AN INFORMATION THEORETIC PROOF OF
THE INTEGRAL REPRESENTATION THEOREM

By

S. Kullback and Peter F. Thall
FSU Statistics Report M331

January, 1975
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

This research was partially supported by the Air Force Office of Scientific
Research, Office of Aerospace Research, U.S. Air Force, under Grant AFOSR-68-1512
and by HEW Public Health Service Grant 5-T01-GM00913-13.
An Information Theoretic Proof of the Integral Representation Theorem

by

S. Kullback and Peter F. Thall

ABSTRACT

The discrimination information in a sigma-algebra is defined and various properties developed, in particular the relation between information and statistical sufficiency. The integral representation theorem is proved by methods which are more inherently information theoretic than others that have been presented.

Key Words: Discrimination Information, Integral Representation Theorem, Sub-sigma-algebra, Likelihood Ratio, Sufficient Statistic
1. INTRODUCTION

Let \( Z_t(w) \) and \( Z(w) \) be non-negative random variables on the probability space \((\Omega, A, P)\) such that

\[
\mu_t(A) = \int_A Z_t(w) dP(w), \quad \mu(A) = \int_A Z(w) dP(w), \quad A \in A
\]

are probability measures on the sigma-algebra \( A \). We also write (1) in the Radon-Nikodym differential formalism as

\[
d\mu_t = Z_t dP, \quad d\mu = ZdP,
\]

and \( Z_t, Z \) may thus be considered as generalized probability densities. If \( \mu_t \) is absolutely continuous with respect to \( \mu \) \((\mu_t \ll \mu)\), then

\[
d\mu_t = W_t d\mu = W_t ZdP = Z_t dP, \quad W_t = Z_t/Z,
\]

so that \( W_t \) is the likelihood ratio. In practice, this is the statistic upon which discrimination between the distributions defined by \( \mu_t \) and \( \mu \) is based.

We shall also deal with the analogous sequences \( \{\mu_t, Z_t\}_n \), \( \{\mu_n, Z_n\}, \{W_t\}_n \), where relationships (1), (2), and (3) are satisfied by \( \mu_t, Z_t, \mu_n, Z_n, W_t \), for each \( n \).

Let \( B \) be a sub-sigma-algebra of \( A \) and \( P|_B (\mu_t, \mu_B) \) the restriction of \( P (\mu_t, \mu) \) to \( B \). By the properties of conditional expectation (we shall use the notation of Loéve, [7]),

\[
E^B Z_t = E^B (Z \cdot (Z_t/Z)) / E^B Z = E^B Z_t / E^B Z
\]

and

\[
\mu_B(B) = \int_B ZdP = \int_B E^B ZdP, \quad B \in B,
\]

so that analogously to (2) and (3) we have
\[ d\mu_t^B = E_{Z_t}^B dP_B, \quad d\mu_B = E_{Z}^B dP_B \] (4)

and

\[ d\mu_t^B = E_{Z_t}^B d\mu_t^\omega = (E_{Z_t}^B)(E^Z dP_B) = E_{Z_t}^B dP_B . \] (5)

If \( B \) is the minimal sigma-algebra generated by the finite, \( A \)-measurable partition \( P_n = \{ B_1, \ldots, B_n \} \) of \( \Omega \) (\( = \sigma(P_n) \)), the discrimination information in \( B \) is defined by (we shall use natural logarithms)

\[ I(B; \mu_t, \mu) = \sum_{i=1}^n \mu_t(B_i) \ln \frac{\mu_t(B_i)}{\mu(B_i)} . \] (6)

In general, however, the sigma-algebra \( A \) does not have a finite generating set, so that because of the convexity of the function \( x \ln \frac{x}{y} \) for non-negative \( x \) and \( y \), and additivity of the measures for disjoint sets the discrimination information in \( A \) is defined by

\[ \overline{I}(A; \mu_t, \mu) = \sup_P \sum_{A_i \in P} \mu_t(A_i) \ln \frac{\mu_t(A_i)}{\mu(A_i)} , \] (7)

where the supremum is taken over all finite \( A \)-measurable partitions \( P \) of \( \Omega \).

It is readily seen that this reduces to expression (6) in the case that \( A = \sigma(P_n) \), for \( P_n \) finite. We shall see that both expressions are non-negative.

