STATISTICAL THERMODYNAMICS AND NATURAL $f$-DIVERGENCES

Luc Knockaert

Abstract: A general class of $f$-divergences, closely related to statistical thermodynamics, is presented. This class includes Kullback-Leibler information and the Jensen-Shannon divergence based on the Shannon entropy. The central idea in deriving this class is the generalization of a contraction principle of large deviation theory from finite state space to infinite state space. The members of this class are in particular characterized by the presence of an unremovable asymptotic negative logarithmic term. Another feature of the members of this class is that best lower bounds in terms of the Bayes probability of error can be explicitly calculated. It is shown that Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics can be derived quite easily from the theory.

Index Terms: $f$-divergence, Maxwell-Boltzmann statistics, Bose-Einstein statistics, Legendre-Fenchel transform, Bayes probability of error.
I. INTRODUCTION

The general notion of $f$-divergence has been introduced by Csiszar [1] and independently by Ali and Silvey [2] and its many properties are studied in [3], [4]. Perhaps the best known $f$-divergence is the Kullback-Leibler information or directed divergence [5], which is also coined cross-entropy in [6], and many results have been established [7], proving its relevance and close relationship with the entropy concept.

In the statistical physics context, Frieden [8] has shown that the concept of cross-entropy can be traced back asymptotically to the multinomial law, and from a statistical thermodynamics point of view [9], it has been established from elementary arguments that the Maxwell-Boltzmann distribution derives from the multinomial law via Stirling’s approximation. Hence there appears to be a very close relationship between Maxwell-Boltzmann statistics and Kullback-Leibler information. From other elementary arguments also involving Stirling’s formula, it is easy to derive the Bose-Einstein and Fermi-Dirac distributions, but until recently, no $f$-divergences have been found which represent Bose-Einstein and Fermi-Dirac statistics in much the same way as Kullback-Leibler information represents Maxwell-Boltzmann statistics.

This author [10] established the existence of a parametrized class of $f$-divergences representing Bose-Einstein statistics, again utilizing Stirling's asymptotic approximation, which class was also independently, but in a quite different context, discovered by Lin [11]. Although it should be relatively easy to apply the same
treatment to Fermi-Dirac statistics, it appears that from a theoretical point of view this would be unsatisfactory for lack of a fundamental asymptotic framework enabling one to obtain more general results. Such an asymptotic framework has been described by Ellis [12] in the context of large deviation theory which deals with convergence properties of certain stochastic systems. For example, large deviation theory shows that under certain circumstances, the weak law of large numbers converges to zero exponentially fast. One of the most salient features of large deviation theory is that the exponential decay rates are computable in terms of entropy functions, which in turn can be calculated directly by the Legendre-Fenchel transform [13], or indirectly by applying a contraction principle involving the I directed divergence.

In this paper we revisit statistical thermodynamics [9], with its hierarchic subdivision in entities, states, levels, microstates and macrostates, its concepts of occupation numbers and degeneracies, but conditioned by a real product measure \( \mu \) over the class of subsets of the natural numbers. In this context, the probability \( P \) of a given macrostate is a well-defined real number.

In Section III we show that in the limit, for the number of entities \( M \) tending to infinity, the quantity \(-\frac{1}{M} \ln P\) tends to an \( f \)-divergence, called natural \( f \)-divergence, where \( f(x) \) is a natural \( f \)-function as defined in that section. The proof is based on a generalization of the Ellis contraction principle from a finite state space [12] (p. 16) to an infinite state space.

The strictly convex function \( f(x) \) is closely related to the generating function of
the measure $\mu$ by the way of the Legendre-Fenchel transform, and moreover in the limit, for $x$ tending to infinity, $f(x)$ behaves essentially as $-\frac{1}{q}\ln x$, where $q$ is a positive natural number.

Finally, under some mild conditions, general best lower bounds and upper bounds in terms of the Bayes probability of error [14] are derived for natural $f$-divergences. Of course the applications of the theory include pertinent results for Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics.

**II. PRELIMINARIES: STATISTICAL THERMODYNAMICS**

In what follows we consider a real measure $\mu$ over the class $\mathcal{G}(N)$ of all subsets of the natural numbers $N$. Clearly, by sigma-additivity, it is only necessary to specify the measure for the empty set $\mu(\emptyset) = 0$ and for all singletons. By a slight abuse of language we put

$$\mu(x) = \mu(\{x\}) \geq 0 \quad x \in N.$$

(2.1)

For any $B \subseteq N$ we have

$$\mu(B) = \sum_{x \in B} \mu(x).$$

(2.2)

The counting measure defined by $\mu_c(x) = 1$ is an important special case: it counts the number of elements in a set.

For the space $N^K$ of vectors

$$x = (x_1, x_2, \ldots, x_K)^T \quad x_i \in N$$

(2.3)

we define the associated product measure by
\[ \mu(x) = \mu(\{x\}) = \prod_{i=1}^{K} \mu(x_i). \]  

(2.4)

Now suppose we can associate the set \( B \subset \mathbb{N}^K \) with an event and the set \( B_0 \subset \mathbb{N}^K \) with the sure event, this requiring \( B \subseteq B_0 \), then we can define the probability of event \( B \) as

\[ Prob(B) = \frac{\mu(B)}{\mu(B_0)}. \]  

(2.5)

This is pretty close to the simple definition of probability in terms of relative frequency: it actually is relative frequency if one takes \( \mu = \mu_c \), the counting measure.

In what follows we will consider all sets \( B \) to belong to

\[ B \subseteq B_0 \subset B_\infty \subset \mathbb{N}^K, \]  

(2.6)

where \( B_\infty \) is the \( K \)-dimensional square

\[ B_\infty = \{ x : x_i \leq M \quad i = 1, 2, \ldots, K \}. \]  

(2.7)

Clearly we have

\[ \mu(B_\infty) = \left[ \sum_{i=0}^{M} \mu(i) \right]^K, \quad \mu_c(B_\infty) = (M + 1)^K. \]  

(2.8)

In the statistical thermodynamics context, a set \( \mathcal{A} \) consisting of a number of \( M \) entities is called an assembly. The assembly is subdivided in levels and the levels are subdivided in states. Schematically:

\[ \text{entity} \rightarrow \text{state} \rightarrow \text{level} \rightarrow \text{assembly}. \]

One can find numerous real-world examples of this hierarchy i.e.
marble → jar → shelve → dresser,

particle → energy state → energy level → assembly,

where the latter is pertaining to statistical physics.

We consider two kinds of entities.

1. Indistinguishable entities: there is no distinctive tag by which any two entities can be discriminated. Examples:
   - marbles having the same color.
   - all particles are photons.

2. Distinguishable entities: there is a distinctive tag by which all entities can be discriminated. Examples:
   - marbles having different colors.
   - particles obeying Maxwell-Boltzmann statistics.

Of course, there can be many more possibilities in between, such as a mixture of red and green marbles, a mixture of fermions and bosons etc., but for our purposes we restrict ourselves to the two kinds of entities defined above. Moreover, as will be shown below, distinguishable entities can be treated in exactly the same way as indistinguishable entities regarding to probability, and hence we will consider exclusively indistinguishable entities.

The mathematical framework is the following. The assembly $\mathcal{A}$ is subdivided in $L$ mutually disjoint levels $\Lambda_i \quad i = 1, 2, \ldots, L$, each containing $N_i$ entities. The
numbers $N_i$ are referred to as the occupation numbers of the levels. The total number of entities is

$$\sum_{i=1}^{L} N_i = M.$$  \hspace{1cm} (2.9)

Any level $\Lambda_i$ is subdivided in $\omega_i$ mutually disjoint states $\sigma_{i,j}$, $j = 1, 2, \ldots, \omega_i$, each containing $N_{i,j}$ entities. The number $\omega_i$ is referred to as the degeneracy of level $\Lambda_i$. In the statistical thermodynamics context, the lowest hierarchical level is the state, even when the entities are distinguishable. This means that the number of entities (counting measure) in a state uniquely defines the state. We therefore define for all states $\sigma$, by a slight abuse of language,

$$\mu(\sigma) \equiv \mu(\mu_c(\sigma)).$$  \hspace{1cm} (2.10)

Whether the entities are distinguishable or not, a specification of the occupation numbers $N_i$ for each level $\Lambda_i$ defines a macrostate of the assembly. The objective of statistical thermodynamics is to obtain an analytical expression for the probability of a macrostate. It is clear that the probability of a given macrostate will depend strongly on the degeneracies $\omega_i$, $i = 1, 2, \ldots, L$.

If the entities are indistinguishable, a specification of the number of entities in each state is said to define a microstate of the assembly. The total number of states being $K = \sum_{i=1}^{L} \omega_i$, a microstate corresponds with a vector $x \in B_0$, $B_0 \subset B_\infty \subseteq \mathbb{N}^K$ given by

$$x = \left( N_{1,1}, N_{1,2}, \ldots, N_{1,\omega_1}, N_{2,1}, N_{2,2}, \ldots, N_{2,\omega_2}, \ldots, N_{L-1,\omega_{L-1}}, N_{L,1}, \ldots, N_{L,\omega_L} \right)^T$$  \hspace{1cm} (2.11)
where

\[ B_0 = \left\{ x: \sum_{i=1}^{K} x_i = M \right\} \quad \text{(2.12)} \]

If one defines

\[ v_0 = 0, \quad v_i = \sum_{k=1}^{i} \omega_k \quad i = 1, 2, \ldots, L, \quad \text{(2.13)} \]

it is easily seen that a macrostate corresponds with the set

\[ B = \{ x: \sum_{k=v_{i-1}+1}^{v_i} x_i = N_i \quad i = 1, 2, \ldots, L \}. \quad \text{(2.14)} \]

Hence the probability of macrostate \( B \) is given by

\[ \text{Prob}(B) = \frac{\mu(B)}{\mu(B_0)} = \frac{1}{a(M, K)} \prod_{i=1}^{L} a(N_i, \omega_i) \quad \text{(2.15)} \]

where

\[ a(N, \omega) = \mu \left( \left\{ x: \sum_{i=1}^{\omega} x_i = N \right\} \right). \quad \text{(2.16)} \]

Note that the probability of a macrostate \( \text{Prob}(B) \) can be interpreted as the conditional probability of the macrostate, given the degeneracies of each level, i.e.

\[ \text{Prob}(B) = P(N_1, N_2, \ldots, N_L \mid \omega_1, \omega_2, \ldots, \omega_L), \quad \text{(2.17)} \]

since

\[ \sum_{\sum_{i=1}^{L} N_i = M} P(N_1, N_2, \ldots, N_L \mid \omega_1, \omega_2, \ldots, \omega_L) = 1 \quad \text{(2.18)} \]

for all values of the degeneracies.

The quantities \( a(N, \omega) \) can be derived from the generating function
\[ A(z) = \sum_{n=0}^{\infty} \mu(n)z^n \]  
\[ (2.19) \]

which is supposed to be analytical with radius of convergence \( R > 0 \). Such measures are called admissible. Note that there exist non-admissible measures with \( R = 0 \), such as \( \mu(n) = n! \).

By inspection we have the fundamental relationship

\[ A(z) = \sum_{n=0}^{\infty} a(n, \omega)z^n. \]  
\[ (2.20) \]

**Theorem 1:** Let \( \mu \) be an admissible measure, \( \alpha, \beta \) positive real numbers and \( \mu' \) the associated measure \( \mu'(n) = \beta \mu(n)\alpha^n \). Then

\[ \text{Prob}_\mu(B) = \text{Prob}_\mu(B) \]  
\[ (2.21) \]

**Proof:** It is clear that \( A'(z) = \beta A(\alpha z) \) and hence

\[ a'(N, \omega) = \beta^\alpha a(N, \omega)\alpha^N \]  
\[ (2.22) \]

\[ \text{Prob}_\mu(B) = \frac{1}{a'(M, K)} \prod_{i=1}^{L} a'(N_i, \omega_i) \]

\[ = \frac{1}{\beta^\alpha a(M, K)\alpha^M} \prod_{i=1}^{L} \beta^\alpha a(N_i, \omega_i)\alpha^{N_i} = \text{Prob}_\mu(B). \]  
\[ (2.23) \]

The proof is complete.

