ON AN ASPECT OF OPTIMAL NONLINEAR ESTIMATION

ERIC B. HALL
Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275

GARY L. WISE
Department of Statistics
University of California, Berkeley
Berkeley, California 94720

ABSTRACT

This paper considers the problem of optimal estimation of a random variable X based on an observation denoted by a random vector Y. A mild restriction on the regular conditional distribution function of X given \( \sigma(Y) \) is presented which ensures that

\[
E[\Phi(X-g(Y))] \text{ is minimized for any cost function } \Phi \text{ that is nonnegative, even, and convex.}
\]

Further, we show that given any real valued Borel measurable function there exist random variables X and Y, possessing a joint density function, so that the chosen function is the optimal estimator, with respect to any nonnegative, even, convex cost function \( \Phi \), of the random variable X as a function of the random variable Y. Finally, the estimation of X via a random variable measurable with respect to a given \( \sigma \)-subalgebra is treated.

INTRODUCTION

In this paper we consider the problem of estimation with respect to nontraditional cost functions. In an estimation problem one is often confronted with two concerns in choosing a cost function: the concern that the cost function adequately reflects the cost one wishes to associate with an error, and the concern that the cost function results in a problem which one finds to be mathematically tractable. Traditional cost functions, such as the extremely popular mean square error cost function, are usually chosen solely on the basis of the second of the above two concerns. As a result, the fidelity demands of the specific problem under consideration are rarely relied upon, and, in fact, are often not even considered, when determining the cost function which will be used. This sacrifice of suitability for mathematical ease in the choice of a cost function should be the cause of some concern since the traditional choices are unsuitable for many problems in estimation.

This paper lessens this problem by extending the domain of mathematical tractability to include many cost functions which, even though pertinent to the subjective demands of many problems, have in the past been excluded from consideration.

**DEVELOPMENT**

In 1958, Seymour Sherman published a paper entitled “Non-Mean-Square Error Criteria” [1] in which he proposed conditions on a conditional distribution which would allow for the simultaneous minimization of a large family of cost functions. Sherman’s effort, although ambitious and widely quoted, lacked the requisite attention to mathematical details and ignored several necessary measure theoretic subtleties. For example, Sherman’s use of the first lemma in [1] results in two integrals which are undefined without the additional assumption that the conditional distribution function under consideration is regular, and, further, [1] incorrectly states that this same lemma holds for any distribution function that is absolutely continuous.

Several subsequent efforts to provide a correct development of the proposition in [1] also proved inadequate. For example, Viterbi [2, pp.308–310], Sakrison [3, pp.10–12], and Van Trees [4, p.61] attempted proofs using integration by parts even though the conditions placed on the cost function by the aforementioned authors were insufficient to allow such a method to be used. After establishing some notation and reviewing some definitions, we will present an example which illustrates this difficulty.

Throughout this paper, almost everywhere (a.e.) will be with respect to Lebesgue measure unless noted otherwise, for a topological space T, we will let \( \mathcal{B}(T) \) denote the Borel subsets of T, and \( 1_S(\cdot) \) will denote the indicator function of the set S. Also, recall that a probability distribution function \( F: \mathbb{R} \to [0, 1] \) is said to be unimodal about \( \gamma \in \mathbb{R} \) if \( F \) is convex on \((-\infty, \gamma) \) and concave on \((\gamma, \infty) \), and a probability distribution function \( F: \mathbb{R} \to [0, 1] \) is said to be symmetric if for all real \( x \), \( F(x) = 1 - \lim_{h \downarrow 0} F(-x-h) \). If \( k \) is a positive integer and \( Y_1, \ldots, Y_k \) are k random variables defined on a common probability space, the random vector \( Y = [Y_1, \ldots, Y_k]^{T} \) induces a probability measure on \( \mathcal{B}(\mathbb{R}^k) \); we will denote this resulting measure by the notation \( \mu_Y \). Finally, we recall that for a random variable \( X \) and a \( \sigma \)-subalgebra \( \mathcal{A} \), a regular conditional distribution function \( F: \mathbb{R} \times \Omega \to [0, 1] \) always exists [5, pp.263-264]; such a function is characterized by the following two conditions: for each \( \omega \in \Omega \), \( F(\cdot, \omega) \) is a probability distribution function, and, for each \( x \in \mathbb{R} \), \( F(x, \omega) = P(X \leq \omega | \mathcal{A})(\omega) \) a.s.

