This Final Technical Report constitutes a summary of the research performed under Grant AFOSR-86-0026 during the period November 1, 1986 through April 30, 1990. First we present a list of the personnel involved in the research effort. Then in the following section we present a brief summary of the research results that have been achieved. Each of these results is well documented in technical articles, and references to these articles are made in the summary of the research results.
INTRODUCTION

This Final Technical Report constitutes a summary of the research performed under Grant AFOSR-86-0026 during the period November 1, 1986 through April 30, 1990. First we present a list of the personnel involved in the research effort. Then in the following section we present a brief summary of the research results that have been achieved. Each of these results is well documented in technical articles, and references to these articles are made in the summary of the research results.
PERSONNEL

Principal Investigator:

Gary L. Wise

Postdoctoral Research Associates:

John M. Morrison

Alan E. Wessel

Graduate Research Associate:

Eric B. Hall
A SURVEY OF RESULTS

In this final technical report we will briefly comment upon our research accomplishments sponsored by the Grant AFOSR-86-026. Much of our work during this period was concerned with various aspects of estimation theory. Additional work was done in the areas of contention resolution for local area computer networks, signal detection theory, data compression for image processing, and linear system theory.

Typically, the problem of estimation is concerned with attempting to approximate a desired quantity by a function of the available data so as to minimize a prescribed fidelity criterion. A commonplace example might be given by attempting to estimate a second order random variable \( X \) (perhaps a signal of interest) by some function \( f(\cdot) \) of the datum \( Y \) (perhaps a noise corrupted version of the signal) so as to minimize the mean square error \( \mathbb{E}(X - f(Y))^2 \). This example appears in many works on the subject of estimation theory.

In earlier work, sponsored by a previous AFOSR Grant, we showed [1] that the best such function is not necessarily given by \( f(Y) = \mathbb{E}(X | Y) \), even though \( X \) and \( Y \) are both bounded random variables. Moreover, it might seem that there is little justification from a practical viewpoint of choosing the mean square error as the appropriate fidelity criterion. Consider a fidelity criterion given by the expectation of a cost function of the error. In the context of estimation theory, 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 attach to an error, and the concern that the cost function results in a problem which one finds to be mathematically tractable. A cursory inspection of the literature in estimation theory might suggest that in many cases the second of the above concerns totally eclipses the first concern. We began a serious study of estimation theory. This work was directed to the very underpinnings of estimation theory, and it is representative of what in many cases in the literature is ignored, is postulated with no concern for the consistency of everything being postulated, or is otherwise swept under the rug. The two such areas in which we have achieved some success are concerned with continuity properties of filtrations of \( \sigma \)-algebras generated by stochastic processes and with the convergence rate of the martingale convergence theorem. We will now briefly comment on our results in these areas.

Let \( (\Omega, \mathcal{F}, P) \) be a probability space. We take a filtration of \( \sigma \)-algebras to be any nondecreasing collection of \( \sigma \)-subalgebras of \( \mathcal{F} \) indexed by \( [0, \infty) \). Let \( \{\mathcal{F}_t : t \geq 0\} \) be a
filtration. Define $F_0^- = F_0$; otherwise, define $F_t^- = \bigvee_{s < t} F_s$ and $F_t^+ = \bigcap_{s > t} F_s$.

We say that a filtration is continuous at $t$ if $F_t^- = F_t^+$; we say that it is left continuous at $t$ if $F_t^- = F_t$; and we say that it is right continuous at $t$ if $F_t^+ = F_t$. The filtration $\{F_t : t \geq 0\}$ is continuous, left continuous, or right continuous if it is continuous, left continuous, or right continuous, respectively, at $t$ for all $t \geq 0$. Let $\{X(t) : t \in [0, \infty)\}$ be a stochastic process defined on $(\Omega, \mathcal{F}, \mathbb{P})$. By a $\mathbb{P}$-null set we mean a measurable set which has probability zero. The canonical filtration of this stochastic process is given by $F_t = \sigma(X(s) : s \leq t) \lor (\mathbb{P}\text{-null sets})$ for $t \geq 0$.

In the context of estimation theory where the data are represented by a stochastic process indexed by an interval of real numbers, much of the current literature is concerned with stochastic differential equations and with martingale theory. Stochastic differential equations often arise as models for stochastic dynamical systems and techniques from martingale theory often arise in the analysis of estimation schemes and their approximation properties. In these areas one often encounters hypotheses stipulating the right continuity of filtrations of $\sigma$-algebras generated by stochastic processes. This is a blanket assumption made by many in the French and Soviet schools of stochastic process theory; see, for example, [2], [3], [4], and [5]. However, the question emerges as to when this assumption is justified or as to what reasonable hypotheses might imply it. It is often tempting and pleasing to the intuition to believe that the regularity of the sample paths of a stochastic process and the continuity of its associated canonical filtration are closely related. For example, separable Brownian motion has almost surely continuous sample paths and with the aid of the Blumenthal Zero-One Law [6] we see that its canonical filtration is continuous. Conversely, martingales with respect to right continuous filtrations have versions that are almost surely c\(^{adag}\) [7]. If we heuristically think of the canonical filtration $F_t$ as the "data" conveyed by the stochastic process $\{X(t) : t \in [0, \infty)\}$ up to and including time $t$, we may be inclined to suppose that the continuity of the sample paths of the process might prevent jumps in the "data" $\{F_t : t \geq 0\}$; and we also might suppose that the continuity of the "data" flow would influence the regularity of the sample paths of the stochastic process. (In [1] we pointed out that this is a totally erroneous concept of data.) In [8] we investigated what properties characterize filtrations of $\sigma$-algebras that are continuous. In this work we showed that the regularity of the sample paths of a stochastic process and the continuity of its associated filtration are logically independent; we presented...
an example of a stochastic process with infinitely differentiable sample paths and a
discontinuous canonical filtration and we also gave an example where a stochastic process
could have an arbitrarily irregularly prescribed sample path (e.g. non-Lebesgue
measurable) and a continuous canonical filtration. We also presented an example of a
stochastic process whose canonical filtration was discontinuous at every point. We then
went on and established conditions guaranteeing the continuity of a filtration of
σ-algebras. Also, we presented necessary and sufficient conditions for a filtration of σ-
algebras to be continuous, right continuous, or left continuous. For example, we
established the following results in [8]:

**Theorem:** Let \((\Omega, \mathcal{F}, P)\) be a probability space and \(\{\mathcal{F}_t : t \geq 0\}\) be a filtration on
\((\Omega, \mathcal{F}, P)\) so that \(\mathcal{F}_0\) contains the \(P\)-null sets. Then the filtration is continuous at \(t_0\) if and
only if for all \(Y \in L_2(\Omega, \mathcal{F}, P)\), the stochastic process defined by
\[Y_t = E(Y | \mathcal{F}_t), \quad t \geq 0,\]
is \(L_2\) continuous at \(t_0\).