**Theorem 2:** Distinguishable entities can be treated as undistinguishable entities provided the following measure is utilized:

\[ \mu_d(n) = \mu(n)/n!. \]  
\[ (2.24) \]
Proof: Consider a microstate $x \in \mathbb{N}^K$. Since all entities are different, say an alphabet of $M$ different letters, there are

$$M! \prod_{i=1}^{K} x_i!$$

possible different realizations of the same microstate $x$. Hence

$$\mu_d(x) = \mu_d(\{x\}) = \prod_{i=1}^{K} (M!)^{1/K} \mu(x_i)/x_i!$$

and the proof is complete, since by the previous theorem the constant factor $(M!)^{1/K}$ is of no importance.

The generating function $A(z)$ plays a crucial part in deriving most of the results of statistical physics and therefore we enumerate some of the most important generating functions now.

1. Bose-Einstein statistics

$\mu$ is the counting measure and the entities are indistinguishable. Hence

$$A(z) = \sum_{k=0}^{\infty} z^k = 1/(1 - z).$$

2. Fermi-Dirac statistics

$\mu(0) = \mu(1) = 1, \quad \mu(k) = 0 \quad k \geq 2$ by the Pauli exclusion principle, and the entities are indistinguishable. Hence

$$A(z) = 1 + z.$$
3. Maxwell-Boltzmann statistics

\( \mu \) is the counting measure and the entities are distinguishable. The relevant measure is \( \mu_d(k) = 1/k! \) and therefore

\[
A(z) = e^z. \tag{2.29}
\]

III. NATURAL \( f \)-FUNCTIONS AND \( f \)-DIVERGENCES

*Definition 1:* A strict measure is an admissible measure such that \( \mu(0) > 0 \) and \( \mu(i_0) > 0 \) for at least one other index \( i_0 > 0 \).

In what follows \( \mu \) is always supposed to be strict unless otherwise stated.

Our basic problem is to provide a framework for obtaining an asymptotic expression for the probability of a macrostate

\[
\text{Prob}(B) = P(N_1, N_2, \ldots, N_L \mid \omega_1, \omega_2, \ldots, \omega_L) \tag{3.1}
\]

given the degeneracies \( \omega_i \).

In general we want the total number of entities \( M \) and the total number of states \( K \) to tend to infinity, with the proviso that the ratio \( \lim_{M \to \infty} K/M = F \) remains finite.

To this end we define the relative frequencies

\[
p_i = N_i/M \tag{3.2}
\]

and the relative degeneracies, which can be assimilated with prior probabilities

\[
q_i = \omega_i/K. \tag{3.3}
\]
\textbf{Theorem 3:}

\[ d(p_1, \ldots, p_L | q_1, \ldots, q_L) = \lim_{M \to \infty} \frac{1}{M} \ln P(N_1, \ldots, N_L | \omega_1, \ldots, \omega_L) \]  \hspace{1cm} (3.4)

is an f-divergence between \( p_i \) and \( q_i \).

\textit{Proof:}

\[ \ln P = \sum_{i=1}^{L} \ln a(N_i, \omega_i) - \ln a(M, K). \] \hspace{1cm} (3.5)

In Theorems 4 and 5 we prove that

\[ \lim_{n \to \infty} \frac{1}{n} \ln a(n, \omega) = -\varphi \left( \lim_{n \to \infty} \frac{n}{\omega} \right), \] \hspace{1cm} (3.6)

whenever \( \lim_{n \to \infty} n/\omega \) exists, and where \( \varphi(x) \) is a strictly convex function over its domain of definition.

Hence

\[ \lim_{M \to \infty} \frac{1}{M} \ln P = F \left[ -\varphi(1/F) + \sum_{i=1}^{L} q_i \varphi(p_i/F q_i) \right]. \] \hspace{1cm} (3.7)

Since \( \varphi(x) \) is strictly convex, the related mirror function \( \overline{\varphi}(x) = x \varphi(1/x) \) is also strictly convex (Theorem 6) and the formula above can be written as

\[ \lim_{M \to \infty} \frac{1}{M} \ln P = d(p_1, \ldots, p_L | q_1, \ldots, q_L) = \sum_{i=1}^{L} p_i \overline{\varphi}(F q_i/p_i) - \overline{\varphi}(F). \] \hspace{1cm} (3.8)

This is an f-divergence with strictly convex f-function

\[ f(x) = \overline{\varphi}(Fx) - \overline{\varphi}(F). \] \hspace{1cm} (3.9)

The proof is complete.

Note that \( f(1) = 0 \) which implies by Jensen’s inequality that \( d(p \mid q) \geq 0 \). In the
sequel we will show that \( d(p \mid q) \) belongs to a particular class of \( f \)-divergences called natural \( f \)-divergences.

**Theorem 4:** Let \( \phi(\tau) \) be given by

\[
\phi(\tau) = \sup_{y \leq \ln R} [\tau y - \ln A(e^y)] \quad \tau \geq 0,
\]

where \( A(z) \) is the generating function of the relevant strict measure \( \mu \).

Then

\[
\lim_{\omega \to \infty} \frac{1}{\omega} \ln a(n, \omega) = -\phi\left( \lim_{\omega \to \infty} \frac{n}{\omega} \right),
\]

whenever \( \lim_{\omega \to \infty} n/\omega = t \) exists.

**Proof:** Note first that for the set \( B_\omega \) defined in Theorem A1 we have explicitly

\[
\mu(B_\omega) = a(n, \omega).
\]

By the contraction principle we have

\[
\lim_{\omega \to \infty} \frac{1}{\omega} \ln a(n, \omega) = \lim_{\omega \to \infty} \frac{1}{\omega} \ln \mu(B_\omega) = \sup_{v \in E} (-I(v, \mu)).
\]

We therefore have to prove that

\[
\phi(t) = \inf_{v \in E} I(v, \mu).
\]

Since the convex set \( E \) is given by

\[
E = \{ v : \Sigma v(i) = 1, \quad \Sigma iv(i) = t \},
\]

the Lagrangian for the minimization problem is

\[
\Lambda(v) = \Sigma v(i) \ln[v(i)/\mu(i)] - \varepsilon(\Sigma v(i) - 1) - \lambda(\Sigma iv(i) - t),
\]
where the sums extend over all \( i \) such that \( \mu(i) > 0 \). Since \( x \ln x \) is a strictly convex function for \( 0 \leq x \leq 1 \), \( \Lambda(v) \) is strictly convex. Hence the necessary conditions

\[
\ln[v(i)/\mu(i)] + 1 - \varepsilon - \lambda i = 0 \quad \mu(i) > 0
\]  

(3.17)

are also sufficient. The optimal values of \( v(i) \) can be written as

\[
v_{opt}(i) = \mu(i)Cx^i,
\]

(3.18)

where \( C = e^{\varepsilon-1} \) and \( x = e^\lambda \). Note that effectively \( v_{opt}(i) = 0 \) whenever \( \mu(i) = 0 \). After some straightforward calculations we obtain

\[
\Lambda(v_{opt}) = -\ln A(x) + t \ln x,
\]

(3.19)

where \( x \) is the solution of the equation

\[
t = xA'(x)/A(x).
\]

(3.20)

That the solution is unique follows from Descartes' rule of signs [15]. To see this we write the above equation as

\[
\sum_{k=0}^{\infty} \mu(k)(t-k)x^k = 0.
\]

(3.21)

Since there is only one change of sign in the sequence \( \mu(k)(t-k) \) \( k = 0, 1, \ldots \) we conclude there is exactly one zero. From this follows that

\[
\Lambda(v_{opt}) = \sup_{x \leq k} \{-\ln A(x) + t \ln x\} = \sup_{y \leq \ln R} \{-\ln A(e^y) + t y\} = \phi(t),
\]

(3.22)

since the necessary condition for this maximum problem is also given by (3.20). That it is a maximum follows from the fact that \( \ln A(e^y) \) is a strictly convex function - see Theorem 5.
**Theorem 5:** Let $\phi(x)$ be the Legendre-Fenchel transform [13] of $\ln A(e^x)$ defined by

$$
\phi(x) = \mathcal{L}(\ln A(e^x)) = \sup_{t \leq \ln R} [xt - \ln A(e^t)] 
$$

(3.23)

where

$$
A(z) = \sum_{k=0}^{\infty} \mu(k)z^k \quad 0 \leq z < R
$$

(3.24)

and $\mu$ is a strict measure.

Then $\phi(x)$ is a strictly convex $C^2$ function over $\text{dom}(\phi)$ defined as $\text{dom}(\phi) = [0, \infty]$ when $A(z)$ is not a polynomial and $\text{dom}(\phi) = [0, m]$ when $A(z)$ is a polynomial of exact degree $m$. Moreover, the inversion formula

$$
\ln A(e^x) = \mathcal{L}(\phi(x)) = \sup_{t \in \text{dom}(\phi)} [tx - \phi(t)]
$$

(3.25)

is valid.

**Proof:** We first show that $h(x) = \ln A(e^x)$ is a strictly increasing, strictly convex $C^2$ function. That it is strictly increasing follows from the fact that the derivative

$$
h'(x) = e^x A'(e^x)/A(e^x) > 0, \quad -\infty < x < \ln R, \quad h'(-\infty) = 0.
$$

(3.26)

That it is strictly convex follows from Hölder’s inequality: take $x \neq y$ and $0 < \lambda < 1$.

We have

$$
e^{h(\lambda x + (1-\lambda)y)} = \sum_{k=0}^{\infty} \mu(k)\lambda^k e^{k\lambda x} \mu(k)^{1-\lambda} e^{(1-\lambda)y} \leq \left( \sum_{k=0}^{\infty} \mu(k)e^{kx} \right)^{\lambda} \left( \sum_{k=0}^{\infty} \mu(k)e^{ky} \right)^{1-\lambda} = e^{\lambda h(x) + (1-\lambda)h(y)},
$$

(3.27)
where the Hölder exponents are \( p = 1/\lambda \) and \( q = 1/(1 - \lambda) \). For equality we should have
\[
\mu(k)e^{by} = E\mu(k)e^{ky} \quad k = 0, 1, \ldots
\]  
(3.28)
for some constant \( E \). Since \( \mu \) is strict we should have \( x = y \) which is not allowed, whence the strict convexity follows. That \( h(x) \) is \( C^2 \) follows from the fact that \( h''(x) \) can be written as
\[
h''(x) = \left( \sum_{k=0}^{\infty} \mu(k)k^2e^{kx} \right) - \left( \sum_{k=0}^{\infty} \mu(k)ke^{kx} \right)^2 / A(e^x)^2 > 0, \quad -\infty < x < \ln R,
\]
\[h''(-\infty) = 0. \]  
(3.29)
Since \( g(x) = h'(x) \) is strictly increasing and continuous, it is a homeomorphism [16] (p. 82) of \([−\infty, \ln R]\) onto \([g(−\infty), g(\ln R)]\), where these intervals are considered to be subsets of the extended real line \([−\infty, \infty]\). Hence the inverse function \( g^{-1}(x) \) exists and is continuous and strictly increasing for \( x \in [g(−\infty), g(\ln R)] \).

To facilitate the further development of the theorem we show at this point that \([g(−\infty), g(\ln R)] = \text{dom}(\phi)\) as defined above. Since \( g(−\infty) = 0 \) we ought to find the value of \( g(\ln R) \) which can be written as
\[
g(\ln R) = B = \lim_{x \to R} xA'(x)/A(x). \]  
(3.30)
Suppose first that \( B < \infty \). Since \( g(t) \) is strictly increasing this would imply that \( xA'(x) \leq BA(x) \) or
\[
\sum_{k > B} \mu(k)(k-B)x^k \leq \sum_{k < B} \mu(k)(B-k)x^k, \quad 0 \leq x < R.
\]  
(3.31)
This is impossible for finite $B$ except when $A(z)$ is a polynomial. For if $R$ is finite, the radius of convergence of the series on the left-hand side of inequality (3.31) is also $R$. By a result in [16] (p. 244) the point $x = R$ is singular for a power series with coefficients which are positive or zero, and hence in the limit the left-hand side would tend to $\infty$ while the right-hand side would remain finite. When $A(x)$ is a polynomial of exact degree $m$, it is a simple matter to show that $B = m$. When $A(x)$ is a non-polynomial entire function we have $B = \infty$, since no polynomial can dominate at $+\infty$ a non-polynomial entire function with coefficients that are positive or zero. This proves that $[g(-\infty), g(\ln R)] = \text{dom}(\phi)$.