Consider two random variables \( X \) and \( Y \) defined on some probability space \((\Omega, \mathcal{S}, P)\) so that \( X \) and \( Y \) are mutually Gaussian random variables with zero means, unit variances, and correlation coefficient \( \rho \). Recall that \( E[X|Y=y] = \rho y \) a.e. Let \( f: [0,1] \to [0,1] \) be the
familiar Cantor–Lebesgue function defined in [6, p.351], and let a cost function 
\( \Phi: \mathbb{R} \to [0, \infty) \) be defined as follows:
\[
\Phi(x) = \begin{cases} 
\Phi(-x) & \text{if } x < 0 \\
(n-1) + f(x-n+1) & \text{if } x \in [n-1,n); \ n \in \mathbb{N}.
\end{cases}
\]
Clearly, \( \Phi(\cdot) \) is even. Further, the well known properties of \( f \) (given in [6]) immediately imply that the cost function \( \Phi \) is also nonnegative, nondecreasing on \([0, \infty)\), continuous, and singular (i.e. its derivative vanishes a.e.) The usual attempt to prove Theorem 1 using integration by parts is based upon a claim that
\[
\mathbb{E}[\Phi(X-\alpha) | Y=y] - \mathbb{E}[\Phi(X-\rho y) | Y=y] = \lim_{x \to \infty} \Phi(z)u(z) \bigg|_0^\infty - \int_0^\infty \frac{d\Phi(t)}{dt} u(t) \, dt
\]
where \( u(t) \) is defined by
\[
u(t) = \int_0^\infty \left[ g(s + t + \rho y | Y=y) - g(s - t + \rho y | Y=y) \right] \, ds,
\]
where \( g(x|Y=y) \) is a conditional Gaussian density function of \( X \) given \( Y = y \), where \( y \) is some fixed real number, and where \( \alpha > \rho y \). The first term above,
\[
\lim_{x \to \infty} \Phi(z)u(z) \bigg|_0^\infty = \lim_{x \to \infty} \Phi(x)u(x) - \Phi(0)u(0),
\]
is equal to zero in this case and equals zero in the indicated references due to restrictions placed on the conditional density. The problem which now arises is that the second term,
\[
\int_0^\infty \frac{d\Phi(t)}{dt} u(t) \, dt,
\]
is also equal to zero since \( \frac{d\Phi(t)}{dt} = 0 \) a.e. Thus, if the above claim were true then it must follow that, for this example, we have \( \mathbb{E}[\Phi(X-\alpha) | Y=y] = \mathbb{E}[\Phi(X-\rho y) | Y=y] \), or that \( \mathbb{E}[\Phi(X-\alpha) - \Phi(X-\rho y) | Y=y] = 0 \) for any choice of \( \alpha > \rho y \). This last equation, which for almost all \( \alpha \) is false, yields a contradiction to the claim that was made earlier. Furthermore, the continuity of the cost function and the conditional density implies that no specious appeal to atomic measures will salvage a method based on integration by parts. Thus, we see that, due to the great generality yielded by Sherman’s conditions on the allowable cost functions, an attempt to prove Theorem 1 via a method which utilizes the previous claim is futile, and hence the developments of Sherman’s proposition given in [2, pp.308–310], [3, pp.10–12], and [4, p.61] are incorrect.

In [7] we put Sherman’s result on a firm mathematical basis and explored several extensions and practical consequences of his proposal. Theorem 1, stated below and
proved in [7], provides a correct statement of Sherman’s original proposal. Here then is
the correct version of “Sherman’s theorem.”

