)-\bar{T}(x,\theta)]D_i(x,\theta)dF(\theta\,|\,x)d\pi\end{aligned} \tag{12}$$

where π is the (marginal) distribution of X. From the condition $E^B(D_i)=0$, it follows that (12) holds whenever $g(\theta)-\bar{T}(x,\theta)=T(x)$, say, where

$T(\cdot)$ is a function of x alone, (and $T \in L^W$) unless $K_0=0$ in which case $\bar{T}=0$, a. e., (and $g=0$ *a fortiori*). For, if $g(\theta)-T(x, \theta)$ depends on θ also on a set of positive $F(\cdot \mid x)$-measure, then by a(ii) [$\{D_i(x, \cdot)\}$ is fundamental], a function $h(\cdot, x)$ of θ exists in M^V (truncate the above suitably and multiply it by the signum of itself) so that on a set H of positive $F(\cdot \mid x)$-measure,

$$\int_H (g(\theta)-\bar{T}(x, \theta))h(\theta, x)dF(\theta \mid x)>0 ,$$

which gives a contradiction to the already established equation (12). [Note that (12) is true if D_i is replaced by any function in the subspace of M^V determined by $\{D_i(x, \cdot)\}$ and this is used in the previous statement.] Hence $\bar{T}(\theta, x)=g(\theta)-T(x)$, and one has

$$K_0=\|G\|=N_W(g(\Theta)-T(X)) . \tag{13}$$

Thus $T(X)$ in L^W is the Bayes estimator of $g(\Theta)$, with risk K_0. This completes both (a) and (b). The other statements being immediate from theorem 1, the proof can be concluded.

Note. The condition (9) and the statement (a) of the above theorem can be imitated from the corresponding one of theorem 1, but it will be more awkward to state. Now the choice of $\{D_i\}$ of the theorem is not entirely easy. A brief indication of it in a special case, which can be extended, is given in ([3], p. 845).

The results of this and the preceding sections are on existence and they do not indicate how an optimal estimator can be constructed. This problem will be examined in the next section.

4. Construction of the optimal estimators

The construction of optimal estimators is a nontrivial task and for the convex $W(\cdot)$ it is somewhat more involved than when $W(x)=|x|^p$, considered in [1]. Even that case was not easy. Due to some technical limitations, some growth conditions will be assumed in contrast to the general character of $W(\cdot)$ of the preceding sections. This however generalizes the above special case considerably.

The (Young's) convex functions W, V to be considered are such that L^W is reflexive. It then follows by ([7], theorem 5) that $M^V=L^W$ *and* $M^V=L^V$ so that W, V are necessarily continuous with continuous derivatives W' and V'. [$W(x)=|x|^p$, $1<p<\infty$ is subsumed. The latter corresponds to the uniform convexity of L^W and L^V, which is more severe than the present assumption. For the present it suffices that $W(2x)\leqq$

$CW(x)$ and $V(2x) \leqq CV(x)$ for large x, (cf., [5])*). Also see ([2], p. 113) and the references there, for various connections dealing with the concepts of the metric geometry here.]

In the following proofs several inequalities appear and constant adjustments are needed. To save some of this trouble it is convenient to normalize W and V as

$$W(1)+V(1)=1\,,$$

so that (from the continuity of W, V) $0<W(1)<1$. This induces a change in some of the previous definitions. They are (cf., [8], p. 173) in equations (4) and (5). Thus

$$(4') \qquad N_W(f)=\inf\left\{k>0,\ \int_\Omega W(f/k)d\mu \leqq W(1)\right\}\,,\ f\in L^W\,,$$

and this gives

$$(5') \qquad W(1)N_W(f) \leqq \|f\|_W \leqq 2N_W(f)\,,$$

where the middle term is defined in (2). This normalization and (4') and (5') will be used hereafter and the norms in both spaces will be $N_W(\cdot)$ and $N_V(\cdot)$. [The reason for not using these before was, first it was not necessary, and secondly it was somewhat inconvenient in the proofs of [7] referred to. However, the results of theorems 1 and 2 are true with this and it is then possible to replace $\|\cdot\|_V$ by $N_V(\cdot)$ without change. It is remarked that this last replacement may not be possible without normalization.] In the following the concept of the derivative of the norm is needed: The norm $\|\cdot\|$ of a B-space $\mathfrak{Y}$ is said to be (*strongly*) *differentiable* at y in $\mathfrak{Y}$, if $\|y+ty_1\|$ is a differentiable function of t for any y_1 in $\mathfrak{Y}$, (i. e., uniformly in y_1).

