Maximum Entropy and Conditional Probability

JAN M. VAN CAMPENHOUT AND THOMAS M. COVER, FELLOW, IEEE

***Abstract*—It is well-known that maximum entropy distributions, subject to appropriate moment constraints, arise in physics and mathematics. In an attempt to find a physical reason for the appearance of maximum entropy distributions, the following theorem is offered. The conditional distribution of X_1 given the empirical observation $(1/n)\Sigma_{i=1}^{n}\boldsymbol{h}(X_i)=\boldsymbol{\alpha}$, where $X_1, X_2, \cdots$ are independent identically distributed random variables with common density g converges to $f_{\boldsymbol{\lambda}}(x)=e^{\boldsymbol{\lambda}^t\boldsymbol{h}(x)}g(x)$ (suitably normalized), where $\boldsymbol{\lambda}$ is chosen to satisfy $\int f_{\boldsymbol{\lambda}}(x)\boldsymbol{h}(x)\,dx=\boldsymbol{\alpha}$. Thus the conditional distribution of a given random variable X is the (normalized) product of the maximum entropy distribution and the initial distribution. This distribution is the maximum entropy distribution when g is uniform. The proof of this and related results relies heavily on the work of Zabell and Lanford.**

I. INTRODUCTION

The differential entropy $H(X)$ of a random variable X with density function $f(x)$ (with respect to Lebesgue measure) is defined by $H(X)=-\int_{-\infty}^{+\infty}f(x)\ln f(x)\,dx$. All of the well-known distributions in statistics are maximum entropy distributions given appropriate simple moment constraints. For example, the maximum entropy distribution under the constraint $EX^2=\sigma^2$ is the normal distribution with mean 0 and variance σ^2. The maximum entropy nonnegative random variable with mean m is exponentially distributed with parameter $\lambda=1/m$. Even the Cauchy distribution is a maximum entropy distribution over all distributions satisfying $E\ln(1+X^2)=\alpha$. In general, the maximum entropy density $f(x)$ under the constraint $\int\boldsymbol{h}(x)f(x)\,dx=\boldsymbol{\alpha}$, where $\boldsymbol{h}$ is a vector-valued function of x, is of the form

$$f(x)=\exp\left(\lambda_0+\boldsymbol{\lambda}^t\boldsymbol{h}(x)\right). \tag{1}$$

The constants λ_0, $\boldsymbol{\lambda}$ are chosen so that $f(x)$ is normalized and satisfies the moment constraint. An easy proof of (1) based on a convexity argument can be found in Kagan *et al.* [1, Theorem 13.2.1, p. 409].

The entropy $H(X)$ is closely related to the disorder or uncertainty associated with making a realization of X. For that reason, maximizing the entropy is a method for finding distributions that represent high uncertainty or, equivalently, a state of high ignorance. For instance, in statistical mechanics, Boltzmann and others found the three-variate normal distribution of velocities in gases as a maximum entropy distribution under an energy constraint. Similarly, one can derive the $p(h)=\lambda e^{-\lambda h}$, $h\geq 0$, distribution of air density as a function of height in the earth's atmosphere under the mean potential energy constraint $\int hp(h)\,dh=E$.

In statistics, the principle of maximum entropy has been used to obtain "uninformative" prior distributions in Bayesian inference. A paper by Jaynes [2] discusses precisely this technique. Although the use of the maximum entropy principle for these purposes may seem ad hoc, maximum entropy distributions have some desirable properties. Jaynes [2] comments, "...the probability distribution which maximizes the entropy is numerically identical with the frequency distribution which can be realized in the greatest number of ways," thus associating maximum entropy with a definite frequency (or maximum likelihood) interpretation.

This note attaches another concrete meaning to the maximum entropy distribution. It characterizes such a distribution as the limit of a sequence of conditional distributions. It is shown that, under certain regularity conditions, the conditional distribution of the first random variable X_1 in a sequence of independent identically distribution (i.i.d.) random variables $X_1, X_2, \cdots$, given the empirical average $(1/n)\Sigma_1^n h(X_i)$, converges to a maximum entropy distribution. More precisely, the limiting distribution f maximizes $H_g(X)=-\int f(x)\ln(f(x)/g(x))\,dx$, the entropy relative to the initial distribution g of X_1, subject to the constraint that $\int h(x)f(x)\,dx$ equals the observed average. The quantity $-H_g(X)$ is also known as the Kullback–Leibler information number of f relative to g. Thus among all distributions satisfying the above moment constraint on $h(X_1)$, the limiting conditional distribution f of X_1 minimizes the Kullback–Leibler number with respect to g. It follows that f is closest to g in a certain hypothesis testing sense.