Let \((\Omega, \mathcal{F}, P)\) be a probability space. A σ-subalgebra \(\mathcal{A}\) of \(\mathcal{F}\) is said to be
essentially countably generated if there exists a countable subset \(K\) of \(\mathcal{F}\) so that
\(\sigma(K) \cap (P\text{-null sets}) = \mathcal{A} \cap (P\text{-null sets})\).

**Theorem:** Let \(M\) be a separable metric space and \(\{X(t) : t \geq 0\}\) be a stochastic
process taking values in \(M\) that is left or right continuous in probability. Then
\(\sigma(X(t) : t \geq 0)\) is essentially countably generated.

**Theorem:** Let \((\Omega, \mathcal{F}, P)\) be a separable probability space. Then if \(\{\mathcal{F}_t : t \geq 0\}\) is a
filtration on \((\Omega, \mathcal{F}, P)\) so that \(\mathcal{F}_0\) contains the \(P\)-null sets, there exists a countable subset
\(C\) of \(\mathbb{R}\) so that for \(t \in C\), \(\mathcal{F}_t^- = \mathcal{F}_t^+ = \mathcal{F}_t\).

Now we comment upon some of our recent results on martingales. Frequently, in
estimation theory one derives a sequence of estimators, say \(Y_n\), and one desires to show
that as \(n \to \infty\), \(Y_n\) converges in an appropriate sense. A typical example arises in an attempt
to estimate a second order random variable \(X\) as a function of the available data, say
\((Z_n : n \in \mathbb{N})\), by choosing \(Y_n = E(X | Z_1, Z_2, \ldots, Z_n)\). In this endeavor, the martingale
convergence theorem often surfaces as a useful tool in establishing convergence.
However, in a practical circumstance, if one were interested in convergence and if \(n\)
corresponded to the progression of time, then the rate of convergence would also be of
concern. This would arise, for instance, if \(Y_n\) represented the estimate after \(n\) samples of
data are taken and data is sampled at regularly spaced intervals. The key to establishing this
rate of convergence is intimately linked with the convergence rate of the martingale convergence theorem. In [9] we examined the convergence rate of the martingale convergence theorem, and we showed that this convergence can be nonuniform and, consequently, arbitrarily slow. This result that the convergence rate of the martingale convergence theorem can be arbitrarily slow is important not only from the obvious practical viewpoint, but also from the viewpoint of the mathematician, since the martingale convergence theorem is one of the key theorems of probability theory.

Another aspect of the martingale convergence theorem which we investigated was concerned with the use of the martingale convergence theorem in estimating a random variable $X$. Let $X$ be a second order random variable, and let $\{Z_n; n \in \mathbb{N}\}$ be a sequence of random variables representing data. Often one may attempt to estimate $X$ based upon the first $n$ terms of the data sequence by $E(X \mid Z_1, Z_2, \ldots, Z_n)$. In [10] we pointed out some problems associated with an overly cavalier usage of the martingale convergence theorem in this context. In particular, we gave an example where each of the above random variables was zero mean Gaussian with a positive variance, $E(X \mid Z_1, Z_2, \ldots, Z_n) = 0$ almost surely for each $n \in \mathbb{N}$, and yet for any positive integer $k$ there exists a function $f_k: \mathbb{R} \to \mathbb{R}$ so that $f_k(Y_k) = X$ pointwise on the underlying probability space.

In a similar context as the above, in [11] we noted that for a second order random variable $X$, the rate of the $L_2$ convergence of $E|X \mid Y_1, Y_2, \ldots, Y_n|$ can crucially depend upon $X$. That is, any $L_2$ perturbation in $X$ could drastically alter the rate of convergence.

Another aspect of estimation theory with which we were concerned dealt with the idea of when an estimator which was optimal under a given fidelity criterion would also be optimal under certain other fidelity criteria. A classical paper on this subject in [12] was written by Sherman, and this result is known in the engineering literature as Sherman's theorem. However, a close inspection of [12] shows some erroneous claims. In [13] we presented a correct derivation of the effort undertaken in [12]. The following theorem is a correct version of Sherman's theorem and we proved it in [13].

**Theorem:** 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 $[0, \infty)$.

Several attempts at a proof of the above result have been presented in the engineering
literature, and each that we know of is wrong; counterexamples to these efforts are given in
[14].

Thus, the result in the above theorem 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 this theorem 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 $E[X|Y](\omega)$ for any fixed $\omega$. This special case explains why Sherman's
theorem 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

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

where $C$ denotes a Cantor subset of $[0, \infty)$ of positive Lebesgue measure, satisfies the
conditions of the above theorem. 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. In [14], we presented a more general
treatment of this general subject. The following results are presented in [14]. Notice that
the first result dispenses with the unimodality assumption, and the second result allows us
to base our estimate upon random variables measurable with respect to a non countably
generated $\sigma$-algebra, such as, for instance, that which may be generated by a random
object.

**Theorem:** Let $k \in \mathbb{N}$, $(\Omega, \mathcal{S}, \mathbb{P})$ be a probability space, and $X, Y_1, \ldots, Y_k$ be random
variables defined on $(\Omega, \mathcal{S}, \mathbb{P})$, 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: R × Ω → [0, 1], such that F(x + M(Y(ω), . . . , Yk(ω), ω), as a function of x with ω fixed, is symmetric. Then M[Y1, . . . , Yk] minimizes the quantity E[Φ(X − f(Y1, . . . , Yk)]] over all Borel measurable functions f: Rk → R where Φ: R → [0, ∞) is even and convex.

**Theorem:** Let (Ω, S, P) be a probability space, A be a σ-subalgebra of S, and X be a random variable defined on (Ω, S, P) such that X is integrable. For each ω ∈ Ω, let

M(ω) = E[X|A](ω), and assume that there exists a regular conditional distribution function of X conditioned on A, F: R × Ω → [0, 1], such that F(x + M(ω), ω), as a function of x with ω fixed, is symmetric. Then M minimizes the quantity E[Φ(X − X)] over all A-measurable random variables X, where Φ: R → [0, ∞) is even and convex.

In [15] and [16] our concern was directed toward fusing, or combining, estimates based upon a finite number of estimates of a fixed second order random variable X in order to achieve a single “best” estimate of X. For example, if X, Y1, Y2, . . . , Yn are random variables and X is second order, how might E[X | Y1], E[X | Y2], . . . , E[X | Yn] be combined so as to approximate X in a minimum mean square sense? Although aspects of this problem have been considered in the literature, we know of no other work in this area that is correct. To illustrate some subtleties in this area, note the following two examples.