First the unbiased optimal estimators will be considered and then the Bayesian case and Bayes lower bounds will be presented. With the same notations as before, one has

THEOREM 3. *Let W, V be normalized convex (Young's) functions such that L^W is reflexive. (Its conjugate is L^V.) Suppose that T in L^W is an unbiased estimator of $g(\theta)$ which is optimal at θ_0 (cf. theorem 1), relative to the given functions $\{D_i\}\subset L^V$ with minimal (positive) risk $K=N_W(T-g(\theta_0))$. Let $\{i_j\}$ be a subsequence of the indices $\{i\}$, and $\{a_i\}$ be a sequence of constants such that $(D_i=D_i(\cdot\,,\theta_0))$,*

*) A *sufficient* condition involving W only is: $\limsup_{x\to\infty}[W(2x)/W(x)]\leqq C<\infty$, and $\liminf_{x\to\infty}[W'(x)/W'(ax)]\geqq M_a>1$, for $3/4<a<1$. [This and related matters are given in [7a], in detail.]

$$\lim_{n\to\infty} \frac{|\sum_{j=1}^{n} a_j\alpha_{i_j}(\theta_0)|}{N_V(\sum_{j=1}^{n} a_jD_{i_j})} = K. \tag{14}$$

Then the sequence of functions T_n *defined by*

$$T_n = \frac{\sum_{j=1}^{n} a_j\alpha_{i_j}(\theta_0)}{N_V(\sum_{j=1}^{n} a_jD_{i_j})} V'\left(\frac{\sum_{j=1}^{n} a_jD_{i_j}}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}\right) sgn\,(\sum_{j=1}^{n} a_jD_{i_j}) \tag{15}$$

has the properties: (*i*) $T_n \in L^W$, *and* (*ii*) $T_n \to T$ *in* L^W*-norm as* $n\to\infty$, *where* T *is the unbiased estimator of* $g(\theta)$ *that is optimal at* θ_0.

Remark. Note that the case $K=0$, which is excluded, is trivial. Recall that $g(\theta_0)=0$ by convension, and that μ, Ω can be interchanged by λ, $\mathfrak{X}$.

PROOF. The proof depends on the following two important results of V. Šmulian (cf., [4], p. 472 and the original reference there) which are stated for convenience. "(I) The norm of a B-space $\mathfrak{Y}$ is (strongly) differentiable at y in $\mathfrak{Y}$ if and only if every sequence $\{y_n^*\}$ in the unit ball of the conjugate space $\mathfrak{Y}^*$ with the property that $y_n^*(y)\to\| y \|$, is itself (strongly) convergent. (II) Similar theorem is true for $\mathfrak{Y}^*$ if y and y^* are interchanged in (I)." (Šmulian has another interesting result in the same paper which is also of interest as it throws some light on l_1 in this context: (III) If the B-space $\mathfrak{Y}$ is weakly complete and weak convergence coincides with the strong convergence, then the weak derivative of the norm $\| \cdot \|$ at each y_0, $\| y_0 \|=1$, implies the strong derivative at the same point.—This follows from (I) above.) Part of the following proof is patterned after the L^p-theory of [1]. Now the details.

To simplify writing let

$$C_n = \frac{\sum_{j=1}^{n} a_j\alpha_{i_j}(\theta_0)}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}, \quad Y_n = \frac{\sum_{j=1}^{n} a_jD_{i_j}}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}, \text{ and } s_n = sgn(\sum_{j=1}^{n} a_jD_{i_j}).$$

To prove (i) consider any $\tilde{T}_n$ in L^W given by

$$\tilde{T}_n = C_nX_n, \quad N_W(X_n)=1.$$

Since $K>0$, from some n on $C_n\neq 0$, so that X_n is well-defined. By Hölder inequality, for Y_n in L^V defined above $N_V(Y_n)=1$, one has

$$\int_\Omega |X_n||Y_n|d\mu \leqq 1, \tag{16}$$

and there is equality in (16) if and only if (a) X_nY_n has a constant sign and (b) either $|X_n|=V'(|Y_n|)$ or $|Y_n|=W'(|X_n|)$, a. e., since in the definition of the norm in (4′) with the particular W, V (satisfying the growth restriction) there is equality after the integral. (cf., [8], p. 175.) Thus there is equality in (16) if $|X_n|=V'(|Y_n|)$, a.e. Substituting this in $\tilde{T}_n$, it follows that $\tilde{T}_n$ in L^W is precisely T_n of the theorem and moreover $N_W(V'(|Y_n|)=1$, so that $N_W(T_n)=|C_n|\to K$, as $n\to\infty$. This finishes (i). It remains to show that $T_n\to T$ is norm. (It is remarked that the equality conditions in Hölder inequality, without normalization, are complicated.)