The convergence problem of conditional distributions has a long history. As early as 1922, in the then fully developing field of statistical mechanics, Darwin and Fowler [11] established their method to derive the energy distribution in large systems of particles with a given total energy. Through the computation of the average occupancy of the discrete energy levels, these authors arrived naturally at the classical energy distributions. Jaynes [12] relates Darwin and Fowler's work to the Shannon maximum entropy principle.

In an attempt to formalize statistical limits in statistical mechanics, Lanford [3] considers the convergence problem

Manuscript received August 22, 1978; revised October 2, 1980. This paper was supported by National Science Foundation under Grant ENG 76-03684 and JSEP DAAG 29-79-C-0047. It was presented at the IEEE International Symposium on Information Theory, Grignano, Italy, June 25–29, 1979.

J. M. Van Campenhout is with the Electronics Laboratory, the State University of Ghent, Ghent, Belgium.

T. M. Cover is with the Departments of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

0018-9448/81/0700-0483$00.75 ©1981 IEEE

of conditional distributions when an empirical average is conditioned in an interval that may or may not contain the mean of the underlying distribution. The same problem is considered by Bartfai [9] and Vincze [10] from a statistical point of view. Zabell [4], on the other hand, studies primarily the convergence of conditional expectations when the conditioning is pointwise, but the points are in the neighborhood of the true mean. We extend Zabell's work to conditioning at points "far" from the mean, and we will reinterpret the work of the above authors from a maximum entropy viewpoint.

In Section II, we study the convergence properties of the special case in which the random variables $X_1, X_2, \cdots$ take values in the set $\{1,2,\cdots,m\}$. For pointwise conditioning far from the mean, the use of Chernoff's tilting idea [5], [6] is clearly illustrated by this example. The idea will be used again in Section III, where we generalize Zabell's result to conditioning at points far from the mean. The convergence of conditional distributions, under the condition that X has only a finite number of mass points, can be obtained by application of Stirling's inequalities, as shown in the work of Vasicek [13]. In Section IV, we provide some examples in the case where the random variables have densities, and in Section V we review and interpret Lanford's work and its implications.

II. A Special Case

In this section we consider the case of discrete bounded random variables, and we give a direct proof of the following convergence theorem.

Theorem 1: Let $X_1, X_2, \cdots$ be i.i.d. discrete random variables with uniform probability mass function $p(x)$ on the range $x \in \{1,2,\cdots,m\}$. Then, for $1 \leq \alpha \leq m$, and for all $x_1 \in \{1,2,\cdots,m\}$, we have

$$\lim_{\substack{n\to\infty \\ n\alpha \text{ is integer}}} P\left\{X_1 = x_1 \middle| \frac{1}{n}\sum_{i=1}^{n} X_i = \alpha\right\} = p^*(x_1), \qquad (2)$$

where

$$p^*(x_1) = e^{\lambda x_1} \Big/ \left(\sum_{i=1}^{m} e^{\lambda i}\right) \qquad (3)$$

is the maximum entropy probability mass function under the constraint $\Sigma x p^*(x) = \alpha$, and λ is chosen to satisfy this constraint.

Proof: First we will prove that for *any* probability mass function $q(x) > 0$ on the range $\{1,2,\cdots,m\}$ with $\alpha = E_q X_1 = \Sigma x_1 q(x_1)$ we have

$$\lim_{\substack{n\to\infty \\ n\alpha \text{ is integer}}} q\left(x_1 \middle| \frac{1}{n}\sum_{i=1}^{n} X_i = \alpha\right) = q(x_1). \qquad (4)$$

Thus conditioning on the expected outcome has an asymptotically negligible effect. This proves (2) in the case where $\alpha = EX_1 = (m+1)/2$. The limiting distribution p^* is obtained by setting $\lambda = 0$ in (3).

We then consider (2) in the case $\alpha \neq (m+1)/2$. We use Chernoff's tilting idea to modify $p(x)$ so that we are again conditioning on the expected outcome. Since the conditional distributions in (2) are invariant under tilting, Theorem 1 will follow.