**Example:** For an integer n > 1, consider a set of random variables {X, Y1, . . . , Yn} with a joint probability density function given by

\[ f(x, y1, . . . , y_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} \exp\left[\frac{-1}{2} \left(x^2 + \sum_{i=1}^{n} y_i^2\right)\right] \left(1 + x \exp\left(-\frac{x^2}{2}\right) \prod_{i=1}^{n} y_i \exp\left(-\frac{y_i^2}{2}\right)\right). \]

It follows straightforwardly that the set \{X, Y1, . . . , Yn\} is not mutually Gaussian and not mutually independent, yet any proper subset of \{X, Y1, . . . , Yn\} containing at least two random variables is mutually independent, mutually Gaussian, and identically distributed with each random variable having zero mean and unit variance. For any nonempty proper subset \(D\) of \{Y1, . . . , Yn\}, we note that \(E[X | D] = 0\) a.s. since X is independent of \(D\). However, it follows quickly that

\[ E[X | Y1, . . . , Y_n] = \frac{1}{2\sqrt{2}} \frac{1}{\sqrt{2}} y_1 . . . y_n \exp\left[\frac{-1}{2} \left(\frac{y_1^2}{2} + \frac{y_2^2}{2} + . . . + \frac{y_n^2}{2}\right)\right] \text{ a.s.} \]

Thus, since any Borel measurable function of the estimates \(E[X | D]\) where \(D\) ranges over all nonempty proper subsets of \{Y1, . . . , Yn\} would be constant almost surely, it would not be reasonable to attempt to estimate \(E[X | Y1, . . . , Y_n]\) based on a combination of these estimates.
Example: Let $\Omega = [0, 1]$, $\mathcal{F}$ denote the Borel subsets of $\Omega$, and $P$ denote Lebesgue measure on $\mathcal{F}$. Let $A$ be a positive real number, $\sigma(Y_1) = \sigma([0, 1/2))$, $\sigma(Y_2) = \sigma([1/4, 3/4])$, and $X(\omega) = A$ for $\omega \in [0, 1/4) \cup [1/2, 3/4)$ and $X(\omega) = -A$ for $\omega \in [1/4, 1/2) \cup [3/4, 1]$. Then it straightforwardly follows that $E[X|Y_1] = E[X|Y_2] = 0$ a.s., but $E[X|Y_1, Y_2] = X$ a.s. Notice that in this special case, any linear combination of $E[X|Y_1]$ and $E[X|Y_2]$ yields an estimate equal to 0 a.s., resulting in a mean square error in approximating $X$ of $A^2$, which can exceed any preassigned real number. Recalling that $E[X|Y_1]$ and $E[X|Y_2]$, respectively, are $\sigma(Y_1)$-measurable and $\sigma(Y_2)$-measurable, we see that $E[X|Y_1] = E[X|Y_2] = 0$ pointwise in $\omega$; similarly, we see that $E[X|Y_1, Y_2] = X$ pointwise in $\omega$. Thus, in this situation, it is fruitless to attempt to approximate $X$ based on any function of $E[X|Y_1]$ and $E[X|Y_2]$.

In [15] and [16] we proved the following theorem.

Theorem: Consider a probability space $(\Omega, \mathcal{F}, P)$ and random variables $X, N_1, \ldots, N_n$ defined on $(\Omega, \mathcal{F}, P)$ where $n$ is a positive integer and $X$ is a second order random variable. Further, assume that for each positive integer $i \leq n$, $N_i$ has a zero mean Gaussian distribution with positive variance given by $\sigma_i^2$, and that $X, N_1, \ldots, N_n$ are mutually independent. Define $Y_i = X + N_i$ for $i = 1, \ldots, n$. Then there exists a Borel measurable function $g: \mathbb{R}^n \to \mathbb{R}$ such that $E[X|Y_1, \ldots, Y_n] = g(E[X|Y_1], \ldots, E[X|Y_n])$ a.s.

A Monte Carlo variance reduction technique known as importance sampling has recently been applied to many problems in data communications. This technique holds the promise of offering vast improvements to traditional Monte Carlo methods. In [17] and [18] we considered importance sampling applied to the estimation of tail probabilities. In this work we gave counterexamples to some commonly used types of importance sampling. Then we introduced a new method of importance sampling, which we called Importance Sampling via a Simulacrum, and we illustrated how it could outperform some other methods of importance sampling.

In other papers we pointed out how wrong certain commonly accepted techniques and results in statistical signal processing can be. In [19] we presented a collection of counterexamples in detection and estimation. In [20] we presented a collection of counterexamples in conditioning. In [21], we presented a collection of counterexamples in maximum likelihood estimation. In [22] and [23] we presented some comments on some problems in Kalman filtering. The papers noted in this paragraph provide several
counterexamples to what is often taken as common knowledge in the literature of statistical signal processing. A copy of [20] is appended to this report.

Another direction of our research efforts was in the area of contention resolution for local area computer networks. In the last few years, packet broadcasting random multiple-access computer communication networks have been commercially available. A typical example of such a network is the Ethernet, developed by Xerox, which was designed based on the idea of carrier sense multiple access with collision detection. In Ethernet, a station among a number of users sharing a common channel will listen before transmitting and defer if the channel is busy; when two or more stations collide, each colliding station waits for a random period of time before retransmitting. Although Ethernet has the advantage of easy interconnection of stations to the common channel and it provides a high level of utilization of the channel, it does not truly address the problem of how to effectively resolve collisions in the channel. Thus, a packet involved in a collision may incur excessive delay due to waiting and abortion of transmission. Recently, a protocol called Enet II was introduced [24] as a candidate for the second generation of Ethernet. This protocol is designed to effectively resolve contention in a broadcast multiple-access network such as Ethernet. We investigated the Enet II protocol in [25], and in this investigation, we gave expressions for the average time required to resolve a collision involving k stations and the average time for a particular station involved in a k-way collision to send its packet successfully. Our results in this area were derived analytically, without recourse to efforts based on approximations or simulations. In [25] we also considered the efficiency of the protocol, and we derived a lower bound for the maximum efficiency.

In the area of image processing, a modest effort was directed toward studying the properties of a data compression scheme for image processing. In [26] we considered a modification of an existing data compression scheme which allowed more general ways of processing the image data while maintaining the favorable data compression rates.