Next consider the strictly concave function of $t$ with parameter $x$:

$$
\xi(x, t) = tx - h(t). \tag{3.32}
$$

For $x < 0$ the function is strictly decreasing and hence obtains its maximum for $t = -\infty$. Since $h(-\infty) = \ln A(0) = \ln \mu(0)$ we obtain $\phi(x) = \infty$ for $x < 0$ and $\phi(0) = -\ln \mu(0) < \infty$. For $x > 0$ the necessary and sufficient condition (strict concavity) for a maximum is given by $t = g^{-1}(x)$. Under these conditions we simply have

$$
\phi(x) = \xi(x, g^{-1}(x)) \tag{3.33}
$$

and

$$
\phi'(x) = \frac{\partial \xi}{\partial x} + \frac{\partial \xi}{\partial t} \frac{dt}{dx} = t = g^{-1}(x). \tag{3.34}
$$

Since $g^{-1}(x)$ is strictly increasing, we conclude that $\phi(x)$ is strictly convex.

After integration we obtain
\[ \phi(x) = \int_0^x g^{-1}(u)\,du - \ln \mu(0). \]  
(3.35)

Note that \( \text{dom}(\phi) \) corresponds with the domain of \( g^{-1}(x) \) and hence with the range of \( g(x) \) i.e. \([g(-\infty), g(\ln R)]\), as has been said before.

To prove that \( \phi(x) \) is \( C^2 \), consider simply that \( g(\phi'(x)) = x \) and hence

\[ \phi''(x) = 1/h''(g^{-1}(x)). \]  
(3.36)

For \( x = 0 \) we have

\[ \phi(0) = -\ln \mu(0), \quad \phi'(0) = -\infty, \quad \phi''(0) = \infty. \]  
(3.37)

To prove that

\[ h(x) = L(\phi(x)) = \sup_{t \in \text{dom}(\phi)} \xi(x, t), \]  
(3.38)

where

\[ \xi(x, t) = xt - \phi(t), \]  
(3.39)

we observe that we have the following necessary and sufficient condition

\[ x = \phi'(t) = g^{-1}(t). \]  
(3.40)

Hence \( t = g(x) \) and

\[ L(\phi(x)) = \xi(x, g(x)) = xg(x) - \phi(g(x)) = xg(x) + \ln \mu(0) - \int_0^{g(x)} g^{-1}(u)\,du. \]  
(3.41)

After the change of variables \( u = g(\tau) \) we have

\[ \xi(x, g(x)) = xg(x) + \ln \mu(0) - \int_{-\infty}^{x} \tau d\,g(\tau), \]  
(3.42)

or after integration by parts

\[ \xi(x, g(x)) = h(x) + \ln \mu(0) - h(-\infty) + \lim_{\tau \to -\infty} \tau g(\tau). \]  
(3.43)
Since \( h(-\infty) = \ln \mu(0) \) and
\[
\lim_{\tau \to \infty} \tau g(\tau) = \lim_{t \to 0} t \ln t A'(t)/A(t) = 0,
\]
with \( t = e^x \), the result follows.

\[ (3.44) \]

**Definition 2:** A function \( f(x) \) defined over the positive reals is called a natural \( f \)-function if
\[
f(1) = 0, \quad \bar{f}(x) = \mathcal{L}[F \ln B(e^x)],
\]
where the mirror function \( \bar{f}(x) \) is defined by
\[
\bar{f}(x) = xf\left(\frac{1}{x}\right),
\]
(3.46)

\( \mathcal{L} \) stands for the Legendre-Fenchel transform, \( F \) is a positive real constant and
\[
B(z) = \sum_{k=0}^{\infty} \mu(k)z^k,
\]
where \( \mu \) is a strict measure.

The function \( f(x) \) is called a strictly natural \( f \)-function if \( B(z) \) is not a polynomial.

**Lemma:** The relation \( \prec \) defined for two functions \( f(x) \) and \( g(x) \) over the positive reals by the definition
\[
f(x) \prec g(x) \iff \exists a \in R : \quad f(x) = g(x) + a(x - 1)
\]
(3.48)
is an equivalence relation. Moreover it is invariant for the mirror operation defined above in the sense that
\[
f(x) \prec g(x) \iff \bar{f}(x) \prec \bar{g}(x).
\]
(3.49)
Proof: By inspection.

Theorem 6: Any natural \( f \)-function \( f(x) \) is strictly convex and can be expressed as \( f(x) \sim \phi(Fx) - \phi(F) \), where \( F > 0 \) and \( \phi(x) = \mathcal{L}(\ln A(e^x)) \), with

\[
A(z) = \sum_{k=0}^{\infty} \mu(k)z^k,
\]

(3.50)

and where \( \mu(k) \) is a strict measure.

Proof: By the definition, the major condition is

\[
\tilde{f}(x) = \sup_{t} \left[ xt - F \ln B(e^t) \right] = \sup_{\tau} \left[ x \ln \tau - F \ln B(\tau) \right].
\]

(3.51)

Define \( B(z) \) as

\[
B(z) = \sum_{k=0}^{\infty} \mu'(k)z^k = CA(cz),
\]

(3.52)

where

\[
\mu'(k) = C \mu(k)c^k, \quad c, C > 0.
\]

(3.53)

If \( \mu' \) is strict, then \( \mu \) is strict and vice versa.

Hence we have

\[
\tilde{f}(x) = -F \ln C - x \ln c + F \phi(x/F).
\]

(3.54)

This means that

\[
f(x) = -xF \ln C - \ln c + \phi(Fx)
\]

(3.55)

or

\[
f(x) \sim \phi(Fx) - \phi(F),
\]

(3.56)
since \( f(1) = 0 \).

By Theorem 5 \( \phi(x) \) is strictly convex. To complete the proof we have to show that the mirror function \( \bar{\phi}(x) = x\phi(1/x) \) is also strictly convex. Note first that \( \phi(x) \) and hence \( \bar{\phi}(x) \) are defined only for \( x \geq 0 \). The mirror operation is its own inverse i.e.

\[ \phi(x) = x\bar{\phi}(1/x) \]

and hence

\[
(\lambda x + (1 - \lambda)y)\bar{\phi}\left(\frac{1}{\lambda x + (1 - \lambda)y}\right) < \lambda x\bar{\phi}\left(\frac{1}{x}\right) + (1 - \lambda)y\bar{\phi}\left(\frac{1}{y}\right), \tag{3.57}
\]

with \( x \neq y \) and \( 0 < \lambda < 1 \).

Taking

\[
\hat{\lambda} = \frac{\lambda x}{\lambda x + (1 - \lambda)y}, \quad \hat{x} = 1/x, \quad \hat{y} = 1/y, \tag{3.58}
\]

the result follows, since

\[
\hat{\lambda}\hat{x} + (1 - \hat{\lambda})\hat{y} = \frac{1}{\lambda x + (1 - \lambda)y}. \tag{3.59}
\]

It is important to note in this context that when \( f(x) \) is not strictly natural, \( A(z) \) is a polynomial of exact degree \( m \), and we can run into trouble since \( \text{dom}(\phi) = [0, m] \) and hence \( \text{dom}(\bar{\phi}) = [1/m, \infty] \), requiring \( F \geq 1/m \) and \( Fx \geq 1/m \).

**Definition 3:** An \( f \)-divergence

\[
d_f(p \mid q) = \sum_{k=1}^{L} p_k f(q_k/p_k) \tag{3.60}
\]

is called natural if \( f(x) \) is a natural \( f \)-function.

It is called strictly natural if \( f(x) \) is a strictly natural \( f \)-function.
Note that this definition makes sense since a natural $f$-function is strictly convex by Theorem 6. Moreover we have $d_f(p \mid q) \geq f(1) = 0$ by Jensen’s inequality and the lower bound 0 is obtained iff $p = q$. If the $f$-divergence is not strictly natural, one must be careful, since in that case it is required that $p_k \leq m F q_k \quad k = 1, 2, \ldots$, which means that $p$ is then a priori bounded above by $q$. Note also that it is an easy matter to show that
\[ f(x) \asymp g(x) \Rightarrow d_f(p \mid q) = d_g(p \mid q). \quad (3.61) \]

**Theorem 7:** A necessary condition for $f(x)$ to be a natural $f$-function is that there exists a real number $\delta$ such that the following representation holds:
\[ f(x) \asymp -\frac{\ln x}{q} + D + \sum_{k=1}^{\infty} d_k(\beta x)^{-k/q}, \quad x > \delta \geq 0, \quad (3.62) \]
where $q$ is a positive natural number.

**Proof:** By Theorem 6 we have that $f(x)$ is strictly convex and
\[ f(x) \asymp \overline{\phi}(Fx) - \overline{\phi}(F), \quad (3.63) \]
where $F > 0$, $\phi(x) = \mathcal{L}(\ln A(e^x))$ and
\[ A(z) = \sum_{k=0}^{\infty} \mu(k) z^k, \quad (3.64) \]
where $\mu$ is strict, which implies there is a smallest index $q > 0$ such that $\mu(q) > 0$.

The Legendre-Fenchel transform can be written as
\[ \phi(x) = \sup_{0 \leq t \leq R} [x \ln t - \ln A(t)]. \quad (3.65) \]
The necessary and sufficient condition for the maximum is
\[ \frac{x}{t} = \frac{A'(t)}{A(t)}. \] (3.66)

By inspection we also have that \( \phi'(x) = \ln t \). Since \( \mu(0) > 0 \) and \( q \) is the next smallest index such that \( \mu(q) > 0 \), the function

\[ g(t) = \left( \frac{A(t) / A'(t)}{\mu(0) / \mu(q) q^{q-1}} \right)^{\frac{1}{q}} \] (3.67)

is analytical in some interval \([0, \varepsilon)\), with \( \varepsilon > 0 \) and \( g(0) = 1 \), and hence has a power series expansion

\[ g(t) = \sum_{k=0}^{\infty} g_k t^k, \quad 0 \leq t < \varepsilon, \] (3.68)

with \( g_0 = 1 \). After some manipulations it appears that the maximum condition can be written as

\[ X = t / g(t), \] (3.69)

where

\[ X = \eta x^{\frac{1}{q}} \] (3.70)

and

\[ \eta = \left[ \frac{\mu(0)}{q \mu(q)} \right]^{\frac{1}{q}}. \] (3.71)

This is the standard form required for the Lagrange inversion formula [17] i.e. \( t \) as a function of \( X \) is analytical in some interval \([0, \varepsilon_1)\), with power series expansion

\[ t(X) = \sum_{k=1}^{\infty} c_k X^k, \quad 0 \leq X < \varepsilon_1, \] (3.72)
where the coefficients \( c_k \) are given by
\[
\frac{c_k}{k!} \left\{ \left( \frac{d}{dt} \right)^{k-1} (g(t)) \right\}_{t=0}
\] (3.73)

Since \( c_1 = 1 \) we have
\[
\phi'(x) = \ln t(X) = \ln X + \ln \left( 1 + \sum_{k=2}^{\infty} c_k X^{k-1} \right), \quad 0 \leq X < \varepsilon_1.
\] (3.74)

Taking into account the power series for \( \ln(1+z) \) we conclude that there is some \( \varepsilon_2 > 0 \) such that
\[
\phi'(x) = \ln X + \sum_{k=1}^{\infty} e_k X^k, \quad 0 \leq X < \varepsilon_2.
\] (3.75)

This leads to
\[
\phi'(x) = \ln \eta + \frac{\ln x}{q} + \sum_{k=1}^{\infty} e_k \eta^k x^{q^k}, \quad 0 \leq x < \varepsilon_3 = (\varepsilon_2/\eta)^q.
\] (3.76)

After integration we have
\[
\phi(x) = -\ln \mu(0) + x \left[ \ln \eta - \frac{1}{q} \right] + \frac{1}{q} x \ln x + \sum_{k=1}^{\infty} d_k \eta^k x^{q^k+1}, \quad 0 \leq x < \varepsilon_3,
\] (3.77)

where the coefficients \( d_k \) are given by
\[
d_k = e_k \frac{q}{k+q}.
\] (3.78)

Finally, after some straightforward calculations we obtain
\[
f(x) \propto -\frac{\ln x}{q} + D + \sum_{k=1}^{\infty} d_k (\beta x)^{-kq}, \quad x > \delta \geq 0,
\] (3.79)

where
\[ D = \ln \eta - \frac{1 + \ln F}{q}, \quad \beta = F \eta^{-q}, \quad \frac{1 \pm i}{b} = F e, \]

and the proof is complete.