**Theorem 1:** Let $k \in \mathbb{N}$, $(\Omega, \mathcal{S}, P)$ be a probability space, and $X, Y_1, \ldots, Y_k$ be random
variables defined on $(\Omega, \mathcal{S}, P)$, with $X$ integrable. Let $M: \mathbb{R}^k \to \mathbb{R}$ be a Borel measurable
function such that $M[Y_1(\omega), \ldots, Y_k(\omega)] = E[X \mid Y_1, \ldots, Y_k](\omega)$ a.s., and assume that
there exists a regular conditional distribution function of $X$ conditioned on $\sigma(Y_1, \ldots, Y_k)$,
$F: \mathbb{R} \times \Omega \to [0,1]$, such that $F(x + M[Y_1(\omega), \ldots, Y_k(\omega)], \omega)$, as a function of $x$ with $\omega$
fixed, is unimodal about the origin and symmetric. Then $M[Y_1, \ldots, Y_k]$ minimizes the
quantity $E[\Phi(X - f(Y_1, \ldots, Y_k))]$ over all Borel measurable functions $f: \mathbb{R}^k \to \mathbb{R}$ where
$\Phi: \mathbb{R} \to [0, \infty)$ is even and nondecreasing on $[0, \infty)$.

Thus, Sherman’s result as correctly stated in Theorem 1 requires a regular conditional
distribution function that, when properly shifted, is symmetric and unimodal about the
origin and a cost function that is nonnegative, even, and nondecreasing to the right of the
origin. It is easy to see that if in Theorem 1 we let $k=1$ and $X$ and $Y$ be mutually Gaussian
random variables then the resulting regular conditional distribution function is symmetric
and unimodal about a version of $E[X \mid Y](\omega)$ for any fixed $\omega$. This special case explains
why Sherman’s result is often invoked to add a token claim of generality to papers that only
consider Gaussian distributions. When one attempts to venture outside this somewhat
limited arena, however, the conditions which Theorem 1 places on the regular conditional
distribution function immediately begin to feel overly restrictive. After all, how
comfortable should we be with the assumption that the regular conditional distribution
function under consideration is unimodal about the conditional mean? The conditions on
the cost function, on the other hand, are extremely nonrestrictive and, in fact, allow for
many interesting, albeit impractical, choices. For example, the cost function given by

$$\Phi(x) = \int_0^{|x|} \mathbb{1}_C(t) \, dt,$$

where $C$ denotes a Cantor subset of $[0, \infty)$ of positive Lebesgue measure, satisfies the
conditions of Theorem 1. This imbalance suggests the possibility of lessening the
restrictions on the regular conditional distribution function by perhaps slightly increasing
the restrictions imposed on the cost function. The following lemma will allow us to present
such a result.
Lemma 2: Assume that $F$ is a symmetric probability distribution function and that $\Phi: \mathbb{R} \to [0, \infty)$ is even and convex. Then $\int \Phi(x) \, dF(x) \leq \int \Phi(x-\alpha) \, dF(x)$ for all $\alpha \in \mathbb{R}$.

Proof: Since $\Phi$ is convex, $\Phi(x) \leq \frac{1}{2} \Phi(x-\alpha) + \frac{1}{2} \Phi(x+\alpha)$. Further, since $F$ is symmetric and $\Phi$ is even, $\int \Phi(x + \alpha) \, dF(x) = \int \Phi(x - \alpha) \, dF(x)$. Thus, it follows that

$$\int \Phi(x) \, dF(x) \leq \frac{1}{2} \int \Phi(x - \alpha) \, dF(x) + \frac{1}{2} \int \Phi(x + \alpha) \, dF(x) = \int \Phi(x - \alpha) \, dF(x).$$

Q.E.D.

As we will now show, Lemma 2 allows a result similar to that of Theorem 1 to be stated for a much less restrictive family of regular conditional distribution functions by slightly restricting the family of allowable cost functions. In particular, we will be able to drop the restriction that the conditional distribution function be unimodal by requiring that the cost function, in addition to the previous restrictions, also be convex. Notice that requiring the cost function to be even and convex implies that it is also nondecreasing to the right of the origin. Also, notice that a close examination of the proof of Lemma 2 reveals that convexity appears to be a slightly more restrictive condition than is needed. In fact, all that is required is that the cost function $\Phi$ be midpoint convex, i.e. that $\Phi((x+y)/2) \leq (\Phi(x)+\Phi(y))/2$ for all $x, y \in \mathbb{R}$. Although, in general, midpoint convexity does not imply convexity, midpoint convexity and monotonicity do imply convexity. Thus, we do not decrease the family of allowable cost functions by stipulating convexity in place of midpoint convexity.