We also devoted some effort to the problem of signal detection. In [27] we studied the problem of choosing the nonlinearity \( g(\cdot) \) when the test statistic was constrained to be of the form

\[
\sum_{i=1}^{n} g(x_i),
\]

where the \( x_i \)'s represented our observations. Observe that in the case of a constant signal
additively corrupted by mutually independent, identically distributed noise, the Neyman-
Pearson test statistic is of the above form. If the noise sequence were not mutually
independent, then the test statistic would not necessarily be of this form. However, it
might seem reasonable to suppose that in some cases, if the noise were "almost mutually
independent" then a test statistic of the above form might be a reasonable approximation to
an appropriate test statistic. In [27] we studied the problem of choosing the function \( g(\cdot) \)
so as to maximize the asymptotic relative efficiency of this detector relative to any other
detector of this form with a different nonlinearity.

In [28] we studied another aspect of statistical hypothesis testing. Consider the
situation of testing one simple hypothesis against another simple hypothesis. The
likelihood ratio (i.e. a Radon-Nikodym derivative) often arises; and it is known that in
several contexts (e.g. Neyman-Pearson, Bayes, minimax) an optimum test is given by
comparing the likelihood ratio against an appropriately chosen threshold. In [28] we
studied the question of when a likelihood ratio with respect to two probability measures \( P_0 \)
and \( P_1 \) might also be the likelihood ratio with respect to another pair of probability
measures \( Q_0 \) and \( Q_1 \) on the same measurable space. In this way, one likelihood ratio
might implement an optimum processing of the data for many pairs of probability
measures; that is, an optimal data processor might be optimal even when different
probability measures are governing the data. For the moment, consider the case where \( P_0 \)
is absolutely continuous with respect to \( P_1 \); we gave examples where the Radon-Nikodym
derivative \( \frac{dP_0}{dP_1} \) was the likelihood ratio not only for testing \( P_0 \) against \( P_1 \), but also for
testing \( Q_0 \) against \( Q_1 \), even when \( P_0 \) was extremely dissimilar from \( Q_0 \) and \( P_1 \) was
extremely dissimilar from \( Q_1 \).

In some recent efforts, we have investigated some aspects of linear systems.
Although the subject of linear systems has truly matured as a research area, we have
uncovered some unappreciated aspects of the theory. In [29] (see also [30]) we
investigated the representation of linear systems. In this work we established the following
result.

**Theorem:** Let \( \Omega \) be a locally compact separable metric space, \( \mu \) be a \( \sigma \)-finite measure on
\( \mathcal{B}(\Omega) \) (where we use \( \mathcal{B}(\cdot) \) to denote the Borel subsets of a topological space), and \( \lambda \) be a
Borel measure on a locally compact separable metric space \( W \). Let
\( T:L^1_{\text{loc}}(\Omega, \mathcal{B}(\Omega), \mu) \rightarrow L^1_{\text{loc}}(W, \mathcal{B}(W), \lambda) \) be a positive, continuous, linear map. Then
there exists \( K: \mathcal{B}(W) \times \Omega \rightarrow [0, \infty) \) so that

(i) For each \( \omega \in \Omega \), \( K(\cdot, \omega) \) is a regular Borel measure on \( \mathcal{B}(W) \).
(ii) For each $E \in \mathcal{B}(W)$, $K(E, \cdot)$ is measurable on $\Omega$.

(iii) For each $A \in \mathcal{B}(\Omega)$ with $\mu(A) < \infty$, the measure

$$\int_A K(\cdot, \omega) \, d\mu(\omega)$$

defined on $\mathcal{B}(W)$ is regular and $\lambda$-absolutely continuous.

(iv) $T(f) = \frac{d}{d\lambda} \int_{\Omega} K(\cdot, \omega)f^+(\omega) \, d\mu(\omega) - \frac{d}{d\lambda} \int_{\Omega} K(\cdot, \omega)f^-(\omega) \, d\mu(\omega)$

for $f \in L^1(\Omega, \mathcal{B}(\Omega), \mu)$, where by this notation, we mean the difference of the Radon-Nikodym derivatives of the measures given by the integrals.

Convolution frequently arises in the study of linear systems. In [31] we constructed two bounded, Lebesgue integrable, nowhere zero functions whose convolution is identically zero. This phenomenon seems to have been overlooked by many working in the area of linear systems. In particular, it dashes any hope of deconvolution in this situation. Also, although it is well known that $L_1(\mathbb{R})$, equipped with the operations of pointwise addition and convolution, is a commutative Banach algebra, this result shows that this commutative Banach algebra $L_1(\mathbb{R})$ is not an integral domain. Indeed, it shows much more than this, since there exist two nowhere zero integrable functions whose convolution is everywhere zero. In [32] we showed the analogous result for sequences. That is, we showed that there exist two summable, nowhere zero sequences whose convolution was identically zero.

This has been a brief survey of our accomplishments; more details can be found in the indicated publications. These accomplishments further our understanding of many aspects of estimation theory, of the performance of a contention resolution scheme for local area computer networks, of data compression for image processing, of signal detection theory, and of linear system theory.

REFERENCES


A copy of [20] is appended.
CONDITIONING: A CRITICAL REVIEW

Eric B. Hall  
Department of Electrical and Computer Engineering  
The University of Texas at Austin  
Austin, Texas 78712

Gary L. Wise  
Department of Electrical and Computer Engineering  
and Department of Mathematics  
The University of Texas at Austin  
Austin, Texas 78712

ABSTRACT

The concept of conditioning in probability theory forms the basis for study in many areas of information sciences and systems. Even so, the subject of conditioning is often shrouded in heuristics, misunderstood, and misused. This paper considers several aspects of conditioning with an emphasis on applications and explores several consequences of an overly cavalier approach to the oft neglected measure-theoretic subtleties involved in this area.

I. INTRODUCTION

Conditioning in probability theory is a widely recurring concept in many areas of information sciences and systems. For example, conditioning is central to many popular techniques in applied probability and, in fact, lies at the heart of many aspects of estimation theory. In spite of this widespread popularity, the subject of conditioning is commonly misunderstood and tools involving conditioning are frequently misapplied. To rephrase Doob [5, p. v], conditioning is simply a branch of measure theory, and no attempt involves the model of data as a filtration of an overly cavalier approach to the oft neglected measure-theoretic subtleties.

Consider the probability space \((\Omega, \mathcal{F}, P)\), where \(\lambda\) denotes Lebesgue measure on \(\mathcal{B}(0, 1)\), and consider the \(\sigma\)-subalgebra \(\mathcal{G}\) given by the family of all subsets of \([0, 1]\) which are either countable or cocontable. Now, for \(B \in \mathcal{B}(0, 1)\), consider the conditional probability \(P(B | \mathcal{G})\). Since \(\mathcal{G}\) contains all singletons \(\{\omega\}\), and hence might be seen as being completely informative, an overly cavalier investigator might suppose that \(P(B | \mathcal{G})\) is equal to \(I_B\). In other words, one might rationalize that to know the sets in \(\mathcal{G}\) implies that one knows itself and hence knows whether or not \(\omega\) is contained in \(B\), leading to the conclusion that \(P(B | \mathcal{G})\) should be one when \(\omega\) is contained in \(B\) and zero otherwise. It follows trivially, however, from the definition of conditional probability, that \(P(B | \mathcal{G}) = P(B)\), except possibly off of a countable subset of \([0, 1]\).