If \( \mu_t \) is not absolutely continuous with respect to \( \mu \), then there exists an \( A \in A \) such that \( \mu(A) = 0, \mu_t(A) > 0 \), so that \( \overline{I}(A) = \infty \). Since in this case the problem of discriminating between \( \mu_t \) and \( \mu \) is no longer one of statistical inference, the occurrence of the event \( A \) being conclusive evidence in favor of \( \mu_t \) over \( \mu \), we shall hereafter assume \( \mu_t \ll \mu \). We note, however, that \( \overline{I}(A) \) may also be infinite in this case, so that \( \mu_t \ll \mu \) is only a necessary condition for \( \overline{I}(A) < \infty \).
It also seems intuitively reasonable to have defined the discrimination
information in $A$ as the Lebesgue integral

$$I(A; \mu_t, \mu) = \int Z_t(w) \ln \frac{Z_t(w)}{Z(w)} \, d\mu(w) = \int W_t(w) \ln W_t(w) \, d\mu(w) \quad .$$

(8)

This definition is appealing in that it is expressed in terms of the likelihood
ratio $W_t$ and satisfies the intuitive requirement that $I(A) \geq 0$, with equality
if and only if $W_t = 1$ a.s. $[\mu]$.

When $A = \sigma(P)$, for $P = \{A_1, \ldots, A_n\}$ a (finite) partition of $\Omega$, the
respective distributions defined by $\mu_t$ and $\mu$ are discrete, with

$$\mu_t(A_i) = Z_t(w)P(A_i), \quad \mu(A_i) = Z(w)P(A_i), \quad w \in A_i,$$

for each $A_i \in P$. In this case, (7) and (8) yield (we hereafter suppress the
arguments $\mu_t, \mu$ when writing $I$ and $I$)

$$I(A) = \sum_{i=1}^{n} \mu_t(A_i) \ln \frac{\mu_t(A_i)}{\mu(A_i)} = I(A) \quad .$$

The question of whether (7) and (8) are equal in general can be answered in the
affirmative when $A$ is generated by a regular sequence of partitions, or equivalently
when $A$ is a separable sigma-algebra (these terms are defined in section 3).

**Theorem 1** (Integral Representation Theorem): If $A$ is a separable sigma-algebra,
then $I(A) = I(A)$.

Proofs of this result to date have involved martingale theory (e.g. [4]) or
the use (in [3]) of the convexity property of the function $x \ln \frac{x}{y}$ in conjunction
with the Darboux-Young approach to the integral. For other approaches see [1],
[2], [8], [9].
The proof to be presented in section 3 of this paper is believed to be intrinsically more information-theoretic in nature. The theorem itself formalizes the intuitive notion that the discrimination information in a general probability space is the supremum of the discrimination informations over a sequence of probability spaces generated by finite partitions, i.e. over each of which \( \mu_t \) and \( \mu \) take on discrete distributions with only finitely many values.

2. DISCRIMINATION INFORMATION IN A SUB-SIGMA-ALGEBRA

Let \( \mathcal{B} \) be a sub-sigma-algebra of \( \mathcal{A} \). We shall make repeated use of Jensen's inequality: If \( g \) is a convex function and \( E|X| < \infty \), then

\[
E(g(X)) \geq g(E(X)) ,
\]

and in its conditional form

\[
E^\mathcal{B}(g(X)) \geq g(E^\mathcal{B}(X)) , \quad \text{a.s.}
\]

It is true in general that both \( I \) and \( \overline{I} \) are non-negative and monotone, and the methods by which these properties are verified exhibit essential qualitative properties of the two representations.

The integral representation of the discrimination information in \( \mathcal{B} \) is defined, in a manner consistent with (8), by

\[
I(\mathcal{B}) = \int E^\mathcal{B}_Z \ln \frac{E^\mathcal{B}_Z}{E^\mathcal{B}_\mathcal{Z}} d\mathcal{P} = \int E^\mathcal{B}_W \ln E^\mathcal{B}_W d\mathcal{U}_B .
\]

By definition of conditional expectation and (10),

\[
I(\mathcal{A}) = \int W_t \ln W_t d\mathcal{U} = \int E^\mathcal{B}_Z(W_t) \ln W_t d\mathcal{U}_B
\]

\[
\geq \int E^\mathcal{B}_Z \ln E^\mathcal{B}_W d\mathcal{U}_B = I(\mathcal{B}) ,
\]
using the convexity of $x \ln x$, $x \geq 0$. On the other hand, since $B \subset A$ implies that any (finite) $B$-measurable partition is also $A$-measurable, $\overline{I}(A) \geq \overline{I}(B)$ follows immediately from the definition.

For the coarsest possible sub-sigma-algebra $\mathcal{B}_0 = \{\phi, \Omega\}$, it is trivially true that $I(\mathcal{B}_0) = \overline{I}(\mathcal{B}_0) = 0$. The non-negativity of $I$ then follows from its monotonicity (12), while the non-negativity of $\overline{I}$ follows, again, directly from the definition.