The following theorem provides the same conclusion as Theorem 1 while greatly relaxing the conditions which Theorem 1 placed on the regular conditional distribution function. Although the family of allowable cost functions is decreased from those considered in Theorem 1, the family under consideration below is still quite large and diverse.

Theorem 3: Let $k \in \mathbb{N}$, $(\Omega, \mathcal{S}, P)$ be a probability space, and $X, Y_1, \ldots, Y_k$ be random variables defined on $(\Omega, \mathcal{S}, P)$, with $X$ integrable. Let $M: \mathbb{R}^k \to \mathbb{R}$ be a Borel measurable function such that $M[Y_1(\omega), \ldots, Y_k(\omega)] = E[X | Y_1, \ldots, Y_k](\omega)$ a.s., and assume that there exists a regular conditional distribution function of $X$ conditioned on $\sigma(Y_1, \ldots, Y_k)$, $F: \mathbb{R} \times \Omega \to [0,1]$, such that $F(x+M[Y_1(\omega), \ldots, Y_k(\omega)],\omega)$, as a function of $x$ with $\omega$
fixed, is symmetric. Then $M[Y_1, \ldots, Y_k]$ minimizes the quantity $E[\Phi(X-f(Y_1, \ldots, Y_k))]$ over all Borel measurable functions $f:R^k \to R$ where $\Phi:R \to [0, \infty)$ is even and convex.

**Proof:** Lemma 2 implies that for each fixed $\omega$ and for $\alpha \in R$,

$$\int_{R} \Phi(x) \, dF(x+M(Y_1(\omega), \ldots, Y_k(\omega)), \omega) \leq \int_{R} \Phi(x-\alpha) \, dF(x+M(Y_1(\omega), \ldots, Y_k(\omega)), \omega).$$

Changing variables yields

$$\int_{R} \Phi(x-M(Y_1(\omega), \ldots, Y_k(\omega))) \, dF(x, \omega) \leq \int_{R} \Phi(x-\alpha-M(Y_1(\omega), \ldots, Y_k(\omega))) \, dF(x, \omega).$$

Let $g:R^k \to R$ be a Borel measurable function by which $X$ is to be estimated and $E[\Phi(X-g(Y_1, \ldots, Y_k))]$ minimized. Note that $E[\Phi(X-g(Y_1, \ldots, Y_k))] = E[ E[ \Phi(X-g(Y_1, \ldots, Y_k)) | \sigma(Y_1, \ldots, Y_k) ] ]$. From the preceding inequality and [8, p.79], the inner expectation, and thus the above expression, is minimized when $g(Y_1, \ldots, Y_k) = M(Y_1, \ldots, Y_k)$.

Q.E.D.

We will next present a useful corollary to Theorem 3. Let $k \in N$, $(\Omega, S, P)$ be a probability space, $X$ be a random variable defined on $(\Omega, S, P)$, and $Y$ be a random vector defined on $(\Omega, S, P)$ taking values in $R^k$. Recall that $F(x|Y=y)$ is said to be a regular conditional distribution function for $X$ given $Y=y$ if for each fixed $y \in R$, $F(x|Y=y)$ is a probability distribution function as a function of $x$, and for each fixed $x \in R$, $F(x|Y=y)$ is a version of the regression function $E[I_{(-\infty, x]}(X)|Y=y]$. Further, recall that a regular conditional distribution function for $X$ given $Y=y$ always exists [cf. 9, pp.372–376]. The next corollary, which follows straightforwardly from Theorem 3, removes the need to work on the underlying probability space.