For another example, consider a probability space \((\Omega, \mathcal{F}, P)\). A commonly used model in estimation theory involves the model of data as a filtration \(\{\mathcal{F}_n, n \in N\}\) of \(\sigma\)-subalgebras of \(\mathcal{F}\). Suppose that the \(\sigma\)-algebra \(\mathcal{F}\) is separable; that is, suppose \(\mathcal{F}\) is generated by a countable family of subsets of \(\Omega\). Does it follow that \(\mathcal{F}_n\) is separable for each \(n\)? As the following example illustrates, \(\sigma\)-subalgebras of separable \(\sigma\)-algebras need not be separable.

Assume that \(\Omega = [0, 1]\) and \(\mathcal{F} = \mathcal{B}(0, 1)\). Further, let \(\mathcal{G}\) be the \(\sigma\)-subalgebra of \(\mathcal{F}\) given by the countable and cocontable subsets of \([0, 1]\). Since \(\mathcal{F} = \sigma(\{a, b\} : 0 \leq a < b \leq 1\) and \((a, b) \in \mathbb{Q}\) it follows that \(\mathcal{F}\) is separable. Assume now that \(\mathcal{G}\) is also separable; that is, assume that \(\mathcal{G} = \sigma(A_n : n \in N)\) where \(A_n\) is a subset of \([0, 1]\) for each \(n\).

Since \(\mathcal{G}\) contains only countable and cocontable subsets of \([0, 1]\), we may assume that \(A_n\) is countable for each \(n\). Let

\[ B = \bigcup_{n=1}^{\infty} A_n, \]

and note that \(B\) is also a countable subset of \([0, 1]\). Hence, there exists a real number \(x\) in \([0, 1]\) which is not an element of \(B\). Notice also that if \(D\) is the family of all subsets of \(B\) and their complements, then \(D\) is a \(\sigma\)-subalgebra such that \(\mathcal{G} \supset D \supset \sigma(A_n : n \in N)\). But, \(D \not\subseteq \mathcal{G}\) since \(\{x\}\) is in \(\mathcal{G}\) but not in \(D\). This contradiction implies that \(\mathcal{G}\) is not separable even though it is a \(\sigma\)-subalgebra of the separable \(\sigma\)-algebra \(\mathcal{F}\).

Now, let \((\Omega, \mathcal{F})\) be a measurable space, and let \(\mathcal{P}\) be a family of probability measures on \((\Omega, \mathcal{F})\). The triple \((\Omega, \mathcal{F}, \mathcal{P})\) is called a probability structure. If \(\mathcal{S}\) is a \(\sigma\)-sub-
algebra of \( \mathcal{F} \), we say that \( \mathcal{S} \) is sufficient if for each
\( \mathcal{F} \)-measurable bounded real valued function \( f \) defined on \( \Omega \),
there exists an \( \mathcal{S} \)-measurable bounded real valued function \( g \)
defined on \( \Omega \) such that \( \int_A f \, d\mathbb{P} = \int_A g \, d\mathbb{P} \) for each \( A \) in \( \mathcal{S} \) and
for all \( P \) in \( \mathcal{P} \). That is, \( g \) is almost surely \( [P] \) equal to the
conditional expectation of \( f \) conditioned on \( \mathcal{S} \) when \( P \) is the
relevant probability measure. Note that although \( g \) does not depend on \( P \), the set of \( P \)-measure zero might depend on \( P \).
It might be tempting and pleasing to the intuition to suppose infinitely differentiable sample paths has a discontinuous
canonical filtration, and a random process taking values in
\([0, 1]\) has a canonical filtration which is everywhere discontinuous.

### III. CONDITIONAL PROBABILITY

Consider a subset \( H \) of the interval \([0, 1]\) with the
properties that the outer Lebesgue measure of \( H \) is 1 and the
inner Lebesgue measure of \( H \) is 0. (For a construction of
such a set, the interested reader is referred to [8, pp.67-70].)
Further, let \( \Omega = [0, 1] \) and let \( \lambda \) denote Lebesgue measure on
\( \mathcal{B}([0, 1]) \). Define \( \mathcal{F} = \mathcal{B}(\mathbb{R}) \) such that \( \mathcal{F} \) is a \( \sigma \)-algebra on \( \mathbb{R} \) and that \( \mathcal{F} \) is a \( \sigma \)-algebra of
\( \mathcal{B}([0, 1]) \). Now, define a probability measure \( \mathbb{P} : \mathcal{F} \to [0, 1] \) on the measurable space \((\mathcal{F}, \mathbb{P})\) via
\( \mathbb{P}(B) = \mathbb{P}(B \cap \mathcal{F}) = \mathbb{P}(B) \mathbb{P}(\mathcal{F}) \) for any probability measure on
\( \mathcal{F} \). (That \( \mathbb{P} \) is well-defined follows from the properties of \( \mathbb{P} \).)

Consider now this probability space \((\Omega, \mathcal{F}, \mathbb{P})\). The following example, adapted from [2, p.464, 33.13], shows that conditional probabilities need not be measures.

Since \( \mathbb{P}(H) = 1/2 \) and \( \mathbb{P}(B) = \lambda(B) \) for \( B \in \mathcal{B}([0, 1]) \) implies that that \( \mathbb{P}(H \cap B) = \lambda(B) \mathbb{P}(H) \mathbb{P}(B) \), it follows that
\( \mathbb{P} \) is independent of \( \mathcal{B}([0, 1]) \). Let \( \mathcal{F} \) be a set in \( \mathcal{F} \) with
probability zero and assume that \( \mathbb{P}(\cdot \cap \mathcal{F}) \) is a
probability measure on \( \mathcal{F} \) for each \( \omega \) outside of the null set \( \mathcal{F} \).
Note that there exists a collection \( \{A_n : n \in \mathbb{N}\} \) of
subsets of \([0, 1]\) such that \( \mathcal{B}([0, 1]) = \sigma(\mathbb{A}_n : n \in \mathbb{N}) \) and
such that \( \{A_n : n \in \mathbb{N}\} \) is closed under finite intersections.