Let us turn to the proof of (4). Let $n\alpha$ be an integer such that $P\{(1/n)\Sigma_{i=1}^{n} X_i = \alpha\} > 0$. Letting $S_n = X_1 + X_2 + \cdots + X_n$, we have

$$\begin{aligned} P\{X_1 = j | S_n = n\alpha\} &= \frac{P\{X_1 = j, S_n = n\alpha\}}{P\{S_n = n\alpha\}} \\ &= \frac{q(j)P\{X_2 + \cdots + X_n = n\alpha - j\}}{P\{S_n = n\alpha\}} \\ &= q(j)\frac{P\{S_{n-1} = n\alpha - j\}}{P\{S_n = n\alpha\}}. \end{aligned} \qquad (5)$$

Therefore, if we can prove that $P\{S_{n-1} = n\alpha - j\}$ is asymptotically independent of $j \in \{1,2,\cdots,m\}$ as $n \to \infty$, it will follow that $P\{X_1 = j | S_n = n\alpha\} \to q(j)$. The desired result is contained in the following lemma.

Lemma 1: Let $X_1, X_2, \cdots$ be i.i.d. random variables with probability mass function $q(x) > 0$ on $x \in \{1,2,\cdots,m\}$. With $\alpha = EX_1$ and $S_n = X_1 + \cdots + X_n$, we have $P\{S_n = k\}/P\{S_n = k+1\} \to 1$ for all integers k satisfying $|n\alpha - k| < A$ for some constant A.

The proof of this lemma follows from a slight generalization of a problem in Chung [7, exercise 24, p. 177], and is, in fact, a form of the Chung–Erdös strong ratio limit theorem.

In the case that $\alpha \neq (m+1)/2$, let $\tilde{p}(x)$ be the tilted probability mass function derived from $p(x)$ as follows:

$$\tilde{p}(x) = ce^{\lambda x}, \qquad x \in \{1,2,\cdots,m\}, \qquad (6)$$

where c and λ are chosen to satisfy

$$\Sigma \tilde{p}(x) = 1, \quad \Sigma x \tilde{p}(x) = \alpha. \qquad (7)$$

Then clearly $\tilde{p}(x) > 0$ for all $x \in \{1,2,\cdots,m\}$; thus (4) is applicable. The properties of the tilting operation allow us to reconnect (4) with the original statement (2) as follows.

First we observe that

$$\begin{aligned} p(x_1,\cdots,x_n) &= \prod_{i=1}^{n} p(x_i) \\ &= \prod_{i=1}^{n} (e^{-\lambda x_i}/c)\tilde{p}(x_i) \\ &= c^{-n}e^{-\lambda \Sigma x_i}\tilde{p}(x_1,\cdots,x_n), \end{aligned}$$

and thus that

$$\begin{aligned} P\{S_n = n\alpha\} &= \sum_{x_1 + \cdots + x_n = n\alpha} p(x_1,\cdots,x_n) \\ &= c^{-n}e^{-\lambda n\alpha} \sum_{x_1 + \cdots + x_n = n\alpha} \tilde{p}(x_1,\cdots,x_n) \\ &= c^{-n}e^{-\lambda n\alpha}\tilde{P}\{S_n = n\alpha\}. \end{aligned} \qquad (8)$$

From (8) it follows easily that the tilting transformation leaves the conditional distributions in (2) invariant. This can be seen by using (5) and (8) as follows:

$$\begin{aligned} P\{X_1 = x_1 | S_n = n\alpha\} &= \frac{p(x_1)P\{S_{n-1} = n\alpha - x_1\}}{P\{S_n = n\alpha\}} \\ &= \frac{c^{-1}e^{-\lambda x_1}\tilde{p}(x_1)c^{1-n}e^{-\lambda(n\alpha - x_1)}\tilde{P}\{S_{n-1} = n\alpha - x_1\}}{c^{-n}e^{-\lambda n\alpha}\tilde{P}\{S_n = n\alpha\}} \\ &= \tilde{P}\{X_1 = x_1 | S_n = n\alpha\}. \qquad (9) \end{aligned}$$

But $\tilde{P}(X_1 = x_1 | S_n = n\alpha\} \to \tilde{P}(X_1 = x_1) = e^{\lambda x_1}/(\Sigma e^{\lambda i})$, by (4), thus proving Theorem 1.