**Corollary**: Let \( f(x) \) be a natural \( f \)-function with

\[ \lim_{x \to \infty} \frac{f(x)}{x} = 0. \tag{3.81} \]

Then \( f(x) \) is strictly decreasing.

**Proof**: By the theorem there is a real constant \( a \) such that

\[ f(x) = a(x - 1) - \frac{1}{q} \ln x + D + O(x^{-1/q}), \quad x \to \infty. \tag{3.82} \]

By the limit condition it appears that \( a = 0 \). Since \( f(x) \) is strictly convex it can have at most one finite minimum. But \( f(\infty) = -\infty \) which contradicts the existence of a finite minimum, whence the result follows.

Note that Theorem 7 allows one to inspect directly whether a given convex function \( f(x) \) and corresponding \( f \)-divergence \( d_f(p \mid q) \) makes a chance of being natural, since the behaviour at infinity should be

\[ f(x) \sim Aw + B - \frac{1}{q} \ln x + O(x^{-1/q}). \tag{3.83} \]

For instance the commonly used Kolmogorov variational distance and Hellinger distance [4] are not natural \( f \)-divergences, since for the first \( f(x) = \frac{1}{2} | 1 - x | \) and for the second \( f(x) = \frac{1}{2}(1 - \sqrt{x})^2 \). It is seen that in both cases the logarithmic term is
lacking.

**IV. BOUNDS IN TERMS OF THE BAYES PROBABILITY OF ERROR**

In binary classification and hypothesis testing problems, the Bayes probability of error [14] is of utmost importance. It is given by

\[ P_e(p, q) = \sum_{k=1}^{L} \min(\lambda p_k, (1-\lambda)q_k), \]  

(4.1)

where \( \lambda \) and \( 1 - \lambda \) are the a priori probabilities with \( 0 < \lambda < 1 \). Note that when \( \lambda = \frac{1}{2} \), we have the simple relationship \( P_e(p, q) = \frac{1}{2} [1 - V(p, q)] \) between the Bayes probability of error and the Kolmogorov variational distance, defined as

\[ V(p, q) = \frac{1}{2} \sum_{k=1}^{L} |p_k - q_k|. \]  

(4.2)

The next two theorems give, under some mild conditions, upper bounds and best lower bounds in terms of the Bayes probability of error for strictly natural \( f \)-divergences.

*Theorem 8:* Let \( f(x) = \bar{\phi}(Fx) - \bar{\phi}(F) \) with \( \phi(x) \) as defined in Theorem 6, the corresponding radius of convergence \( R = 1, F > 0 \) and \( \mu(0) = 1 \).

Then the following upper bound is valid:

\[ d_f(p \| q) \leq |\bar{\phi}(F)| \frac{1}{\lambda} |\phi\left( F \frac{\lambda}{1-\lambda} \right)| P_e(p, q). \]  

(4.3)

*Proof:* Clearly \( f(x) \) is natural and hence strictly convex. It is also strictly decreasing by the Corollary of Theorem 7 since
\[
\lim_{x \to \infty} \frac{f(x)}{x} = \lim_{x \to \infty} \frac{\overline{\phi}(Fx)}{x} = F \lim_{x \to 0} \phi(x) = -F \ln \mu(0) = 0. 
\tag{4.4}
\]

Moreover we have
\[
\overline{\phi}(x) = x \phi \left( \frac{1}{x} \right) = \sup_{0 \leq t \leq R} [\ln t - x \ln A(t)] 
\tag{4.5}
\]
and hence \( \overline{\phi}(0) = \ln R = 0 \).

This means that the function \( g(x) = \overline{\phi}(Fx) \) is strictly convex and strictly decreasing from 0 to \(-\infty\) for all values \( F > 0 \).

The Bayes probability of error can be written as
\[
P_e(p, q) = -\sum_{k=1}^{L} p_k f_e(q_k/p_k),
\tag{4.6}
\]
where
\[
f_e(x) = -(1-\lambda)x, \quad x \leq \frac{\lambda}{1-\lambda}, \tag{4.7a}
\]
\[
= -\lambda, \quad x \geq \frac{\lambda}{1-\lambda}. \tag{4.7b}
\]

Clearly \( f_e(x) \) is a convex function with \( f_e(0) = 0 \) and \( f_e(\infty) = -\lambda \). For any positive constant \( A \) a sufficient condition to have \( g(x) \leq Af_e(x) \) is therefore \( g \left( \frac{\lambda}{1-\lambda} \right) \leq -A \lambda \).

Under these conditions it is easy to show that
\[
d_f(p \mid q) + \overline{\phi}(F) \leq -A P_e(p, q),
\tag{4.8}
\]
and the result follows, if one actually takes
\[
A = -\frac{1}{\lambda} \overline{\phi} \left( F \frac{\lambda}{1-\lambda} \right). \tag{4.9}
\]
Theorem 9: Let \( d_f(p \mid q) \) be a strictly natural \( f \)-divergence with \( \phi(x) \) and \( F \) as usual. Let \( P_0(p, q) \) be the Bayes probability of error with \( \lambda \leq \frac{1}{2} \) and let \( \kappa = \frac{1}{\lambda} - 1 \geq 1 \). Then the lower bound

\[
J_\kappa[1 - (\kappa + 1)P_0(p, q)] \leq d_f(p \mid q),
\]

(4.10)

where

\[
J_\kappa(\zeta) = \inf_{\alpha - \kappa \beta = \zeta} T(\alpha, \beta),
\]

(4.11a)

\[
T(\alpha, \beta) = \alpha \bar{\phi}[F \beta / \alpha] + (1 - \alpha)\bar{\phi}[F (1 - \beta) / (1 - \alpha)] - \bar{\phi}(F),
\]

(4.11b)

is the best possible lower bound.

Proof: By Theorem 6 we have \( f(x) = \bar{\phi}(Fx) - \bar{\phi}(F) \), where \( F > 0 \) and \( \phi(x) = L(\ln A(e^x)) \), with

\[
A(z) = \sum_{k=0}^{\infty} \mu(k)z^k
\]

(4.12)

and \( \mu \) a strict measure.

Let \( h(x) = \ln A(e^x) \). By the definition of the Legendre-Fenchel transform we have

\[
\tau x - h(\tau) \leq \phi(x), \quad \tau \leq \ln R, \quad x \geq 0,
\]

(4.13)

and hence

\[
\tau - Fxh(\tau) \leq \bar{\phi}(Fx), \quad \tau \leq \ln R, \quad x \geq 0.
\]

(4.14)

Applying this to \( x = q_k / p_k \), multiplying by \( p_k \) and summing up, leads to the following basic inequality:

\[
\sum_{k=1}^{L} [p_k \tau_k - Fq_k h(\tau_k)] \leq d_f(p \mid q) + \bar{\phi}(F), \quad \tau_k \leq \ln R.
\]

(4.15)
Taking $\tau_k = t_1$ whenever $p_k \geq \kappa q_k$ and $\tau_k = t_2$ whenever $p_k < \kappa q_k$, we obtain the inequality
\[\alpha t_1 - \beta h(t_1) + (1 - \alpha) t_2 - \beta h(t_2) \leq d_f(p \mid q) + \bar{\phi}(F), \quad t_1, t_2 \leq \ln R, \tag{4.16}\]
where
\[\alpha = \sum_{p_k \geq \kappa q_k} p_k, \quad \beta = \sum_{p_k \geq \kappa q_k} q_k. \tag{4.17}\]

It is straightforward to show that $\alpha - \kappa \beta \geq \max(0, 1 - \kappa) = 0$ and
\[P_\epsilon(p, q) = \frac{1 - (\alpha - \kappa \beta)}{1 + \kappa}. \tag{4.18}\]

Since $t_1, t_2 \leq \ln R$ can be freely chosen, we can take the supremum with respect to these arguments, leading to
\[T(\alpha, \beta) \leq d_f(p \mid q), \tag{4.19}\]
where
\[T(\alpha, \beta) = \alpha \bar{\phi}(F \beta / \alpha) + (1 - \alpha) \bar{\phi}(F (1 - \beta) / (1 - \alpha)) - \bar{\phi}(F). \tag{4.20}\]

This lower bound is actually the best possible since it can be attained for the distributions
\[p_1 = \alpha - \delta, \quad p_2 = 1 - \alpha - \delta, \quad q_1 = \beta - \delta, \quad q_2 = 1 - \beta - \delta, \quad p_i = q_i = \epsilon^{i-2} \quad i = 3 \ldots L, \tag{4.21}\]
where
\[\delta = \frac{1 - \epsilon^{L-1} - \epsilon}{2(\epsilon - 1)}. \tag{4.22}\]
Since \( f(1) = 0 \) we have in the limit for \( \epsilon \to 0 \) that \( d_f(p \mid q) = T(\alpha, \beta) \).

For a given value of \( P_\epsilon(p, q) \), \( \alpha \) and \( \beta \) lie on the line \( \alpha - \kappa \beta = \zeta \), where \( \zeta = 1 - P_\epsilon(p, q)(1 + \kappa) \). Note that \( 0 \leq \zeta \leq 1 \), which follows from the fact that \( 0 \leq P_\epsilon(p, q) \leq 1 \). Since \( \beta \) is unknown, the inequality (4.19) can be interpreted in the following way:

\[
\exists \beta: \ T(\zeta + \kappa \beta, \beta) \leq d_f(p \mid q),
\]  
(4.23)

and hence the result follows, provided we can prove the existence and uniqueness of the function

\[
J_\kappa(\zeta) = \inf_{\alpha - \kappa \beta = \zeta} T(\alpha, \beta).
\]  
(4.24)

To this effect consider the Lagrangian

\[
\Lambda(\alpha, \beta) = T(\alpha, \beta) - \rho(\alpha - \kappa \beta).
\]  
(4.25)

The necessary conditions are

\[
\frac{\partial \Lambda}{\partial \alpha} = \phi'\left(\frac{\alpha}{F \beta}\right) - \phi'\left(\frac{1 - \alpha}{F(1 - \beta)}\right) - \rho = 0
\]  
(4.26)

and

\[
\frac{\partial \Lambda}{\partial \beta} = F \phi'\left(\frac{F \beta}{\alpha}\right) - F \phi'\left(\frac{F(1 - \beta)}{1 - \alpha}\right) + \kappa \rho = 0.
\]  
(4.27)

The Lagrange multiplier \( \rho \) can be eliminated easily yielding the equation

\[
G(x_0) = G(x_1), \quad x_0 = \frac{1 - \alpha}{F(1 - \beta)}, \quad x_1 = \frac{\alpha}{F \beta},
\]  
(4.28)

where
\[ G(x) = \kappa \phi'(x) + F \bar{\phi}\left(\frac{1}{x}\right) = F \phi(x) + (\kappa - F x) \phi'(x). \] (4.29)

The interesting point here is that
\[ G'(x) = (\kappa - F x) \phi''(x) \] (4.30)
and hence that \( G(x) \) is strictly increasing for \( x < \kappa/F = \gamma \) and strictly decreasing for \( x > \gamma \). Since by inspection \( x_0 \leq \gamma \leq x_1 \), the condition \( G(x_0) = G(x_1) \) establishes a one to one correspondence between \( x_0 \) and \( x_1 \), say \( x_1 = \theta(x_0) \), where \( \theta(x) \) is a strictly decreasing continuous function and \( \theta(0) = \infty \) - see also Corollary 2.