Define \( K_n = \{\omega \in \Omega : \mathbb{P}(A_n \cap \mathcal{F}) = 0\} \) and note that \( K_n \in \mathcal{B}([0, 1]) \) and \( \mathbb{P}(K_n) = 1 \) for all \( n \in \mathbb{N} \) since
\( \mathbb{P}(\bigcap_{n=1}^{\infty} K_n \cap \mathcal{F}) = \mathbb{P}(\bigcap_{n=1}^{\infty} K_n) \mathbb{P}(\mathcal{F}) = 0 \).
Thus, \( \mathbb{P}(\mathcal{F} \cap \mathcal{B}([0, 1])) = \mathbb{P}(\mathcal{F}) = 1 \), a.s.

Note that \( \mathbb{P}(\mathcal{K} \cap \mathcal{F}) = 1 \) and that \( \mathbb{P}(\mathcal{F}) = 1 \). Further, note that the function which,
for a fixed \( \omega \) in \( \mathcal{K} \), maps an element \( B \) of \( \mathcal{B}([0, 1]) \) to \( I_B(\omega) \)
is a probability measure on \( \mathcal{B}([0, 1]) \) which agrees with
\( \mathbb{P}(B \cap \mathcal{F}) \) whenever \( B \in \mathcal{B}([0, 1]) \). Thus, the
Dyson system theorem [1, p.169] implies that for \( \omega \in \mathbb{K} \),
\( \mathbb{P}(B \cap \mathcal{F}) = \mathbb{E}(I_B(\omega)) = 1 \) for any set \( B \) in \( \mathcal{B}([0, 1]) \). Thus, in particular, if \( \omega \in \mathbb{K} \) then
\( \mathbb{P}(\omega \cap \mathcal{F}) = \mathbb{P}(\omega) = 1 \). Now, recalling we assumed that
P(\cdot \mathcal{B}([0, 1]))(\omega) is a probability measure on \mathcal{F} for each \omega outside of the null set \mathcal{F}, we see that if \omega \in H \cap \mathcal{K} then P(H \mathcal{B}([0, 1]))(\omega) \geq P(\omega) \mathcal{B}([0, 1]))(\omega) = 1, and if \omega \in H \cap \mathcal{K} then P(H \mathcal{B}([0, 1]))(\omega) \leq P(\omega) \mathcal{B}([0, 1]))(\omega) = 0. Thus, if \omega \in \mathcal{K}, then P(H \mathcal{B}([0, 1]))(\omega) = \mathcal{H}(\omega). But H and \mathcal{B}([0, 1]) are independent, and hence P(H \mathcal{B}([0, 1])) = P(H) = \frac{1}{2} a.s. This contradiction implies that P(\cdot \mathcal{B}([0, 1]))(\omega) is not almost surely a probability measure on \mathcal{F}. Hence, a conditional probability is not necessarily a measure.

A regular conditional probability allows one to sidestep many of the undesirable aspects of conditional probability since a regular conditional probability is by definition required to be a measure for each fixed \omega \in \Omega. Unfortunately, however, regular conditional probabilities do not always exist. In fact, the situation detailed above, in addition to showing that a conditional probability need not be a measure, also provides an example in which a regular conditional probability does not exist.

IV. CONDITIONAL INDEPENDENCE

The concept of conditional independence arises frequently in many aspects of probability theory. For example, the concept plays an important role in the study of Markov processes. Unfortunately, misconceptions often arise regarding the relationship between conditional independence and independence. As the following examples adapted (with a correction) from [4, p.221] indicate, the notions of independence and conditional independence taken in conjunction concerning versions of conditional expectations. In particular, a random variable is given which is equal a.s. to a conditional expectation yet is not a version of the conditional expectation.

Consider a probability space \((\Omega, \mathcal{F}, P)\) and a \sigma-subalgebra \mathcal{H} of \mathcal{F}. Further, let \mathcal{H}^1 and \mathcal{H}^2 be two families each composed of elements from \mathcal{F}. The families \mathcal{H}^1 and \mathcal{H}^2 are said to be conditionally independent given \mathcal{H} if P(A_1 \mathcal{H}^1) = P(A_1) \mathcal{H}^2 P(A_2) \mathcal{H}^2 a.s. for all \mathcal{A}_1 \in \mathcal{H}^1 and \mathcal{A}_2 \in \mathcal{H}^2. Further, two random variables \text{X} and \text{Y} defined on \((\Omega, \mathcal{F}, P)\) are said to be conditionally independent given \mathcal{H} if \sigma(\text{X}) and \sigma(\text{Y}) are conditionally independent given \mathcal{H}.

Let \text{X}_1 and \text{X}_2 be two independent identically distributed random variables such that P(\text{X}_1 = 1) = P(\text{X}_1 = -1) = \frac{1}{2}. Further, let \text{Z} = \text{X}_1 + \text{X}_2, and let \text{A}_i = \text{X}_i^{-1}(\{1\}) for i = 1 and 2. In this case, P(\text{A}_1 \mathcal{Z}) = \frac{1}{2} on \mathcal{Z}^{-1}(\{0\}) for i = 1 or 2, and P(\text{A}_1 \cap \text{Z} \mathcal{Z}) = 0 on \mathcal{Z}^{-1}(\{0\}). In particular, P(\text{A}_1 \mathcal{Z}) \mathcal{Z} \mathcal{Z} P(\text{A}_1 \mathcal{Z}) \mathcal{Z} on an event of positive probability. Thus, the independent random variables \text{X}_1 and \text{X}_2 are not conditionally independent given \sigma(\text{Z}).

Consider now the mutually independent random variables \text{Y}_1, \text{Y}_2, and \text{Y}_3 such that each random variable takes on only integer values and such that P(\text{Y}_i = m) < 1 for all integers m and for i = 1, 2, 3. Further, let \text{S}_2 = \text{Y}_1 + \text{Y}_2 and \text{S}_3 = \text{Y}_1 + \text{Y}_2 + \text{Y}_3 and notice that \text{Y}_1 and \text{S}_3 are dependent random variables. Let \text{B}_i = \text{S}_2^{-1}(i) for each integer i. There exists k such that \text{B}_k has positive probability. On such a set \text{B}_k we have that P(\text{Y}_1 = 1, \text{S}_3 = j | \text{B}_k) = P(\text{Y}_1 = 1, \text{S}_3 = j | \text{B}_k). Hence, P(- \text{Y}_1 | \text{Y}_1 \mathcal{B}_k) = P(- \text{Y}_1 | \text{Y}_1 \mathcal{B}_k). This contradicts the fact that \text{Y}_1 and \text{S}_3 are conditionally independent given \sigma(\text{S}_2).