Remark: The smooth behavior of the probability distribution of S_n at small deviations from its mean is crucial to the convergence in (4). These ideas are also borne out by the restrictions on the random variables X_i imposed by Zabell in the more general case (see Section III). Unlike the central limit theorem or the law of large numbers, the additional conditions deal with the fine structure of S_n/n in the sense that deviations of the order $1/n$ from the mean are considered.

III. A Limit Theorem for Pointwise Conditioning

We proceed with the generalization of the special case studied in Section II to lattice random variables that may be unbounded and to random variables with density functions. We start by reminding the reader of Zabell's results [4] concerning the convergence of conditional expectations. We then apply the tilting transformation to these results to obtain the desired generalization.

Zabell considers a sequence $U, X_1, X_2, \cdots$ of random variables where U has finite expectation and the pair (U, X_1) is independent of $X_2, X_3, \cdots$. He derives a set of sufficient conditions under which it is true that

$$E(U | X_1 + \cdots + X_n = A_n) \to E(U), \qquad (10)$$

as $n \to \infty$. These conditions can be summarized as follows.

1) The random variables $X_1, X_2, \cdots$ take values in the same additive subgroup of $\mathbb{R}$ (or, more generally, in the same coset of an additive subgroup of $\mathbb{R}$).

2) Consider the normalized sums $Y_n = (X_2 + \cdots + X_n - A_n)/B_n$. Then there exist sequences $\{A_n\}, \{B_n\}$ with $B_n \to \infty$ such that Y_n converges in distribution to a (nondegenerate) random variable Y.

3) Let $\psi_n(t) = E(\exp(itY_n))$ and $\psi(t) = E(\exp(itY))$ denote the characteristic functions of Y_n and Y, respectively. Then either a) ψ_n is periodic, and $P\{Y_n = 0\} > 0$ for n sufficiently large; or b) ψ is absolutely integrable, ψ_n is absolutely integrable for n sufficiently large, $\psi_n \to \psi$ in L_1, and $\int \psi > 0$.

Condition 1) is the generalization of the regularity properties of S_n derived in Section II. For i.i.d. random variables X_i, $i = 1, 2, \cdots$, this condition together with condition 3) implies that the random variables are either of the lattice type (i.e., the X_i take values in $\{a + kb\colon k = 0, \pm 1, \pm 2, \cdots\}$) or are real-valued and have a density function $f(x)$ with respect to Lebesgue measure.

For example, these conditions preclude the case $X_i \in \{0, \pi, 5\}$. In this case the event $\{X_1 + \cdots + X_n = n\pi\}$ implies the event $\{X_1 = \pi\}$, for all probability mass functions $p(x_1)$. Thus clearly for all n,

$$p(x_1 | X_1 + \cdots + X_n = n\pi) = \begin{cases} 1, & x_1 = \pi \\ 0, & \text{otherwise,} \end{cases}$$

and Theorem 1 fails to hold even if π were the true mean of $p(x_1)$.

Letting U range over all bounded continuous functions of X_1, we see that Zabell's work implies the convergence of the conditional probability distribution of X_1 to its unconditional distribution (see, for example, Chung [7, Theorem 4.4.2, p. 89]).

We now limit our attention to i.i.d. random variables $X_1, X_2, \cdots$ having a density function $f(x)$ with respect to Lebesgue measure. Rather than conditioning on $\Sigma_1^n X_i$, we consider a (Borel-measurable) function $h\colon \mathbb{R} \to \mathbb{R}$ and condition on $S_n = \Sigma_1^n h(X_i)$. We assume that $h(X_1)$ has a density with mean μ. The case of discrete conditioning variables is completely analogous and was covered to some extent in Section II.

It follows from (10) and conditions 1)–3) that the centering constants A_n must be "close" to $n\mu$ in order to ensure the nondegeneracy of the limit of $(S_n - A_n)/B_n$. Conditioning on the event $\{(1/n)S_n = \alpha\}$, $\alpha \neq \mu$, results in centering constants $n\alpha$ that are too far from $n\mu$ for (10) to hold. Under certain restrictions, an application of Chernoff's tilting idea allows us to move the mean of $h(X_1)$ to the conditioning point α, rendering (10) applicable. Again, the tilting leaves the conditional distributions invariant and thus provides us with the limit of these conditional distributions, conditioned at points off the true mean.