The expressions for \( \alpha, \beta \) and \( \zeta \) in terms of \( x_0 \) and \( x_1 \) are
\[ \alpha = x_1 \frac{1 - F x_0}{x_1 - x_0}, \quad \beta = \frac{1}{F} \frac{1 - F x_0}{x_1 - x_0}, \] (4.31a)
\[ \zeta = \eta(x_0) = (\theta(x_0) - \gamma) \frac{1 - F x_0}{\theta(x_0) - x_0}. \] (4.31b)

Since \( 0 \leq (x_1 - \gamma)/(x_1 - x_0) \leq 1 \) it appears that in order to have \( 0 \leq \zeta \leq 1 \), we have to impose the condition \( 0 \leq x_0 \leq 1/F \). Since \( \theta(x_0) \) is continuous it is seen that \( \eta(x_0) \) is continuous and \( \eta(0) = 1, \eta(1/F) = 0 \). To prove that \( \eta(x_0) \) is strictly decreasing it is sufficient to show that it is injective or one-to-one [16], (p. 82).

Suppose we have \( 0 < \tilde{x}_0 \leq \tilde{x}_1 < 1/F \) with \( \eta(\tilde{x}_0) = \eta(\tilde{x}_1) = \zeta_0 \). Below we prove that \( T(\alpha, \beta) \) and hence \( \Lambda(\alpha, \beta) \), which we minimize, are convex functions. Now the set of points at which a convex function takes on its global minimum is a convex set [18]. Consider the values \( \tilde{x}_1, \tilde{x}_1, \zeta_1 \) and \( \beta = \zeta_0/F(\tilde{x}_1 - \gamma), \beta = \zeta_0/F(\tilde{x}_1 - \gamma) \) corresponding
to $\bar{x}_0, \tilde{x}_0$. The convexity of $\Lambda(\alpha, \beta)$ implies that the same value $\zeta_0$ must be obtained for all $\bar{\beta} \leq \beta \leq \tilde{\beta}$ and hence for all $\bar{x}_0 \leq x_0 \leq \tilde{x}_0$. This means that

$$x_1 = \frac{\gamma - x_0(\kappa + \zeta_0)}{1 - \zeta_0 - Fx_0}, \quad \bar{x}_0 \leq x_0 \leq \tilde{x}_0$$  \hspace{1cm} (4.32)

and

$$\frac{dx_1}{dx_0} = \frac{\zeta_0(\kappa - 1 + \zeta_0)}{[1 - \zeta_0 - Fx_0]^2} \geq 0, \quad \bar{x}_0 < x_0 < \tilde{x}_0.$$  \hspace{1cm} (4.33)

Since $\theta(x_0)$ is strictly decreasing this is only possible if $(\bar{x}_0, \tilde{x}_0) = \emptyset$ i.e. $\bar{x}_0 = \tilde{x}_0$. This proves that $\eta(x_0)$ is one-to-one and hence strictly decreasing, and therefore the inverse function $x_0 = \eta^{-1}(\zeta)$ exists and is continuous. After some trivial substitutions we obtain the following expression for $J_\kappa(\zeta)$ in terms of $x_0$ and $x_1$.

$$J_\kappa(\zeta) = \Xi(x_0) = F \left[ \phi(x_0) - \phi\left( \frac{1}{F} \right) \right] + (1 - Fx_0) \frac{\phi(x_1) - \phi(x_0)}{x_1 - x_0}.$$  \hspace{1cm} (4.34)

Note that the couple of equations $\zeta = \eta(x_0)$ and $J_\kappa(\zeta) = \Xi(x_0)$ form a parametric description of $J_\kappa(\zeta)$.

To complete the theorem we prove that $T(\alpha, \beta)$ is convex. We can write $T(\alpha, \beta)$ as

$$T(\alpha, \beta) = S(\alpha, \beta) + S(1 - \alpha, 1 - \beta) - \bar{\varphi}(F),$$  \hspace{1cm} (4.35)

where

$$S(\alpha, \beta) = F \beta \varphi(\alpha/F \beta).$$  \hspace{1cm} (4.36)

If $S(\alpha, \beta)$ is convex, then clearly $S(1 - \alpha, 1 - \beta)$ and $T(\alpha, \beta)$ are convex. To prove that $S(\alpha, \beta)$ is convex, we show that its Hessian is positive semidefinite. We have
\[
\frac{\partial^2 S}{\partial \alpha^2} = \frac{1}{F \beta^2} \phi''(\alpha/F \beta), \tag{4.37a}
\]
\[
\frac{\partial^2 S}{\partial \beta^2} = \frac{\alpha^2}{F \beta^3} \phi''(\alpha/F \beta), \tag{4.37b}
\]
\[
\frac{\partial^2 S}{\partial \alpha \partial \beta} = -\frac{\alpha}{F \beta^2} \phi''(\alpha/F \beta). \tag{4.37c}
\]

The eigenvalues are \( s_1 = 0 \) and
\[
s_2 = \frac{\alpha^2 + \beta^2}{F \beta^3} \phi''(\alpha/F \beta) > 0, \tag{4.38}
\]
whence the result follows. The proof is complete.

Note that the case \( \lambda > \frac{1}{2} \) can be treated by interchanging \( p \) and \( q \), and that for \( \lambda = \frac{1}{2} \), \( \kappa = 1 \) and by inspection the best lower bound in terms of \( V(p, q) \) is simply \( J_1[V(p, q)] \).

**Corollary 1:** Let \( \tilde{\phi}(x) \equiv \phi(x) \) and \( F = \kappa = \frac{1}{\lambda} - 1 \geq 1 \). Then the best lower bound is given by
\[
\frac{1}{\lambda} [\xi(P, (p, q)) - \xi(\lambda)] \leq d_f(p \mid q), \tag{4.39}
\]
where
\[
\xi(x) = x \phi\left(\frac{1}{x} - 1\right). \tag{4.40}
\]
Proof: We first show that it is actually possible to have \( \overline{\phi}(x) = \phi(x) \) for certain natural \( f \)-functions. The function \( \overline{\phi}(x) \) can be written as
\[
\overline{\phi}(x) = \sup_{t \in [-\ln R, -\ln R]} [-t - x h(-t)].
\] (4.41)

The function \( -h(-t) \) is a strictly increasing, strictly concave homeomorphism of \([-\ln R, \infty]\) onto \([-\infty, -\ln \mu(0)]\). Hence its inverse function \( \hat{h}(t) \) is a strictly increasing, strictly convex homeomorphism of \([-\infty, -\ln \mu(0)]\) onto \([-\ln R, \infty]\). This implies that
\[
\overline{\phi}(x) = \sup_{t \leq -\ln \mu(0)} [xt - \hat{h}(t)] = \mathcal{L}(\hat{h}(t)),
\] (4.42)

and by Theorem 5 we should therefore have that \( h(t) = \hat{h}(t) = -h^{-1}(-t) \) or equivalently
\[
\frac{1}{A(z)} = \frac{1}{z}. \quad (4.43)
\]

This seems to be a strong condition, imposed on the generating function of the strict measure \( \mu \), but it is not an impossible condition, since it is readily verified that it is satisfied by generating functions of the form \((1 + \delta x)/(1 - x)\), where \( \delta > -1 \). This includes the counting measure when \( \delta = 0 \). Note that entire functions are de facto excluded, since we should have \( \mu(0)R = 1 \).

The function \( G(x) \) is given by
\[
G(x) = F[\phi(x) + (1 - x)\phi'(x)]. \quad (4.44)
\]

The condition \( \overline{\phi}(x) = \phi(x) \) implies \( G(x) = G(1/x) \), and hence \( x_1 = 1/x_0 \) or
\[
\frac{F \beta}{\alpha} = \frac{1 - \alpha}{F(1 - \beta)}. \tag{4.45}
\]

The relevant function \( T(\alpha, \beta) \) is then
\[
T(\alpha, \beta) = (\alpha - F \beta + F) \phi(F \beta / \alpha) - \phi(F), \tag{4.46}
\]

and the result follows, since
\[
\alpha - F \beta = 1 - (F + 1) \rho_e \tag{4.47}
\]

implies that
\[
\alpha = (1 - \rho_e) \left( 1 - \rho_e / \lambda \right) / (1 - 2 \rho_e), \tag{4.48}
\]
\[
F \beta = \rho_e (1 - \rho_e / \lambda) / (1 - 2 \rho_e). \tag{4.49}
\]

**Corollary 2:** \( J_\kappa(\zeta) \) is a strictly increasing, strictly convex function. Moreover the following parametric representations are valid:
\[
J_\kappa(\zeta) = -\tau(\Delta) + \Delta \frac{d \tau}{d \Delta}, \quad \zeta = \frac{d \tau}{d \Delta}, \quad \Delta \geq \Delta_{\text{min}} \geq 0, \tag{4.50}
\]

where the strictly increasing and strictly convex function \( \tau(\Delta) \) is defined by
\[
\gamma \Delta = h(u_0 + \Delta) - h(u_0), \quad \tau(\Delta) = F h(u_0) - u_0 + \bar{\phi}(F), \quad u_0 \leq u_{\text{max}} = \phi'(1/F), \tag{4.51}
\]

and \( \Delta_{\text{min}} \) satisfies the equation
\[
\gamma \Delta_{\text{min}} = h(u_{\text{max}} + \Delta_{\text{min}}) - h(u_{\text{max}}). \tag{4.52}
\]

**Proof:** Since \( \phi(x) = x \phi'(x) - h(\phi'(x)) \) and \( \phi'(x) = h'^{-1}(x) \), we have the following relationship:
\[
G(h'(u)) = \kappa u - F h(u), \quad u \leq \ln R, \tag{4.53}
\]
implying that \( G(0) = G(\infty) = -\infty \).

Since \( h'(u) \) is strictly increasing and \( \kappa u - F h(u) \) is strictly concave, we can find two points \( u_0 \) and \( u_1 \) such that \( u_0 \leq u_1 \) and
\[
h'(u_0) = x_0, \quad h'(u_1) = x_1, \quad \gamma u_0 - h(u_0) = \gamma u_1 - h(u_1).
\]  
\tag{4.54}

Putting \( \Delta = u_1 - u_0 \geq 0 \), we obtain the fundamental relationships
\[
\gamma \Delta = h(u_0 + \Delta) - h(u_0)
\]  
\tag{4.55}

and
\[
d\frac{u_0}{\Delta} = \frac{\gamma - x_1}{x_1 - x_0} < 0.
\]  
\tag{4.56}

The function \( \eta(x_0) \) can therefore be written as
\[
\eta(x_0) = (F h'(u_0) - 1) \frac{d u_0}{d \Delta} = \frac{d \tau}{d \Delta} > 0, \quad u_0 \leq u_{\text{max}}.
\]  
\tag{4.57}

As for the function \( J_\kappa(\zeta) \), it can be written as
\[
J_\kappa(\zeta) = F \phi(x_0) - \phi(F) + \eta(x_0) \frac{\phi(x_1) - \phi(x_0)}{x_1 - \gamma}.
\]  
\tag{4.58}

We have that \( h(u_1) = h(u_0) + \gamma \Delta \) and
\[
h'(u_1) = \left( \gamma + h'(u_0) \frac{d u_0}{d \Delta} \right) \left( 1 + \frac{d u_0}{d \Delta} \right).
\]  
\tag{4.59}

After some straightforward calculations we obtain
\[
\phi(x_1) - \phi(x_0) = (x_0 - \gamma) \left( \frac{d u_0}{d \Delta} - u_0 \right) \left( \frac{d u_0}{d \Delta} + 1 \right)
\]  
\tag{4.60}

and
\[ x_1 - \gamma = (x_0 - \gamma) \frac{du_0}{d\Delta} \left( 1 + \frac{du_0}{d\Delta} \right), \tag{4.61} \]

yielding

\[ \frac{\phi(x_1) - \phi(x_0)}{x_1 - \gamma} = \Delta - u_0 \frac{d\Delta}{du_0}. \tag{4.62} \]

Moreover it is easy to prove that

\[ F \phi(x_0) = u_0 \frac{d\tau}{du_0} - \tau + \overline{\phi(F)}, \tag{4.63} \]

and the result follows, since

\[ J_x(\zeta) = u_0 \frac{d\tau}{du_0} - \tau + \frac{d\tau}{d\Delta} \left( \Delta - u_0 \frac{d\Delta}{du_0} \right) = -\tau + \Delta \frac{d\tau}{d\Delta}. \tag{4.64} \]

The strict convexity of \( \tau(\Delta) \) follows from the fact that \( \eta(x_0) \) is strictly decreasing with respect to \( x_0 \) and \( u_0 \) and hence strictly increasing with respect to \( \Delta \). Note that

\[ \tau(\Delta_{\text{min}}) = F h(u_{\text{max}}) - u_{\text{max}} + \overline{\phi(F)} = 0 \tag{4.65} \]

and

\[ \frac{dJ_x}{d\zeta} = \Delta, \tag{4.66} \]

implying that \( J_x(\zeta) \) is strictly increasing. By inspection it appears that

\[ J_x(\zeta) = \mathcal{L}(\tau(\Delta)) = \sup_{\Delta} [\zeta \Delta - \tau(\Delta)], \tag{4.67} \]

which implies, from Legendre-Fenchel transform theory [12],[13], that \( J_x(\zeta) \) is strictly convex.
V. APPLICATIONS

In the preceding section we have shown, among other results, that

\[ d(p \mid q) = \lim_{M \to \infty} -\frac{1}{M} \ln P(N_1, \ldots, N_L \mid \omega_1, \ldots, \omega_L) \]  

is a natural $f$-divergence. The right-hand side of the above equation is the limit of a negative log-likelihood. Hence minimizing $d(p \mid q)$ is equivalent to maximizing a likelihood.