We next define the **conditional discrimination information** in $A$ given $B$ by

$$I(A|B) = \int_{\mathcal{W}_t} \ln(E_{\mathcal{W}_t}^{R_B})d\mu .$$

Using the fact that

$$\int_{\mathcal{W}_t} \ln(E_{\mathcal{W}_t}^{B})d\mu = \int_{E_{\mathcal{W}_t}^{B}} \ln(E_{\mathcal{W}_t}^{B})d\mu_B =$$

$$\int_{E_{\mathcal{W}_t}^{B}} \ln(E_{\mathcal{W}_t}^{B})d\mu_B = I(B) ,$$

we have

$$I(A) - I(B) = \int_{\mathcal{W}_t} \ln W_t d\mu - \int_{\mathcal{W}_t} \ln(E_{\mathcal{W}_t}^{B})d\mu =$$

$$\int_{\mathcal{W}_t} \ln(E_{\mathcal{W}_t}^{B})d\mu = I(A|B) .$$

We can now state

**Theorem 2**: If $B$ is a sub-sigma-algebra of $A$, then $I(A) = I(B) + I(A|B)$, with $I(A) = I(B)$ if and only if $I(A|B) = 0$.

**Proof**: We need only show that $I(A|B)$ is a bona fide discrimination information, and as such non-negative. This follows from the facts that
\[ \int W_t \, d\mu = \int (Z_t / Z) \, d\mu = \int Z_t \, dP = 1 \]
and
\[ \int_{E} W_t \, d\mu_B = \int (E^B Z_t / E^B Z) \, d\mu_B = \int E^B Z_t \, dP_B = \int Z_t \, dP = 1 . \]

If \( A \) and \( B \) satisfy (14), then \( B \) is said to be a sufficient sub-sigma-algebra for \( A \). This includes the classical concept of a sufficient statistic (see [5]). For the statistical problem of discrimination between \( \mu_t \) and \( \mu \), the natural statistic of interest is the likelihood ratio \( W_t \), where \( B_t = \{ W_t^{-1}(B) \colon B \text{ a linear Borel-set} \} \) is the corresponding sub-sigma algebra. The sufficiency of \( W_t \) can thus be shown by dealing exclusively with the informational properties of \( B_t \).

**Theorem 3:** \( B_t \) is a sufficient sub-sigma-algebra; that is, \( I(A) = I(B_t) \).

**Proof:** Since \( W_t \) is \( B_t \)-measurable by definition, \( E^B Z_t W_t = W_t \) a.s. [\( \mu \)]. Hence
\[ I(B_t) = \int E^B Z_t W_t \ln E^B Z_t W_t d\mu_B = \int W_t \ln W_t d\mu = I(A) . \]

3. **PROOF OF THE INTEGRAL REPRESENTATION THEOREM**

**Preliminaries:** We shall require two results from probability theory. Denote convergence in \( L_1 \) by \( Z_n \overset{L_1}{\rightarrow} Z \) and convergence in probability \( Z_n \overset{P}{\rightarrow} Z \).

**Lemma 1:** \( Z_n \overset{L_1}{\rightarrow} Z \) as \( n \to \infty \) if and only if
\[ \lim_{n \to \infty} \int W dP_n = \int Z dP, \text{ uniformly in } A \in A . \]
Lemma 2: If \( Z_n \xrightarrow{P} Z \) as \( n \to \infty \), then

\[
\lim_{n \to \infty} \int_A Z_n \, dP = \int_A Z \, dP, \text{ uniformly in } A \in \mathcal{A}.
\]

A sigma-algebra is said to be separable if and only if it is generated by a countable collection of sets. A sequence \( \{P_n\}_{n=1}^{\infty} \) of partitions is said to be regular if and only if each \( P_n \) is finite and, for each \( n \), \( P_{n+1} \) is a sub-partition of \( P_n \) (i.e. \( A \in P_{n+1} \implies A \subseteq B \), for some \( B \in P_n \)). If \( \mathcal{B}_n = \sigma(P_n) \) then, for each \( n \), \( \mathcal{B}_n \subseteq \mathcal{B}_{n+1} \), and \( \mathcal{B} = \sigma(\bigcup_{n=1}^{\infty} \mathcal{B}_n) \) is said to be generated by a regular sequence of partitions, denoted \( \mathcal{B} = \mathcal{V} \mathcal{B}_n \).