**Corollary 4:** Let $k \in N$, $(\Omega, S, P)$ be a probability space, $X$ be an integrable random variable defined on $(\Omega, S, P)$, $Y$ be a random vector defined on $(\Omega, S, P)$ taking values in $R^k$, and $M:R^k \to R$ be any Borel measurable function equal a.e.[$\mu_Y$] to $E[X|Y=y]$. Further, assume that, as a function of $x$ with $y$ fixed, a regular conditional distribution function of $X$ given $Y=y$, denoted by $F(x|Y=y)$, is such that $F(x+M[y]|Y=y)$ is symmetric. Then $g(y) = M[y]$ minimizes $E[\Phi(X-g(Y))]$ over all Borel measurable functions $g:R^k \to R$ where $\Phi:R \to [0,\infty)$ is even and convex.
Theorem 3 and Corollary 4 thus offer the trade off that was suggested earlier. In particular, by slightly increasing the restrictions on the family of allowable cost functions, we may drop the requirement that the regular conditional distribution function be unimodal, which seems overly restrictive from a utilitarian viewpoint. Further, the family of cost functions considered by Theorem 3 and Corollary 4, although smaller than the family which Sherman considered, is still quite large and, in fact, includes the $L^p$ norm for $1 \leq p < \infty$.

The following example illustrates the usefulness of Theorem 3 and Corollary 4 and shows how these results may be applied to non-Gaussian distributions. In particular, given any real valued Borel measurable function $g(\cdot)$, we show that there exists a random variable $X$ and a random variable $Y$, possessing a joint density function, so that $E[\Phi(X-h(Y))]$ is minimized when $h(Y) = g(Y)$ for any cost function $\Phi$ that is nonnegative, even, and convex.

**Example:** Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be Borel measurable and define

$$f(x, y) = \frac{1}{8} \exp(-\exp(yl) |x - g(y) + K|) + \frac{1}{8} \exp(-\exp(yl) |x - g(y) - K|)$$

where $K$ is some fixed real number. Note that $f(x, y)$ is a joint probability density function since

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) \, dx \, dy = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{8} \exp(-\exp(yl) |x - g(y) + K|) + \frac{1}{8} \exp(-\exp(yl) |x - g(y) - K|) \, dx \, dy$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{1}{4} \exp(-\exp(yl) |zl|) \, dz \, dy = \int_{\mathbb{R}} \frac{1}{2} \exp(-yl) \, dy = 1.$$

Let $X$ and $Y$ be random variables such that the pair $(X, Y)$ has a joint density function given by $f(x, y)$. Notice from the above calculation that the second marginal density of $f(x, y)$ is given by $f_Y(y) = \frac{1}{2} \exp(-yl)$. Also, notice that $f(x + g(y), y)$, as a function of $x$ with $y$ fixed, is even. Recall that a version of $E[X|Y=y]$ is given by

$$\int_{\mathbb{R}} x \frac{f(x, y)}{f_Y(y)} \, dx.$$

Fixing this version and using the above expression for $f_Y(y)$ implies that $E[X|Y=y]$

$$= 2 \exp(yl) \int_{\mathbb{R}} \frac{1}{8} (\exp(-\exp(yl) |x - g(y) + K|) + \exp(-\exp(yl) |x - g(y) - K|)) \, dx.$$
Since \( f(x, y) \), as a function of \( x \) for \( y \) fixed, is even about \( g(w) \), it follows that the conditional density function \( f_{y}(y) \) shares this same property. Thus, it is easy to see that the associated regular conditional distribution function, when properly shifted, is symmetric (and not unimodal if \( K \) is nonzero).

Corollary 4 may thus be applied to see that \( h(y) = g(y) \), which we recall was an arbitrary Borel measurable function, minimizes \( E[X-h(Y)] \) over all Borel measurable functions \( h: \mathbb{R} \rightarrow \mathbb{R} \) where \( \Phi: \mathbb{R} \rightarrow [0, \infty) \) is even and convex. Notice that this example illustrates the applicability of Theorem 3 in a situation where Theorem 1 would not apply, and, in addition, demonstrates the applicability of these results to non-Gaussian distributions. Notice further that this example also points out that the existence of a joint density function in no way guarantees that a regression function will obey any regularity property other than Borel measurability.