In the statistical physics context, one minimizes with respect to $p$ belonging to some constraint set, which is in fact a maximum a posteriori (MAP) procedure, due to the fact that the degeneracies and hence $q$ are assumed known.

But in other cases, $p$ can represent a known distribution, and $q$ belonging to some constraint set, has to be estimated. Since $q$ is to the right of the conditional bar, this is in fact a maximum likelihood (ML) procedure.

In the following theorem we propose explicit solutions for some simple ML and MAP problems.

**Theorem 10:** Let $d_f(p \mid q)$ be a natural $f$-divergence. Let the MAP problem be defined as

\[ \inf_{p} d_f(p \mid q), \quad \sum_{k=1}^{L} p_k = 1, \quad \sum_{k=1}^{L} \epsilon_k p_k = E, \]  

and the corresponding ML problem be defined as
\[
\inf_q d_f(p \mid q), \quad \sum_{k=1}^{L} q_k = 1, \quad \sum_{k=1}^{L} \varepsilon_k q_k = E. \quad (5.3)
\]

A necessary condition for the existence of a solution is that
\[
p_k / q_k = F h'(\lambda + \delta \varepsilon_k), \quad (5.4)
\]
in the case of the MAP problem, and
\[
p_k / q_k = F h'(h^{-1}(\lambda + \delta \varepsilon_k)), \quad (5.5)
\]
in the case of the ML problem. Here \(\lambda, \delta, \lambda, \bar{\delta}\) are suitable real constants. The parameter \(F > 0\) and \(h(x) = \ln A(e^x)\) are directly related to \(f(x)\) in the usual manner.

**Proof**: First consider the MAP problem. We have
\[
d_f(p \mid q) = \sum_{k=1}^{L} p_k f(q_k / p_k) = \sum_{k=1}^{L} q_k \bar{f}(p_k / q_k), \quad (5.6)
\]
and hence the Lagrangian is given by
\[
\Lambda(p) = \sum_{k=1}^{L} [q_k \bar{f}(p_k / q_k) - \lambda' p_k - \delta p_k \varepsilon_k]. \quad (5.7)
\]
The necessary condition for a minimum is therefore
\[
\bar{f}'(p_k / q_k) = \lambda' + \delta \varepsilon_k. \quad (5.8)
\]
Since \(f(x) = \bar{\phi}(Fx) - \bar{\phi}(F)\) and \(\bar{f}(x) = F \phi(x/F) - x \bar{\phi}(F)\), the necessary condition can be written as
\[
\phi'(p_k / F q_k) = \lambda + \delta \varepsilon_k, \quad (5.9)
\]
with \(\lambda = \lambda' + \bar{\phi}(F)\).

Since \(\phi^{-1}(x) = h'(x)\), the first result follows.

For the ML problem the Lagrangian is
\[ \Lambda(q) = \sum_{k=1}^{L} \left[ p_k f(q_k/p_k) - \lambda_1 q_k - \delta_1 q_k \varepsilon_k \right], \quad (5.10) \]

and hence the necessary conditions are
\[ f'(q_k/p_k) = \lambda_1 + \delta_1 \varepsilon_k, \quad (5.11) \]

or
\[ \bar{\phi}'(F q_k/p_k) = \lambda_2 + \delta_2 \varepsilon_k, \quad (5.12) \]

where \( \lambda_2 = \lambda_1/F \) and \( \delta_2 = \delta_1/F \). The function \( \bar{\phi}(x) \) can be written as
\[ \bar{\phi}(x) = \sup_t [t - x h(t)]. \quad (5.13) \]

The maximum is obtained for \( h'(t) = 1/x \) and moreover \( \bar{\phi}'(x) = -h(t) \). Since \( h(t) \)
is strictly increasing and strictly convex, the inverse functions \( h^{-1}(t) \) and \( h'^{-1}(t) \)
exist and are strictly increasing. Hence
\[ \bar{\phi}'(x) = -h(h'^{-1}(1/x)) \quad (5.14) \]

and
\[ \bar{\phi}^{-1}(x) = 1/h'(h^{-1}(-x)), \quad (5.15) \]

whence the result follows, if one takes \( \bar{\lambda} = -\lambda_2 \) and \( \bar{\delta} = -\delta_2 \).

It is seen that the above approach remains appropriate for any number of linear equality constraints.

We shall now derive the most salient results pertaining to statistical physics.
1. Maxwell-Boltzmann statistics

The relevant measure is \( \mu(k) = 1/k! \) and therefore \( A(x) = e^x \). Hence

\[
h(x) = \ln A(e^x) = e^x, \tag{5.16}
\]

\[
\phi(x) = \mathcal{L}(h(x)) = x \ln x - x, \tag{5.17}
\]

\[
f(x) = \tilde{\phi}(Fx) - \tilde{\phi}(F) = -\ln x, \tag{5.18}
\]

and \( d_f(p \mid q) = I(p \mid q) \). Note that the parameter \( F \) has disappeared. By Theorem 10 the solution to the MAP problem is

\[
p_k = q_k e^{\lambda + \delta \xi_k}, \tag{5.19}
\]

and the solution to the corresponding ML problem is

\[
q_k = \frac{p_k}{\lambda + \delta \xi_k}. \tag{5.20}
\]

The MAP equation gives the celebrated Maxwell-Boltzmann term and the ML equation has been utilized in spectrum estimation [19] to find the nearest spectrum, in the Kullback-Leibler sense, to a given spectrum under linear autocovariance constraints.

The best lower bound in terms of the Bayes probability of error can be found by exploiting Theorem 9 Corollary 2. After some simple manipulations we obtain the function \( \tau(\Delta) \), which uniquely defines \( J_\lambda(\zeta) = \mathcal{L}(\tau(\Delta)) \) as

\[
\tau(\Delta) = \kappa \Delta/(e^\Delta - 1) - \ln[\kappa \Delta/(e^\Delta - 1)] - 1, \quad \Delta \geq \Delta_{\text{min}}, \tag{5.21}
\]

where \( \Delta_{\text{min}} \) is the solution of the equation
\[ \frac{1}{\kappa} = \Lambda_{\text{min}}(e^{\Lambda_{\text{min}}} - 1). \] (5.22)

Note that an explicit analytical expression for \( J_k(\zeta) \) seems difficult to find, but nevertheless the bound \( J_1[V(p, q)] \) in terms of the Kolmogorov variational distance is the best possible, and hence better than the bounds given in [20] and [21]. An upper bound in terms of the Bayes probability of error is probably impossible to find since to apply Theorem 8 one should have a radius of convergence 1 whereas here it is infinity.

2. Bose-Einstein statistics

The relevant measure \( \mu \) is the counting measure and therefore \( A(z) = 1/(1 - z) \). Hence

\[ h(x) = \ln A(e^x) = -\ln(1 - e^x), \] (5.23)

\[ \phi(x) = \overline{\phi}(x) = L(h(x)) = x \ln x - (x + 1) \ln(x + 1), \] (5.24)

and \( f(x) = \overline{\phi}(Fx) - \overline{\phi}(F) \).

The strictly natural \( f \)-divergence \( d_f(p \mid q) \) can be neatly expressed as

\[ d_f(p \mid q) = K[H(np + (1 - n)q) - nH(p) - (1 - n)H(q)], \] (5.25)

where \( H(p) \) is the Shannon entropy

\[ H(p) = - \sum_{k=1}^{L} p_k \log_2 p_k \] (5.26)

and \( n = 1/(F + 1), \quad K = (F + 1) \ln 2. \)

Many properties of \( d_f(p \mid q) \), which is called the Bose-Einstein number in [10],
and the related number

\[ JS_\lambda(p, q) = d_f(p \mid q) / K, \]  \hspace{1cm} (5.27)

which is called the Jensen-Shannon divergence in [11], are given in those papers.

In order to solve the MAP and ML problems we need the functions

\[ h'(x) = 1/(e^{-x} - 1), \quad h^{-1}(x) = \ln(1 - e^{-x}) \quad \text{and} \quad h'(h^{-1}(x)) = e^x - 1. \]

By Theorem 10 the solution to the MAP problem is therefore

\[ p_k = F \cdot q_k \left( e^{-\lambda - \delta_k} - 1 \right), \]  \hspace{1cm} (5.28)

and the solution to the corresponding ML problem is

\[ q_k = \frac{1}{F} \cdot p_k \left( e^{\lambda + \delta_k} - 1 \right). \]  \hspace{1cm} (5.29)

In both MAP and ML equations we recognize the celebrated Bose-Einstein term, with the slight difference that in the first \( F \) occurs in the numerator, whereas in the second it occurs in the denominator.

Lower and upper bounds in terms of the Bayes probability of error are given in the following theorem, which is also the last of this paper.

**Theorem 11:** Let \( JS_\lambda(p, q) \) be the Jensen-Shannon divergence defined above and let \( P_\epsilon(p, q) \) be the Bayes probability of error with the same parameter \( \lambda \leq \frac{1}{2} \). Then

\[ H_2(\lambda) - H_2(P_\epsilon(p, q)) \leq JS_\lambda(p, q) \leq H_2(\lambda) - 2P_\epsilon(p, q), \]  \hspace{1cm} (5.30)

where

\[ H_2(x) = -x \log_2 x - (1 - x) \log_2 (1 - x), \quad 0 \leq x \leq 1. \]  \hspace{1cm} (5.31)
Proof: Note that we can write

\[ F = \frac{1}{\lambda} - 1, \quad d_f(p \mid q) = \frac{\ln 2}{\lambda} JS_\lambda(p, q). \quad (5.32) \]

We have \( \mu(0) = R = 1 \) and therefore we can apply Theorem 8, yielding

\[ d_f(p \mid q) \leq |\phi(F) - \frac{1}{\lambda} \phi \left( F \frac{\lambda}{1 - \lambda} \right)| P_q(p, q). \quad (5.33) \]

Since |\( \bar{\phi}(1) = \phi(1) = -2 \ln 2 \) the upper bound follows from

\[ \xi(\lambda) = \lambda \phi \left( \frac{1}{\lambda} - 1 \right) = -H_2(\lambda) \ln 2. \quad (5.34) \]

This equation also provides the best lower bound by Theorem 9 Corollary 1.

Note that the upper bound has also been obtained in [11].