More concretely, let $f(x)$ and $g(t)$ denote the probability densities of X_1 and $h(X_1)$, respectively. Consider the exponential family $\mathcal{G}$ of densities indexed by λ,

$$\mathcal{G} = \left\{ e^{\lambda t}g(t)/c(\lambda)\colon c(\lambda) = \int e^{\lambda t}g(t)\,dt < \infty \right\}. \qquad (11)$$

Assume that $\mathcal{G}$ contains a density $g^*(t) = e^{\lambda t}g(t)/c(\lambda)$ with mean α. The desired tilting operation, then, changes the underlying probability measure P to a measure P^*, under which $h(X_1)$ has the density $g^*(t)$. One can easily verify that changing the density of X_1 to

$$f^*(x) = e^{\lambda h(x)}f(x)/c(\lambda) \qquad (12)$$

induces the density $g^*(t)$ on $h(X_1)$.

Thus, applying (10) under the measure P^* induced by f^*, we have for all bounded continuous functions $U(\cdot)$,

$$E^*\left(U(X_1) \middle| \sum_{i=1}^{n} h(X_i) = n\alpha\right) \to E^*(U(X_1)), \qquad (13)$$

where $E^*(U(X_1)) = (\int U(x)e^{\lambda h(x)}f(x)\,dx)/c(\lambda)$. The conditional expectation in (13) can be written as

$$E^*(U(X_1) | S_n = n\alpha) = \int U(x)P^*\{X_1 \in dx | S_n = n\alpha\}. \qquad (14)$$

On the support set of S_n, the conditional distribution of X_1 can be written in terms of the conditional density of X_1 as

$$P^*\{X_1 \in dx | S_n = t\} = f^*(x) g^*_{n-1}(t - h(x))\, dx / g^*_n(t), \tag{15}$$

where $g^*_n(t)$ denotes the tilted density of $S_n = \Sigma_1^n h(X_i)$. Equation (15) can be directly verified by the defining relation for conditional distributions

$$P^*\{X_1 \in A, S_n \in B\} = \int_B \left(\int_A P^*\{X_1 \in dx | S_n = t\} \right) g^*_n(t)\, dt. \tag{16}$$

As in Section II, it is easy to verify that the tilting transformation and convolutions commute, i.e.,

$$g^*_n(t) = (c(\lambda))^{-n} e^{\lambda t} g_n(t). \tag{17}$$

From (12), (15), and (17), it follows readily that

$$P^*\{X_1 \in dx | S_n = n\alpha\} = P\{X_1 \in dx | S_n = n\alpha\}, \tag{18}$$

and thus that

$$E^*\{U(X_1) | S_n = n\alpha\} = E\{U(X_1) | S_n = n\alpha\}, \tag{19}$$

from which it follows that

$$E\{U(X_1) | S_n = n\alpha\} \to \left(\int U(x) e^{\lambda h(x)} f(x)\, dx \right) / c(\lambda), \tag{20}$$

the desired extension of Zabell's result (10). Thus we have proved the following.

Theorem 2: Let $X_1, X_2, \cdots$ be i.i.d. random variables with density $f(x)$, and let $h: \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Let the random variables $h(X_1), h(X_2), \cdots$ have a density $g(t)$. If there exists a real number λ such that

$$c(\lambda) = E \exp(\lambda h(X_1)) < \infty,$$

and

$$\alpha = \left(\int h(x) e^{\lambda h(x)} f(x)\, dx \right) / c(\lambda),$$

and furthermore, if $g_\lambda(t) = e^{\lambda t} g(t)/c(\lambda)$ satisfies Zabell's conditions 1)–3), then as $n \to \infty$,

$$P\left\{ X_1 \le x \middle| \sum_{i=1}^n h(X_i) = n\alpha \right\} \to \left(\int_{-\infty}^x e^{\lambda h(x)} f(x)\, dx \right) / c(\lambda).$$

IV. Some Examples

We now give a few easy examples to illustrate Theorem 2. Note that Examples 1, 2, and 3 are established by direct calculation rather than by using Theorem 2.