To prove (ii), which will be accomplished in two stages, first the norm differentiability of L^W and L^V will be established. By duality, it suffices to consider one of them, say L^W. Since L^W is reflexive, as noted before, $M^V=L^W$ and $M^V=L^V$ and W' and V' exist as continuous functions. (i. e., in the terminology of [5], W, V satisfy the so-called Δ_2-condition.) From this and ([5], p. 188) it follows that the norm is (strongly) differentiable at every point except the origin. [Before noting the result in [5], the author has established this in a different way using some results of R. C. James (*Trans. Am. Math. Soc.*, 1947) and [7]. It seems to have other interesting implications on the differentiability of norms, but they are not needed here.] With this the proof can be completed as follows.

Let $r_n=sgn(\sum_{j=1}^{n} a_j\alpha_{i_j})$, so that one has

$$F_n(T)=\int_\Omega TY_nr_nd\mu=|C_n|\to K=N_W(T), \tag{17}$$

as $n\to\infty$. It is clear that $||F_n||=1$. Since $T\neq 0$ and the norm in L^W is differentiable at T, by the preceding paragraph, it follows, by Šmulian's theorem (I), that $F_n\to F$ in norm and that $||F||=1$. So $F(T)=K$ and by the representation theorem in [7], there is a Y_0 in L^V, and $N_V(Y_0)=||F||$,

$$F(T)=\int_\Omega TY_0d\mu=N_W(T). \tag{18}$$

Since there is equality (in the Hölder inequality) in (18), repeating the by now familiar argument, one has

$$Y_0=V'\left(\frac{|T|}{K}\right)sgn(T). \tag{19}$$

But this means that $Y_nr_n\to Y_0$ in L^V-norm. With this the final step can

now be completed.

Consider again the linear functionals F_n on L^W, defined above.

$$F_n(T_n)=\int_\Omega T_n Y_n r_n d\mu=|C_n|\int_\Omega Y_n V'(|Y_n|)s_n d\mu$$

(20)

$$=|C_n|\int_\Omega |Y_n| V'(|Y_n|)d\mu=|C_n|\to K,$$

as $n\to\infty$. Here T_n are the functions given by (15), and the value of the integral $=1$ (found after (16)) is used. In the preceding paragraph it was shown that $F_n\to F$ in norm, and $F_n(T)\to F(T)=K$. So,

$$|F(T_n)-K|\leqq \|F-F_n\| N_W(T_n)+|F_n(T_n)-K|\to 0, \tag{21}$$

as $n\to\infty$ by (20) and the fact that $\{N_W(T_n)\}$ is bounded, being a convergent sequence. (21) may be written as, by letting $T_n'=(T_n/|C_n|)$ so that $N_W(T_n')=1$,

$$\lim_{n\to\infty} F(T_n')=\lim_{n\to\infty}\int_\Omega T_n' Y_0 d\mu=1=N_V(Y_0). \tag{22}$$

Since $Y_0\neq 0$, the norm in L^V is differentiable at Y_0 and the second theorem of Šmulian is applicable here. From this it follows that $T_n'\to T_0$ is norm. Hence $T_n\to KT_0$ since $|C_n|\to K$ by (14). Thus (22) yields

$$\int_\Omega Y_0 T_0 d\mu=N_V(Y_0)=1. \tag{23}$$

From the equality in (23), for the same reasons as in (19), one has

$$T_0=W'(|Y_0|)\,sgn\,(Y_0). \tag{23'}$$

Using (19) for Y_0 and noting that V' and W' are inverse functions to each other, it follows immediately from (23') that $T=T_0K$, a. e., or $T_n\to T$ in L^W-norm. This completes the proof.

Note. The above proof shows that the theorem is true if W, V are such that both the norms of L^W and L^V are (strongly) differentiable. But there must be other conditions (as in theorem 1) for which the result holds.

The foregoing proof has the following consequence (cf. (19)).

COROLLARY. *Under the hypothesis of the theorem $Y_n r_n\to Y_0$ in L^V-norm where $Y_0=V'(|T|/K)\,sgn\,(T)$, or $T=K\cdot W'(|Y_0|)\,sgn\,(Y_0)$, a. e.*

The construction of the Bayes estimator, whose existence is assured by theorem 2, is given by the following.

THEOREM 4. *Let W, V be normalized (Young's) convex functions*

such that L^W is reflexive. If T is the Bayes estimator of $g(\Theta)$ relative to the functions $\{D_i\}\subset L^V$, as in theorem 2, with (positive) Bayes risk $K=N_W(T-g)$, let $\{i_j\}$ be a subsequence of the indices $\{i\}$, and $\{a_i\}$ be a sequence of constants such that

$$\lim_{n\to\infty}\frac{|\sum_{j=1}^{n} a_j\alpha_{i_j}|}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}=K. \tag{24}$$

Then the sequence of functions $\bar{T}_n$ defined by

$$\bar{T}_n=\frac{\sum_{j=1}^{n} a_j\alpha_{i_j}}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}\,V'\left(\frac{\sum_{j=1}^{n} a_jD_{i_j}}{N_V(\sum_{j=1}^{n} a_jD_{i_j})}\right)sgn\,(\sum_{j=1}^{n} a_jD_{i_j}) \tag{25}$$

has the properties: (i) $\bar{T}_n\in L^W$ and (ii) $\bar{T}_n\to\bar{T}=g(\Theta)-T$, in L^W-norm.