3. Fermi-Dirac statistics

The relevant measure is given by \( \mu(0) = \mu(1) = 1, \quad \mu(k) = 0, \quad k \geq 2 \) and therefore \( A(z) = 1 + z \). Note that this is closely related to the coin-tossing problem in [12] where \( A(z) = (1 + z)/2 \). The fact that \( A(z) \) is a polynomial of exact degree 1 implies that \( \text{dom}(\phi) = [0, 1] \). We have

\[ h(x) = \ln A(e^x) = \ln(1 + e^x), \quad (5.35) \]

\[ \phi(x) = \mathcal{L}(h(x)) = x \ln x + (1 - x) \ln(1 - x) = -H_2(x) \ln 2, \quad 0 \leq x \leq 1, \quad (5.36) \]

\[ \bar{\phi}(x) = (x - 1) \ln(x - 1) - x \ln x, \quad x \geq 1 \quad (5.37) \]

and
\[ f(x) = \overline{\phi}(Fx) - \overline{\phi}(F), \quad F \geq 1, \quad Fx \geq 1. \quad (5.38) \]

The \( f \)-divergence \( d_f(p \mid q) \) makes sense only when \( F \geq 1 \) and \( p_k \leq Fq_k \) for all \( k \). Quite surprisingly, it can be written in terms of the Jensen-Shannon divergence with a negative index i.e.

\[ d_f(p \mid q) = \frac{\ln 2}{\nu} JS_\nu(p,q), \quad \nu = \frac{1}{1 - F}. \quad (5.39) \]

This means that both Fermi-Dirac and Bose-Einstein statistics can be taken into account by the same \( f \)-divergence

\[ d_f(p \mid q) = \frac{\ln 2}{\nu} JS_\nu(p,q), \quad -\infty < \nu \leq 1. \quad (5.40) \]

The value \( \nu = 0 \) corresponds in both cases with \( F = \infty \), and one deduces easily [10] that

\[ \lim_{\nu \to 0} d_f(p \mid q) = I(p,q), \quad (5.41) \]

implying that Maxwell-Boltzmann statistics is included as well.

In order to solve the MAP and ML problems we need the functions

\[ h'(x) = 1/(e^{-x} + 1), \quad h^{-1}(x) = \ln(e^x - 1) \text{ and } h'(h^{-1}(x)) = 1 - e^{-x}. \]

By Theorem 10 the solution to the MAP problem is therefore

\[ p_k = Fq_k(e^{-\lambda - \delta e_k} + 1), \quad (5.42) \]

and the solution to the corresponding ML problem is

\[ q_k = \frac{1}{F}P_k(1 - e^{-\lambda - \delta e_k}). \quad (5.43) \]
In the MAP equation we recognize the celebrated Fermi-Dirac term, and the ML equation has the aspects of an inverted Bose-Einstein term.

Note that no bounds in terms of the Bayes probability of error can be found. The reasons are: $R = \infty$ and $f(x)$ is not strictly natural since $A(z)$ is a polynomial.

**CONCLUDING REMARKS**

We discuss some possible generalizations of the theory.

It is obvious that letting the number of levels $L$ tend to infinity should not pose much problems. For general probability measures $P$ and $Q$ we should redefine $d_f(p \mid q)$ as

$$d_f(P \mid Q) = \int f(dQ/dP) dP,$$

where $dQ/dP$ is the Radon-Nikodym derivative [3],[4] of $Q$ with respect to $P$. This author conjectures in that context that the bounds in terms of the Bayes probability of error will remain valid, but maybe not necessarily that the lower bound is the best possible. In any case this requires some thorough measure-theoretic investigations.

Another field of investigation would be to generalize the results in the Appendix to other kinds of permutation sets or to non-permutation sets, since it is clear that the contraction principle grants the multinomial law and hence the $I$ directed divergence its central role only because of the fact that one deals exclusively with permutation sets.
APPENDIX

Definition A1: \( Y \subseteq \mathbb{N}^p \) is a permutation set iff

\[
x = (x_1, x_2, \ldots, x_p)^T \in Y \Rightarrow (x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(p)})^T \in Y
\]

(A.1)

for all permutations \( \pi \) of \((1, 2, \ldots, p)\).

Example: The \( p \)-dimensional square \( Y_n = \{ x: \ x_i \leq n, \ i = 1, 2, \ldots, p \} \) is a permutation set.

In what follows we consider functions \( f, g \geq 0 \) taking values in \( \mathbb{R} \) and defined over non-empty subsets

\[
X_\omega \subseteq \mathbb{N}^\delta \quad \omega, \delta \in \mathbb{N},
\]

(A.2)

such that \( \omega \) and \( \delta \) tend to \( \infty \).

Definition A2: The function \( f(x) \geq 0 \) is \( \omega \)-negligible iff

\[
\lim_{\omega \to \infty} \left[ \sum_{x \in X_\omega} f(x) \right]^{\frac{1}{\omega}} = 1.
\]

(A.3)

Definition A3: The function \( f(x) \geq 0 \) is absolutely \( \omega \)-negligible if it is \( \omega \)-negligible and if moreover

\[
\lim_{\omega \to \infty} \left[ \inf_{x \in X_\omega} f(x) \right]^{\frac{1}{\omega}} = 1.
\]

(A.4)
We suppose of course that the limits exist. Note that \( \omega \)-negligibility is a joint property involving both \( f(x) \) and \( X_\omega \). It is seen that for a function \( f(x) \) to be absolutely \( \omega \)-negligible it is necessary that \( f(x) > 0, \quad \forall x \in X_\omega \). If the function \( t(x) = \mu_c(x) \equiv 1 \) is \( \omega \)-negligible then it is also absolutely \( \omega \)-negligible.

**Lemma A1:** With each permutation set \( Y \subseteq Y_n \subset \mathbb{N}^p \) and product measure \( \mu \) corresponds a unique set \( Y^f \subseteq Y_n^f \subset \mathbb{N}^{n+1} \), hereafter called frequency set, with

\[
Y^f_n = \left\{ y : \sum_{i=0}^n y_i = p \right\}, \tag{A.5}
\]

such that

\[
\mu(Y) = p! \sum_{y \in Y^f} \prod_{i=0}^n \mu(i)^{y_i}/y_i!. \tag{A.6}
\]

**Proof:** Let \( k_i(x) \) be the frequency with which a given outcome \( 0 \leq i \leq n \) occurs in a vector \( x \in Y \). In other words

\[
k_i(x) = \sum_{j=1}^{p} \delta(x_j - i), \quad i = 0, 1, \ldots, n, \tag{A.7}
\]

where the Kronecker \( \delta \) function is defined as

\[
\delta(0) = 1, \quad \delta(j) = 0, \quad j \in \mathbb{Z}\setminus\{0\}. \tag{A.8}
\]

Clearly for all \( x \in Y_n \), the vector

\[
k(x) = (k_0(x), k_1(x), \ldots, k_n(x))^T \tag{A.9}
\]

belongs to \( Y^f_n \). Define the frequency set \( Y^f \) as

\[
Y^f = \{ k(x) : x \in Y \}. \tag{A.10}
\]
Since $Y$ is a permutation set, the number of vectors $x \in Y$ mapped onto the same vector $y \in Y' \cap Y$ is given by the multinomial coefficient

$$\mu_c(\{x: k(x) = y\}) = p^1 / \prod_{j=0}^{n-1} y_j!,$$  \hspace{1cm} (A.11)$$

and the result follows from the fact that $\mu$ is a product measure.

**Lemma A2:** If $\tau(x) = \mu_c(x)$ is $\omega$-negligible and $g(x)$ is absolutely $\omega$-negligible then for all $f(x) \geq 0$,

$$\lim_{\omega \to \infty} \left[ \sum_{x \in X_\omega} f(x)g(x) \right]^{\frac{1}{\omega}} = \lim_{\omega \to \infty} \left[ \sup_{x \in X_\omega} f(x) \right]^{\frac{1}{\omega}}. \hspace{1cm} (A.12)$$

**Proof:** We consider first the case $g(x) = \tau(x)$. We have

$$\sup_{x \in X_\omega} f(x) \leq \sum_{x \in X_\omega} f(x) \leq \mu_c(X_\omega) \sup_{x \in X_\omega} f(x). \hspace{1cm} (A.13)$$

Since $\tau(x)$ is $\omega$-negligible, the result follows after taking the appropriate limits.

Since this is valid for all $f(x)$ we have in particular the following intermediate result:

$$\lim_{\omega \to \infty} \left[ \inf_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} = \lim_{\omega \to \infty} \left[ \sum_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} = \lim_{\omega \to \infty} \left[ \sup_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} = 1 \hspace{1cm} (A.14)$$

valid for all absolutely $\omega$-negligible functions $g(x)$.

Since

$$\inf_{x \in X_\omega} g(x) \sum_{x \in X_\omega} f(x) \leq \sum_{x \in X_\omega} f(x)g(x) \leq \sup_{x \in X_\omega} g(x) \sum_{x \in X_\omega} f(x), \hspace{1cm} (A.15)$$

the result follows after taking the appropriate limits.
Corollary 1: If \( t(x) \) is \( \omega \)-negligible and \( g(x) > 0 \) is a function such that

\[
\lim_{\omega \to \omega} \left[ \sum_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} \leq 1, \quad \lim_{\omega \to \omega} \left[ \sum_{x \in X_\omega} 1/g(x) \right]^{\frac{1}{\omega}} \leq 1,
\]

then \( g(x) \) is absolutely \( \omega \)-negligible.

**Proof:** We have

\[
\lim_{\omega \to \omega} \left[ \sup_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} \leq 1, \quad \lim_{\omega \to \omega} \left[ \sup_{x \in X_\omega} 1/g(x) \right]^{\frac{1}{\omega}} = \lim_{\omega \to \omega} \left[ 1/\inf_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} \leq 1,
\]

and hence

\[
\lim_{\omega \to \omega} \left[ \sup_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}} \leq 1 \leq \lim_{\omega \to \omega} \left[ \inf_{x \in X_\omega} g(x) \right]^{\frac{1}{\omega}},
\]

whence the result follows.

Corollary 2: If \( t(x) \) is \( \omega \)-negligible and the function \( g(x) \geq 1 \) is \( \omega \)-negligible then it is absolutely \( \omega \)-negligible.

**Proof:** This is simply a consequence of the preceding corollary and the fact that \( 1/g(x) \leq 1 \leq g(x) \).

Corollary 3: If \( t(x) \) is \( \omega \)-negligible and the functions \( g_1(x) \geq g_2(x) \geq 1 \) are such that \( g_1(x) \) is \( \omega \)-negligible then \( g_2(x) \) is absolutely \( \omega \)-negligible.

The proof is obvious.

Lemma A3: Let \( C_\omega \subseteq \mathbb{N}^{\omega+1} \) be the set
\[ C_\omega = \left\{ y : \sum_{i=0}^{n} y_i = \omega, \sum_{i=0}^{n} i y_i = n \right\} \quad (A.19) \]

and let \( f(x) \) be multiplicative i.e.

\[ f(y) = \prod_{i=0}^{n} \beta(y_i), \quad (A.20) \]

where

\[ \beta(0) = 1, \quad \beta(i) > 0 \quad i = 1, 2, \ldots \quad (A.21) \]

and

\[ \eta(z) = \sum_{i=0}^{\infty} \beta(i) z^i \quad (A.22) \]

is analytical with radius of convergence \( r \geq 1 \).