Let \( Z_1 \) and \( Z_2 \) be the respective generalized densities of the probability measures \( \mu_1 \) and \( \mu_2 \), with respect to \( P \). The total variation between \( \mu_1 \) and \( \mu_2 \) is defined as

\[
V(\mu_1, \mu_2) = \int |Z_1 - Z_2| \, dP.
\]

Kullback [6] has given the following lower bound for the discrimination information in terms of \( V(\mu_1, \mu_2) \).

Lemma 3: \( I(A; \mu_1, \mu_2) \geq \frac{V^2(\mu_1, \mu_2)}{2} \).

Proof of Theorem 1: We shall prove the result for an arbitrary sub-sigma-algebra \( \mathcal{B} \subseteq \mathcal{A} \), so that it will hold in particular for \( \mathcal{A} \).

Since \( \mathcal{A} \) is separable, so is \( \mathcal{B} \), and hence \( \mathcal{B} = \mathcal{V} \mathcal{B}_n \). By properties of conditional expectation
$$I(\mathcal{B}|\mathcal{B}_n) = \int_{E^n_{Z,t}} \ln \left( \frac{E^n_{Z,t}}{E^n_{Z,n}(U_t)} \right) d\mu_{\mathcal{B}} =$$
$$\int_{E^n_{Z,t}} \ln \left( \frac{E^n_{Z,t}}{E^n_{Z,n+1}(U_t)} \right) d\mu_{\mathcal{B}} +$$
$$\int_{E^n_{Z,t}} \ln \left( \frac{E^n_{Z,n+1}(U_t)}{E^n_{Z,n}(U_t)} \right) d\mu_{\mathcal{B}} =$$
$$I(\mathcal{B}|\mathcal{B}_{n+1}) + \int_{E^n_{Z,t}} \ln \left( \frac{E^n_{Z,n+1}(U_t)}{E^n_{Z,n}(U_t)} \right) d\mu_{\mathcal{B}} =$$
$$I(\mathcal{B}|\mathcal{B}_{n+1}) + \int_{E^n_{Z,t}} \ln \left( \frac{E^n_{Z,n+1}(U_t)}{E^n_{Z,n}(U_t)} \right) d\mu_{\mathcal{B}} =$$

since $I(\mathcal{B}_{n+1}|\mathcal{B}_n) \geq 0$. We thus have the chain of inequalities

$$I(\mathcal{B}) = I(\mathcal{B}|\mathcal{B}_0) \geq I(\mathcal{B}|\mathcal{B}_1) \geq \ldots \geq I(\mathcal{B}|\mathcal{B}_n) \geq I(\mathcal{B}|\mathcal{B}_{n+1}) \geq \ldots \geq 0$$

Since this monotonically decreasing sequence is bounded below it must converge and so must satisfy the Cauchy criterion,

$$\lim_{n,m \to \infty} (I(\mathcal{B}|\mathcal{B}_n) - I(\mathcal{B}|\mathcal{B}_{n+m})) = \lim_{n,m \to \infty} I(\mathcal{B}_{n+m}|\mathcal{B}_n) = 0 \quad (15)$$

Applying the result in Lemma 3 to (15) we have

$$\lim_{n,m \to \infty} \int_{E^n_{Z}(U_t) - E^{n+m}_{Z}(U_t)} d\mu_{\mathcal{B}} = 0$$

that is, \( \{E^n_{Z}(U_t)\} \) is \( L_1 \)-fundamental and there must exist an \( \mathcal{X} \in L_1 \) such that

$$E^n_{Z}(U_t) \xrightarrow{L_1} \mathcal{X} \quad \text{as} \ n \to \infty \quad (16)$$

By applying Lemma 1 with \( A = \Omega \),
\[
\lim_{n \to \infty} \int_{E_Z^n(U_t)} d\mu_B = \int X d\mu_B ,
\]

which implies

\[
\int X d\mu_B = 1 \tag{17}
\]
since

\[
\int_{E_Z^n(U_t)} d\mu_B = 1 , \quad \text{for all } n .
\]

Without loss of generality we may take \( X \) to be \( \mathcal{B} \)-measurable, by the following argument. Since \( X \xrightarrow{L^1} X \) implies \( E_X \xrightarrow{L^1} E_X \), we have from (16) that

\[
E^n_{Z} (U_t) = E^n_{Z} (E^n_{Z} (U_t)) \xrightarrow{L^1} E^n_{Z} X. \quad \text{In the case that } X \text{ is } \mathcal{F}\text{-measurable, } X = E^n_{Z} X \quad \text{a.s. } \{\mu\}.
\]