**THE GENERAL CASE**

The preceding development was concerned with an attempt to estimate the integrable random variable \( X \) based on a Borel measurable function of the random variables \( Y_{1}, \ldots, Y_{k} \), where \( k \) is a positive integer. In this case, our estimate was a \( \sigma(Y_{1}, \ldots, Y_{k}) \)-measurable random variable. In many cases, we might wish to estimate \( X \) by a random variable which is measurable with respect to some other \( \sigma \)-algebra. For example, consider the case where \( \{Y_{t}: t \in [0,1]\} \) is a random process and we wish to estimate \( X \) via a random variable which is \( \sigma([0,1]) \)-measurable. Also, consider the case where \( H \) is a real, separable Hilbert space and \( Z \) is an \( H \)-valued random variable; here we might wish to estimate \( X \) via a \( \sigma(Z) \)-measurable random variable. In the general case, \( Z \) could be a random object; that is, a random variable taking values in a measurable space \((G, \mathcal{G})\), and we would be interested in estimating \( X \) via a random variable which is measurable with respect to \( \sigma(Z) = Z^{-1}(\mathcal{G}) \).

The following theorem addresses the estimation of \( X \) via a random variable which is measurable with respect to a \( \sigma \)-subalgebra of \( S \).

**Theorem 5:** Let \((\Omega, S, P)\) be a probability space, \( \mathcal{A} \) be a \( \sigma \)-subalgebra of \( S \), and \( X \) be a random variable defined on \((\Omega, S, P)\) such that \( X \) is integrable. For each \( \omega \in \Omega \), let \( M(\omega) = E[X|\mathcal{A}] (\omega) \), and assume that there exists a regular conditional distribution function
of \(X\) conditioned on \(A\), \(F:R \times \Omega \to [0,1]\), such that \(F(x+M(\omega),\omega)\), as a function of \(x\) with \(\omega\) fixed, is symmetric. Then \(M\) minimizes the quantity \(E[\Phi(X-\hat{X})]\) over all \(A\)-measurable random variables \(\hat{X}\), where \(\Phi:R \to [0, \infty)\) is even and convex.

**Proof:** Lemma 2 implies that for each fixed \(\omega\) and for \(\alpha \in R\),
\[
\int_R \Phi(x) \, dF(x+M(\omega),\omega) \leq \int_R \Phi(x-\alpha) \, dF(x+M(\omega),\omega).
\]
Changing variables yields
\[
\int_R \Phi(x-M(\omega)) \, dF(x,\omega) \leq \int_R \Phi(x-\alpha-M(\omega)) \, dF(x,\omega).
\]
Let \(\hat{X}\) be an \(A\)-measurable random variable by which \(X\) is to be estimated and \(E[\Phi(X-\hat{X})]\) minimized. Note that \(E[\Phi(X-\hat{X})] = E[ E[ \Phi(X-\hat{X}) | A ]]\). From the preceding inequality and [8, p.79], the inner expectation, and thus the above expression, is minimized when \(\hat{X} = M\).

**Q.E.D.**

**CONCLUSION**

Thus, in conclusion, this paper has considered the problem of optimal estimation of a random variable \(X\) based on an observation denoted by a random vector \(Y\). We gave a mild restriction on the regular conditional distribution function of \(X\) given \(\sigma(Y)\) which ensured that \(E[\Phi(X-g(Y))]\) is minimized for any cost function \(\Phi\) that is nonnegative, even, and convex. Further, we showed that given *any* real valued Borel measurable function there exist random variables \(X\) and \(Y\), possessing a joint density function, so that the chosen function is the optimal estimator, with respect to any of the cost functions described above, of the random variable \(X\) based on the random variable \(Y\). The results were then extended to estimation of \(X\) based upon a random variable which was measurable with respect to a given \(\sigma\)-subalgebra.

Finally, we hope that the results presented in this paper will contribute to a concept of estimation theory based on more meaningful cost functions than the overused and often abused mean square cost function. Despite the popularity of mean square estimation, its use is often a signal that the fidelity considerations of the problem have been ignored which, of course, should cast some doubt on the quality of the estimate. For those who still insist upon using mean square estimation techniques we note in closing that, even for bounded random variables, conditional expectation fails to minimize mean square error [10].
ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research under Grant AFOSR-86-0026.

REFERENCES