Then the following upper bound is valid:

\[ \ln \left[ \sum_{y \in C_\omega} f(y) \right] \leq \ln \eta(z) - \omega \ln z + 2\sqrt{nJ(z)}, \quad 0 < z < r, \quad (A.23) \]

where the increasing function \( J(z) \) is given by

\[ J(z) = \int_0^z \ln \eta(t) \frac{dt}{t}. \quad (A.24) \]

**Proof:** Consider the generating function

\[ F(q, z) = \sum_{y \in \mathbb{N}^{n+1}} \prod_{i=0}^{n} \beta(y_i) q^{\sum_{i=0}^{n} i y_i} z^{\sum_{i=0}^{n} y_i} = \sum_{m=0}^{\infty} \sum_{\omega=0}^{\infty} B_{m, \omega} q^{m} z^{\omega}. \quad (A.25) \]

Clearly

\[ B_{n, \omega} = \sum_{y \in C_\omega} f(y). \quad (A.26) \]

The generating function \( F(q, z) \) can be explicitly written as
\[ F(q, z) = \prod_{i=0}^{n} \eta(q^i z). \quad (A.27) \]

It is clear that \( F(q, z) \) is analytical for \(| q | < 1, \quad | z | < r \). Since all \( B_{n, \omega} \) are positive, it is straightforward to show that
\[ B_{n, \omega} \leq F(q, z)q^{-n}z^{-\omega}, \quad 0 < q < 1, \quad 0 < z < r. \quad (A.28) \]

Taking logarithms, we find
\[ \ln B_{n, \omega} \leq -n \ln q - \omega \ln z + \sum_{i=0}^{n} \ln \eta(q^i z). \quad (A.29) \]

Since \( \eta(z) \) is a strictly increasing function of \( z \) and \( q^i \) is a strictly decreasing function of \( i \), \( \ln \eta(q^i z) \) is a strictly decreasing function of \( i \). Hence we have
\[ \sum_{i=0}^{n} \ln \eta(q^i z) \leq \ln \eta(z) + \int_{0}^{n} \ln \eta(q^i z) dx \leq \ln \eta(z) - \frac{1}{\ln q} \int_{0}^{z} \ln \eta(t) \frac{dt}{t}, \quad (A.30) \]
where the last inequality follows from
\[ \eta(0) = \beta(0) = 1, \quad \lim_{t \to 0} \frac{\ln \eta(t)}{t} = \beta(1) > 0, \quad \eta(t) > 1 \text{ for } t > 0. \quad (A.31) \]

Putting \(- \ln q = u\) we have therefore
\[ \ln B_{n, \omega} \leq \ln \eta(z) + nu - \omega \ln z + \frac{1}{u} J(z), \quad 0 < z < r, \quad 0 < u < \infty. \quad (A.32) \]

For \( z \) fixed the minimum of \( J(z)/u + nu \) is obtained for \( u = \sqrt{J(z)/n} \), whence the result follows.

Note that we can be more specific when \( r > 1 \) since then the value \( z = 1 \) is allowed. In that case
\[ \ln B_{n,\omega} \leq \ln \eta(1) + 2\sqrt{nJ(1)}. \]  
\[ (A.33) \]

**Corollary 1:** Let \( \beta(i) = 1 + \kappa i, \quad i = 0, 1, 2, \ldots \), where \( \kappa \geq 0 \). Then

\[ \ln B_{n,\omega} \leq Z(\omega) + 2\sqrt{nJ(1)}, \]  
\[ (A.34) \]

where

\[ Z(\omega) = (\omega + 1) \ln(\omega + 1) - \omega \ln \omega, \]  
\[ (A.35) \]

when \( \kappa = 0 \) and

\[ Z(\omega) = (\omega + 2) \ln(\omega + 2) - \omega \ln \omega + \ln \left[ 1 + (\kappa - 1) \frac{\omega}{\omega + 2} \right] - 2 \ln 2, \]  
\[ (A.36) \]

when \( \kappa > 0 \).

In this context \( J(1) = \text{diln}(\kappa) - 2\text{diln}(0) \), where diln(x) is the dilogarithm \([22]\) defined by

\[ \text{diln}(x) = \int_1^x \frac{\ln t}{t - 1} \, dt. \]  
\[ (A.37) \]

**Proof:** Here

\[ \eta(z) = \sum_{i=0}^{\infty} (1 + \kappa i) z^i = \frac{1 + (\kappa - 1)z}{(1 - z)^2}. \]  
\[ (A.38) \]

Clearly \( r = 1 \). After some calculations we have

\[ J(z) = \text{diln}(1 - (1 - \kappa)z) - 2\text{diln}(1 - z). \]  
\[ (A.39) \]

Note that \( \text{diln}(x) \) is a strictly increasing function with \( \text{diln}(0) = -\pi^2/6 \) which implies that \( J(1) \geq \pi^2/6 \). Hence we have
\ln B_{n,x} \leq \ln \eta(z) - \omega \ln z + 2\sqrt{nJ(z)} \leq \ln \eta(z) - \omega \ln z + 2\sqrt{nJ(1)}. \quad (A.40)

Regarding the term
\ln \eta(z) - \omega \ln z = \ln[1 + (\kappa - 1)z] - 2\ln(1 - z) - \omega \ln z, \quad (A.41)
we take \( z = \omega/(\omega + 1) < 1 \) when \( \kappa = 0 \) and \( z = \omega/(\omega + 2) < 1 \) when \( \kappa > 0 \), whence the result follows.

**Corollary 2:** If
\[
\lim_{\omega \to \infty} \frac{\sqrt{n}}{\omega} = 0, \quad (A.42)
\]
then the function \( t(x) = \mu_c(x) \) is \( \omega \)-negligible with respect to \( C_\omega \).

**Proof:** This corresponds to the case \( \kappa = 0 \) in the above corollary. We have
\[
\ln \mu_c(C_\omega) \leq (\omega + 1) \ln(\omega + 1) - \omega \ln \omega + 2\pi\sqrt{n/6}. \quad (A.43)
\]
Since \( C_\omega \) is never empty, e.g. it always contains the point \( (\omega - 1, 0, \ldots, 0, 1)^T \), and since
\[
\lim_{\omega \to \infty} \frac{1}{\omega} [(\omega + 1) \ln(\omega + 1) - \omega \ln \omega] = \lim_{\omega \to \infty} \frac{\ln \omega}{\omega} = 0, \quad (A.44)
\]
we have
\[
1 \leq \lim_{\omega \to \infty} [\mu_c(C_\omega)]^{\frac{1}{\omega}} \leq \exp \left( \lim_{\omega \to \infty} \frac{2\pi}{\omega} \sqrt{\frac{n}{6}} \right) \quad (A.45)
\]
and the result follows.
**Corollary 3:** In the same context as Corollary 2 the multiplicative function defined by $\beta(i) = 1 + \kappa i$ is absolutely $\omega$-negligible.

**Proof:** When $\kappa > 0$ the dominating term of $Z(\omega)$ is $2 \ln \omega$ whence the result follows by Lemma A2, Corollary 2.

**Corollary 4:** In the same context as Corollary 2 the multiplicative function defined by $\beta(i) = \rho(i)$, $i \geq 0$, where

\[\rho(0) = 1, \quad \rho(i) = i! i^{-i} e^i, \quad i > 0,\]  

is absolutely $\omega$-negligible.

**Proof:** Since $0^0 = 1$, we have

\[\frac{\beta(i+1)}{\beta(i)} = e \left( \frac{i}{i+1} \right)^i, \quad i \geq 0.\]  

(A.47)

It is a simple matter to show that $\beta(i+1)/\beta(i)$ is strictly decreasing from $\beta(1)/\beta(0) = e$ to $\lim_{i \to \infty} \beta(i+1)/\beta(i) = 1$. Hence $\beta(i)$ is increasing and $\beta(i) \geq 1, \quad i \geq 0$. We next show that there is a constant $\kappa > 0$, such that $1 + \kappa i \geq \beta(i)$ for all $i \geq 0$. In other words

\[\kappa = \sup_{i \geq 1} \frac{\beta(i) - 1}{i} < \infty.\]  

(A.48)

That $\kappa < \infty$ follows from Stirling's formula, indicating that $\beta(i) \sim \sqrt{2\pi i}, \quad i \to \infty$.

Note that the odds are high that $\kappa = e - 1$, but what matters here is that $\kappa$ exists.

By Corollary 3 and Lemma A2, Corollary 3, the result follows.
Theorem A1: Contraction principle. Let $\mu$ be a strict product measure over $\mathcal{P}(\mathbb{N})$ and let $B_{\omega} \subseteq \mathbb{N}^{\omega}$ be the set
\[
B_{\omega} = \left\{ x : \sum_{i=1}^{\omega} x_i = n \right\},
\]
where $n$ is such that $\lim_{\omega \to \infty} n/\omega = t \geq 0$.

Let $\nu$ be a real probability measure over $\mathcal{P}(\mathbb{N})$ belonging to the admissible set
\[
E = \left\{ \nu : \sum_{i=0}^{\infty} i\nu(i) = t \right\}.
\]

Let further $I(\nu, \mu)$ be the Kullback-Leibler information
\[
I(\nu, \mu) = \sum_{\mu(i) > 0} \nu(i) \ln(\nu(i)/\mu(i)),
\]
where we suppose $0 \cdot \ln 0 = 0$.

Then
\[
\lim_{\omega \to \infty} \frac{1}{\omega} \ln \mu(B_{\omega}) = \sup_{\nu \in E} (-I(\nu, \mu)).
\]

Proof: Since $B_{\omega}$ is a permutation set, there exists by Lemma A1 a frequency set $B_{\omega}^f \subseteq \mathbb{N}^{n+1}$ defined by
\[
B_{\omega}^f = \left\{ y : \sum_{i=0}^{n} y_i = \omega, \sum_{i=0}^{n} i y_i = n \right\} = C_{\omega},
\]
with
\[
\mu(B_{\omega}) = \sum_{y \in C_{\omega}} \omega! \prod_{i=0}^{n} \mu(i)^{y_i}/y_i!.
\]
Note that, for this to make any sense, we should assign $y_i = 0$ whenever $\mu(i) = 0$. In order that $C_\omega \neq \emptyset$ it is therefore necessary that there be at least two values $i, j$ such that $\mu(i), \mu(j) > 0$. This is assured, since we have assumed that $\mu$ is strict. We have, with $\rho(i)$ defined as in Lemma A3 Corollary 4,

$$
\mu(B_\omega) = \rho(\omega) \sum_{y \in C_\omega} \left[ \prod_{i = 0}^{n} \frac{1}{\rho(y_i) \mu(i) > 0} \mu(i)^{y_i} \left( \frac{\omega}{y_i} \right)^{y_i} \right].
$$

(A.55)

Defining

$$
g(y) = \prod_{i = 0}^{n} \frac{1}{\rho(y_i)},
$$

(A.56)

and

$$
f(y) = \prod_{\mu(i) > 0} \mu(i)^{y_i} \left( \frac{\omega}{y_i} \right)^{y_i},
$$

(A.57)

we have

$$
\mu(B_\omega)^{\frac{1}{\alpha}} = \rho(\omega)^{\frac{1}{\alpha}} \left[ \sum_{y \in C_\omega} g(y) f(y) \right]^{\frac{1}{\alpha}}.
$$

(A.58)

By Lemma A3 Corollaries 2, 4, $u(y)$ and $1/g(y) \geq 1$ are absolutely $\omega$-negligible with respect to $C_\omega$, provided $\lim_{\omega \to \infty} \sqrt{n/\omega} = \lim_{\omega \to \infty} \sqrt{t/\omega} = 0$, which is the case. By the intermediate result in Lemma A2 the same can be said about $g(y)$. Since

$$
\lim_{\omega \to \infty} \rho(\omega)^{\frac{1}{\alpha}} = \lim_{\omega \to \infty} (2\pi\omega)^{\frac{1}{2\alpha}} = 1,
$$

(A.59)

application of Lemma A2 yields

$$
\lim_{\omega \to \infty} \mu(B_\omega)^{\frac{1}{\alpha}} = \lim_{\omega \to \infty} \left[ \sum_{y \in C_\omega} g(y) f(y) \right]^{\frac{1}{\alpha}} = \lim_{\omega \to \infty} \left[ \sup_{y \in C_\omega} f(y) \right]^{\frac{1}{\alpha}}.
$$

(A.60)
After taking logarithms this becomes

\[
\lim_{\omega \to \infty} \frac{1}{\ln \mu(B_o)} = \lim_{\omega \to \infty} \sup_{\omega \to \infty} \sum_{y \in C_\omega \mu(i) > 0} \frac{y_i}{\omega} \left[ \ln \mu(i) + \ln \left( \frac{\omega}{y_i} \right) \right] \\
= \sup_{y \in C_\omega \mu(i) > 0} \sum v(i) \left[ \ln \mu(i) - \ln v(i) \right] = \sup_{y \in C_\omega} -I(v, \mu) \tag{A.62}
\]

where \(v(i) = y_i/\omega\) are the relative frequencies, and the result follows since we can redefine \(C_\omega\) as

\[
C_\omega = \left\{ v : \sum_{i=0}^{n} v(i) = 1, \quad \sum_{i=0}^{n} iv(i) = n/\omega \right\} \tag{A.63}
\]

implying that \(C_\omega = E\).
REFERENCES