We shall now show that \( X = E^n_{Z} W_t \) a.s. \( \{\mu\} \). Applying Lemma 1 to (16),

\[
\lim_{n \to \infty} \int_{B} E^n_{Z} (U_t) d\mu_B = \int X d\mu_B , \quad \text{uniformly in } B \in \mathcal{B} .
\]

If \( B \in \mathcal{B}_k \), then for all \( n \geq k \),

\[
\int_{B} E^n_{Z} (U_t) d\mu_B = \int_{B} E^n_{Z} W_t d\mu_B ,
\]

which implies

\[
\int_{B} E^n_{Z} W_t d\mu_B = \int X d\mu_B . \tag{18}
\]

Since (18) must be true for arbitrary \( k \) and \( B \in \mathcal{B}_k \), it also holds for

\( B \in \bigcup_k \mathcal{B}_k \). The probability measures defined by the integrals in (18) are thus identical on the field \( \bigcup_k \mathcal{B}_k \), and hence by the Kolmogorov extension theorem (18) holds for all \( B \in \mathcal{B} \). By the uniqueness of the Radon–Nikodym derivative, (18) then yields
\[ x = E_Z^n W_t \text{ a.s. } [\mu] \] \quad (19)

We next show that

\[ \lim_{n \to \infty} I(B_n) = I(B) \] \quad (20)

By (16) and (19) it follows that \( E_Z^n W_t \overset{L_1}{\to} E_Z^B W_t \), which implies that \( E_Z^n W_t \overset{p}{\to} E_Z^B W_t \). Hence there exists a subsequence \( \{ n_k \} \) of positive integers increasing to infinity such that

\[ \lim_{k \to \infty} E_Z^{n_k} W_t = E_Z^B W_t \text{ a.s. } [\mu] \] .

Since the convex function \( x \ln x, \ x \geq 0 \), has lower bound \(-e^{-1}\), we can apply the Fatou-Lebesgue Theorem to obtain

\[ \liminf_{k \to \infty} \int E_Z^{n_k} W_t \ln E_Z^{n_k} W_t d\nu_B \geq \int E_Z^B W_t \ln E_Z^B W_t d\nu_B \] \quad (21)

By the convexity, inequality (10), and the smoothing property of conditional expectation,

\[ \int E_Z^B W_t \ln E_Z^B W_t d\nu_B = \int E^n [E_Z^B W_t \ln E_Z^B W_t] d\nu_B \geq \]

\[ \int E^n [E_Z^B W_t \ln E_Z^B W_t] d\nu_B = \int E^m [E_Z^m W_t \ln E_Z^m W_t] d\nu_B \geq \]

\[ \int E^m [E_Z^m W_t \ln E_Z^m W_t] d\nu_B , \text{ for all } n \geq m \geq 1 \] .
From the monotonic property in (22) combined with (21) we conclude that

\[
\lim_{k \to \infty} \int_{\mathbb{W}_t} E_{\mathbb{W}_t}^{B_k} \ln E_{\mathbb{W}_t}^{B_k} d\mu_{B_k} = \int_{\mathbb{W}_t} E_{\mathbb{W}_t}^{B_k} \ln E_{\mathbb{W}_t}^{B_k} d\mu_{B_k},
\]

or

\[
\lim_{k \to \infty} I(B_n) = I(B). \tag{23}
\]

Since \( I(B|B_n) \) converges, it must converge to the same limit as \( I(B|B_n) \), which by (23) is 0. We have thus shown (20).

Since each \( B_n \) is generated by a finite partition, \( I(B_n) = \overline{I}(B_n) \). Combining this with the monotonicity property and (20), we have

\[
\overline{I}(B) \geq \sup_n I(B_n) = \lim_{n \to \infty} I(B_n) = I(B). \tag{24}
\]

By the additivity of the integral and Jensen's inequality, for any finite \( \mathcal{B} \)-measurable partition \( \{A_1, \ldots, A_n\} \) of \( \Omega \),

\[
I(B) = \sum_{i=1}^{n} \int_{A_i} E_{\mathbb{W}_t}^{B} \ln E_{\mathbb{W}_t}^{B} d\mu_{B} \geq \\
\sum_{i=1}^{n} \int_{A_i} E_{\mathbb{W}_t}^{B} d\mu_{B} \ln \frac{\int_{A_i} E_{\mathbb{W}_t}^{B} d\mu_{B}}{\mu_{B}(A_i)} = \\
\sum_{i=1}^{n} \mu_{B}(A_i) \ln \frac{\mu_{B}(A_i)}{\mu_{B}(A_i)}. 
\]

Hence

\[
I(B) \geq \overline{I}(B), \tag{25}
\]

and by (24) and (25), we have the desired result. \( \Box \)
REFERENCES