Remark. The integrals in the proof will be relative to the joint distibution $\nu(x,\theta)$, and the actual details are similar to those of theorem 3. The proof that $\bar{T}_n\to\bar{T}$ being the same, the special hypothesis on $\{D_i\}$ now ensures (as shown in theorem 2) that $g(\Theta)-\bar{T}$ is a function of x alone and that it must be T.

From this result, a class of (Bayes) lower bounds can be given. A sample result is the following.

THEOREM 5. *Let W, V be (arbitrary Young's) convex functions, and L^W and L^V be the corresponding Orlicz spaces. Suppose $\{D_i\}\subset L^V$ are such that $E^B(D_i)=0$, $i\geqq 1$, where $B(=B(X))$ is the σ-field generated by X, and $g(\Theta)$ is the parameter in L^W satisfying $E(gD_i)=\alpha_i$, $i\geqq 1$. If T is a Bayes estimator of $g(\Theta)$ relative to W, with risk K, then for any set of constants $\{a_i\}$, the following class of (Bayes) lower bounds obtain:*

$$K=N_W(T(X)-g(\Theta))\geqq\frac{|\sum_{j=1}^{n} a_j\alpha_{i_j}|}{\|\sum_{j=1}^{n} a_jD_{i_j}\|_V}. \tag{26}$$

The best (i. e. largest) lower bounds are obtained by maximizing the right side of (26) relative to $\{a_i\}$ and n.

This is an immediate consequence of the direct part of theorem 2(a). That the bound is achieved for a particular set of $\{a_i\}$ and $\{D_i\}$, if W, V are restricted, is a consequence of theorem 4. A special case of the above result was given in ([3], theorem 5) if $n=1$, $W(x)=|x|^p$, $p\geqq 1$. Thus one may specialize this result to get various interesting

lower bounds for different n and different forms of the loss function $W(\cdot)$ of the theorem.

5. Final remarks

In some problems, with a certain prior information, the (locally) optimal unbiased estimators are of interest. Besides this, however it admits relatively deep mathematical tools yielding a complete solution. (E. g., theorems 1 and 2 are in the final forms.) But from the point of view of (practical) interpretation Bayes estimators may be more satisfactory in many cases, e. g., when the prior information is available in the form of the prior distribution of the parameter. However, the points of view expressed in theorems 1 and 2, make it plain that the problem of (locally) optimal unbiased estimation and the Bayes estimation subject to restraints *on the parameter*, are in a certain sense, "dual" to each other. Thus, whatever be the relative merits of unbiased and Bayes estimations, the results in one can suggest the corresponding results and extensions in the other. From the technical point of view at least, it seems advantageous to study both problems in their own right.

The reader may also be interested in [4a] which treats some problems in the spirit of this paper.

CARNEGIE INSTITUTE OF TECHNOLOGY

REFERENCES

[A] T. Andô, "Linear functionals on Orlicz spaces," *Nieuw Arch. v. Wisk.* (3), 8 (1960), 1-16.

[1] E. W. Barankin, "Locally best unbiased estimates," *Ann. Math. Statist.*, 20 (1949), 477-501.

[2] M. M. Day, *Normed Linear Spaces*, Springer-Verlag, Berlin, 1958.

[3] M. H. DeGroot and M. M. Rao, "Bayes estimation with convex loss," *Ann. Math. Statist.*, 34 (1963), 839-846.

[4] N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory Interscience*, New York, 1952.

[4a] K. Isii, "Inequalities of the types of Chebyshev and Cramér-Rao and mathematical programming," *Ann. Inst. Stat. Math.*, 16 (1964), 277-293.

[5] M. A. Kransnosel'skii and Ya. B. Rutickii, *Convex Functions and Orlicz Spaces*, (Translation), P. Noordhoff Ltd., Groningen, 1961.

[6] M. M. Rao, "Lower bounds for risk functions in estimation," *Proc. Nat. Acad. Sci.* (U.S.A.), 45 (1959), 1168-1171.

[7] M. M. Rao, "Linear functionals on Orlicz spaces," *Nieuw Arch. v. Wisk.* (3), 12 (1964), 77-98.

[7a] M. M. Rao, "Smoothness of Orlicz spaces," *Indag. Math.*, 27 (1965), (to appear).

[8] A. Zygmund, *Trigonometric Series, Vol. I*, (2nd ed.), Cambridge University Press, 1959.