V. CONDITIONAL EXPECTATION

Let \text{X} be an integrable random variable defined on the probability space \((\Omega, \mathcal{F}, P)\), and let \mathcal{H} be a \sigma-subalgebra of \mathcal{F}. Then can the conditional expectation \text{E}[\text{X} | \mathcal{H}] be expressed as \text{E}[\text{X} | \mathcal{H}] = \int_{\mathcal{H}} d\mathcal{P} (\cdot | \mathcal{H})$, where \mathcal{P} (\cdot | \mathcal{H}) denotes \text{E}[\text{X} | \mathcal{H}] a.s. conditional probability given \mathcal{H}? The alert reader will immediately give a negative response to this question, since, recalling Section III, \text{E}[\cdot | \mathcal{H}] might not be a measure and hence the preceding integral might not even be defined.

The following example counters a common misconception regarding conditional probability given \mathcal{H}. Consider the probability space consisting of \([0, 1]\), \mathcal{B}([0, 1]), and Lebesgue measure on \mathcal{B}([0, 1]), and let \mathcal{G} denote the \sigma-algebra consisting of the countable and cocountable subsets of \([0, 1]\). Let \text{X} be the identity map on \([0, 1]\) and note that \text{E}[\text{X} | \mathcal{G}] = \frac{1}{2} a.s. Further, let \text{Y} = \frac{1}{2} (\text{X} - 1_{\text{C}}) where \text{C} denotes the Cantor ternary set. Note that \text{Y} = \text{E}[\text{X} | \mathcal{G}] a.s., yet \text{Y} is not \mathcal{G}-measurable (since \text{C} is neither countable nor cocountable) and hence is not a version of \text{E}[\text{X} | \mathcal{G}].

Another commonly occurring misconception regarding conditional expectation is that it is a "smoothing" operator. Consider, for example, a random process \(\text{X}(t): t \in \mathbb{R}\) defined on a probability space \((\Omega, \mathcal{F}, P)\) and a \sigma-subalgebra \mathcal{H} of \mathcal{F}. It has been argued by some (see for instance several recent papers in the area of perturbation analysis) that \text{E}[\text{X}(t) | \mathcal{H}] is "smoother" than \text{X}(t) as a function of t. To dispel this absurd notion simply let \text{X}(t) be an \mathcal{H}-measurable random process which is discontinuous everywhere; the version of \text{E}[\text{X}(t) | \mathcal{H}] given by \text{X}(t) obviously retains this same property.

Perhaps a little less obvious is the fact that, for a random variable \text{X} on \((\Omega, \mathcal{F}, P)\) and a \sigma-subalgebra \mathcal{G} of \mathcal{F}, \text{E}[\text{X} | \mathcal{G}] need not be as "smooth" a function of \text{X} as \text{X}. Consider for instance the probability space given by the.
interval $[0, 1]$, the $\sigma$-algebra $G$ given by the countable and
cocountable subsets of $[0, 1]$, and Lebesgue measure on $G$.
If we let $X = 1$, then a version of $E[X | G]$ is given by
$1 - I_B$ where $B$ equals the set of rationals in $[0, 1]$. Hence,
even though $X$ is everywhere continuous, there exists a
version of $E[X | G]$ which is everywhere discontinuous.

A commonly encountered property of conditional
expectation is the so-called nesting property. Unfortunately,
this property is sometimes misapplied. In this example,
from [6], it is shown that $E[E[X | Y]]$ may exist even when the
expectation of $X$ does not exist. In other words, before calculating $E[E[X | Y]]$ and claiming one has found the mean of $X$, it is necessary to first ascertain that the mean of $X$ actually exists.

Consider random variables $X$ and $Y$ defined on the same
probability space such that $Y$ possesses a probability density
function $g(y)$ given by $g(y) = \frac{1}{\sqrt{2\pi}y^2}$; $y > 0$, and, for
each $y > 0$, $Y$ is such that a conditional density function of $X$
given $Y = y$, denoted by $f(x | y)$, exists and is given by
$$f(x | y) = \sqrt{\frac{y}{2\pi}} \exp\left(-\frac{yx^2}{2}\right); \quad y > 0 \text{ and } x \in \mathbb{R}.$$
It follows immediately that $E[X | Y] = 0$ a.s. and therefore $E[E[X | Y]] = 0$. Notice, however, that the mean of $X$ does not exist since $X$ has a Cauchy density $h(x)$ given by
$$h(x) = \frac{1}{\pi (1 + x^2)} \text{ for } x \in \mathbb{R}.$$

For another example, consider random variables $X$ and $Y$
either defined on the same probability space $(Q, F, P)$ and
a $\sigma$-subalgebra $M$ of $F$. Another commonly encountered
misconception concerning conditional expectations is that $E[X | M]$ and $E[Y | M]$ are independent if $X$ and $Y$ are
independent. The following counterexample, which we see that, even though the random variables $X_n$ are uniformly independent, $\lim_{n \to \infty} E[X_n | M]$ is $1$ a.s. and
$$\lim_{n \to \infty} E[X_n | M] = 0 \text{ a.s.}$$

VI. REGRESSION FUNCTIONS

Given two random variables $X$ and $Y$ defined on the same
probability space, a common problem concerns the
determination of the form of the regression function $E[X | Y]$. For example, [13] considers this problem when both $X$ and $Y$ are uniformly distributed. In this example, we show that the existence of a joint probability density function for $X$ and $Y$ in no way guarantees that the regression function will obey any regularity property, other than Borel measurability.

Let $g: \mathbb{R} \to \mathbb{R}$ be Borel measurable and define $f(x, y) = \frac{1}{4} \exp(-\exp(yl) x - g(y))$. Note that $f(x, y)$ is a joint
probability density function since
$$\int \int f(x, y) \, dx \, dy$$
$R$ $R$
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 a second marginal density of \( f(x, y) \) is given by \( f_Y(y) = \int f(x, y) dx \). Recalling that

\[
E[X | Y = y] = \int \frac{f(x, y)}{f_Y(y)} dx \quad \text{and substituting for } f_Y(y)
\]

implies that \( E[X | Y = y] = g(y) \).

\[
= 2 \exp(y) \int \frac{A \exp(-\exp(-y) x) - g(y)}{4} dx
\]

\[
= 2 \exp(y) \int \frac{(x + g(y)) \frac{1}{4} \exp(-\exp(-y) x) dx}{2} \exp(y)
\]

Hence, the random variables \( X \) and \( Y \) with the joint density function \( f(x, y) \) are such that \( E[X | Y = y] = g(y) \) where we recall that \( g(\cdot) \) was an arbitrarily selected Borel measurable function.

**VII. MEAN SQUARE ESTIMATION**

One of the most common misconceptions in estimation theory is that conditional expectation minimizes mean square error. This mistaken concept arises in estimation and filtering applications in engineering as well as in many \( L_2 \) minimization problems in probability and statistics. As the following example from [1] indicates, even for bounded random variables, conditional expectation may not even come Gaussian and not mutually independent, yet any proper subset of \{X, Y, \ldots, \} containing at least two random variables is mutually independent, mutually Gaussian, and identically distributed with each random variable having zero mean and unit variance. For any nonempty proper subset \( D \) of \{Y_1, \ldots, Y_N\}, we note that \( E[X | D] = 0 \) a.s. since \( X \) is independent of \( D \). However, that \( E[X | Y_1, \ldots, Y_n] = \frac{\frac{1}{2} - \frac{1}{2} Y_1 \cdots Y_n \exp(-[1 \frac{1}{2} Y_1 + \frac{1}{2} Y_2 + \cdots + \frac{1}{2} Y_n]) \} \) a.s. follows easily. Thus, since any Borel measurable function of the estimates \( E[X | D] \) where \( D \) ranges over all nonempty proper subsets of \{Y_1, \ldots, Y_n\} would be constant almost surely, it would be absurd to attempt to estimate \( E[X | Y_1, \ldots, Y_n] \) based on a combination of these estimates. Once again, notice that the oft used and much abused Gaussian assumption does not alleviate this difficulty.

**IX. MARTINGALES**

The subject of martingale theory is an important aspect of conditioning which finds many applications in information sciences and systems. The following example shows that a martingale may have a constant positive mean, converge a.s. to zero in finite time, and yet with positive probability exceed any real number.

Let \( \{X_n; n \in \mathbb{N}\} \) be a sequence of mutually independent identically distributed random variables such that \( P(X_1 = 0) = P(X_1 = 2) = 1/2 \). Now, for each positive integer \( n \), define \( Y_n = X_1 X_2 \cdots X_n \) and note that \( \{Y_n; n \in \mathbb{N}\} \) is a martingale and that \( E[Y_n] = 1 \) for all \( n \in \mathbb{N} \). Further, notice that not only does the sequence \( \{Y_n; n \in \mathbb{N}\} \) converge almost surely to zero, but with probability one, only a finite number of terms of the sequence \( \{Y_n; n \in \mathbb{N}\} \) are nonzero. Even so, it follows easily that \( Y_n \) exceeds any real value with positive probability since \( P(Y_n = 2^n) > 0 \) for all \( n \in \mathbb{N} \).

Consider now the following example from [14] which illustrates a pathology concerning the martingale convergence theorem. In particular, it shows that in certain circumstances the martingale convergence theorem might be useless as an estimation technique.
Consider the probability space \((R, \mathcal{B}(R), P)\) where \(P\) denotes zero mean, unit variance Gaussian measure on \((R, \mathcal{B}(R))\). Let \(P_\mu\) denote the inner \(P\) measure on \((R, \mathcal{B}(R))\). Let \(S\) be a subset of \(R\) such that \(P_\mu(S) = P_\mu(S^c) = 0\). (That such sets exist is shown in [14].) Further, let \(\mathcal{W} = (S \cap B_1) \cup (S^c \cap B_2)\), \(B_1, B_2 \in \mathcal{B}(R)\) and note that \(\mathcal{W}\) is a \(\sigma\)-algebra on \(R\) which includes \(\mathcal{B}(R)\). Define a probability measure \(\mu\) on \((R, \mathcal{W})\) via
\[
\mu((S \cap B_1) \cup (S^c \cap B_2)) = (P(B_1) + P(B_2))/2. 
\]
(That \(\mu\) is well-defined follows from the properties of \(S\).) Note that the restriction of \(\mu\) to \(\mathcal{B}(R)\) is \(P\).

Consider now the probability space \((R, \mathcal{W}, \mu)\). Note first that \(S\) and \(S^c\) are each independent of \(\mathcal{B}(R)\) since, for any Borel set \(B\), \(\mu(S \cap B) = \mu(B) / \mu(S) \mu(B)\) and \(\mu(S^c \cap B) = \mu(B) / \mu(S^c) \mu(B)\). Now define a random variable \(X\) on \((R, \mathcal{W}, \mu)\) via \(X(x) = x I\{x < 2\} - x I\{x \geq 2\}\) and notice that, for any Borel set \(B\), \(\mu(X \in B) = \mu((S \cap B) \cup (S^c \cap \{x \in R : x \leq 2\})) = P(B)\) since \(P\) is symmetric. Hence, \(X\) is a Gaussian random variable with zero mean and unit variance. Further, note that
\[
E(X(x) | \mathcal{B}(R)) = x E[I\{x - I\{x = 0\}\} | \mathcal{B}(R)] =
\]
\[
= x E[I\{x = 0\}] = 0. 
\]
Since the identity map is Borel measurable, \(S\) and \(S^c\) are independent of \(\mathcal{B}(R)\), and \(P(S) = P(S^c) = 1/2\).

Now, let \(\{Y_k: k \in N\}\) be any sequence of Borel measurable functions mapping \(R\) into \(R\). Note that \(\{Y_k: k \in N\}\) is a sequence of random variables on \((R, \mathcal{W}, \mu)\). Consider the martingale
\[
[\mathcal{W} = E[X | Y_1, \ldots, Y_k]: k \in N]. 
\]
Since, given any \(k \in N\), \(\mathcal{B}(R)\) includes \(\sigma(Y_1, \ldots, Y_k)\), it follows that
\[
E[X | \mathcal{B}(R)] = \sum_{k \geq 0} E[X | \mathcal{B}(R) | Y_1, \ldots, Y_k] = 0
\]
a.s. using the previous result. Hence, the martingale convergence theorem is completely useless in estimating the random variable \(X\) in terms of the random variables \(\{Y_k: k \in N\}\). Furthermore, note that for any sequence \(\{s_k: k \in N\}\) of positive real numbers, we could let \(Y_k(x) = s_k x\). In this case, the above phenomena is exhibited when all of the random variables of concern are Gaussian. Finally, we note that, yet again, the ubiquitous Gaussian assumption does not protect us from this disturbing problem.

Another disturbing result concerning the martingale convergence theorem is detailed in [10]. There it is shown that the convergence rate guaranteed by the martingale convergence theorem can be arbitrarily slow. This result contrasts with the previous example in which the convergence was instantaneous, yet to the wrong random variable.

\section{Conclusion}

We hope these comments will be helpful to those using conditioning as a tool in investigations. Although some of these examples are undoubtedly well known to the specialist in measure theory, as previously mentioned, our experience indicates that these caveats have been overlooked by many working in the area of information sciences and systems. In conclusion, if this paper serves no other purpose, we hope it will serve as a reminder that conditioning can be a dangerous tool in the hands of amateurs.

**Acknowledgement**

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

**References**