Example 1: Gaussian Random Variables

Let $\phi(x; \mu, \sigma^2)$ denote the normal density with mean μ and variance σ^2. Let $X_1, X_2, \cdots$ be i.i.d. with density $\phi(x; 0, 1)$ and let $f(x | \Sigma_1^n X_i = n\alpha)$ denote the conditional density of X_1 given $\Sigma_1^n X_i = n\alpha$. The sum $S_n = X_1 + \cdots + X_n$ has density $\phi(x; 0, n)$, and according to (15) we have

$$f(x | S_n = n\alpha) = \phi(x; 0, 1) \phi(n\alpha - x; 0, n-1) / \phi(n\alpha; 0, n) = \phi\left(x; \alpha, \frac{n-1}{n} \right).$$

Thus

$$f(x | S_n = n\alpha) \to \phi(x; \alpha, 1) = e^{x\alpha} \phi(x; 0, 1) / e^{\alpha^2/2},$$

in accordance with Theorem 2.

Example 2: Exponential Random Variables

Let $X_1, X_2, \cdots$ be i.i.d. exponential random variables with parameter λ. Let $f_g(x; n, \mu)$ denote the gamma density with parameters n, μ. Since S_n has a gamma $(n, n\lambda)$ distribution we have

$$f(x | S_n = n\alpha) = \frac{f_g(x; 1, \lambda) f_g(n\alpha - x; n-1, \lambda)}{f_g(n\alpha; n, \lambda)} = \frac{n-1}{n\alpha} \left(\frac{n\alpha - x}{n\alpha} \right)^{n-2}.$$

Thus

$$f(x | S_n = n\alpha) \to \alpha^{-1} \exp(-x/\alpha) = \exp((\lambda - 1/\alpha)x) f_g(x; 1, \lambda)/\lambda\alpha.$$

Example 3: An Exception (Cauchy Random Variables $(E|X| = \infty)$)

Let $X_1, X_2, \cdots$ be i.i.d. with Cauchy density $f_c(x; 0, 1)$, where $f_c(x; \alpha, \beta) = \beta/\pi(\beta^2 + (x - \alpha)^2)$. It is well known that S_n has density $f_c(x; 0, n)$. Thus we find

$$\begin{aligned} f(x | S_n = n\alpha) &= f_c(x; 0, 1) f_c(n\alpha - x; 0, n-1) / f_c(n\alpha; 0, n) \\ &= f_c(x; 0, 1)(n-1)\left(n^2 + (n\alpha)^2\right) \\ &\quad / n\left((n-1)^2 + (n\alpha - x)^2\right). \end{aligned}$$

It follows that for every finite value of α we have

$$f(x | S_n = n\alpha) \to f_c(x; 0, 1),$$

pointwise in x. Thus conditioning on *any* α has an asymptotically negligible effect.

The reason for this exceptional behavior is that X_1 has no mean and thus Theorem 2 is not applicable. It should be noted that even if $E|X_1|^k < \infty$ for some $k > 1$ but $P\{X_1 < -A\}$ and $P\{X_1 > A\}$ approach zero less than exponentially fast, Zabell's result is applicable, but not its extension in Theorem 2. In this case the exponential family $\mathcal{G}$ in (11) contains only one density corresponding to $\lambda = 0$. No tilted density $g^*(t)$ with mean α can be found.

Example 4: The Maxwell–Boltzmann Distribution

Let the velocities $V_1, V_2, \cdots$ be i.i.d. vector valued random variables (r.v.), each drawn according to a uniform distribution over the cube $[-A, A]^3$. Then, by Theorem 2,

$$f\left(v \Big| \frac{1}{n}\sum_{i=1}^{n}\|V_i\|^2 = E\right) \to ce^{-\|v\|^2/2E}, \qquad v \in [-A, A]^3.$$

Thus the limiting density is the multivariate normal density truncated to the prior range.

V. Conditioning on Intervals

In this section we review the limiting behavior of conditional distribution of a random variable X_1, given that the empirical average of n independent observations $h(X_i)$, $i = 1, 2, \cdots, n$ lies in an interval (a, b). Although the results presented here have the same flavor as the results discussed in Section III, they are quite distinct. For instance, the rather strong regularity conditions on S_n imposed by Zabell are absent here. (Essentially what is left is the additional condition allowing tilting imposed in Theorem 2.) Thus Zabell's result cannot be obtained from the results established in this section. Conversely, one might be tempted to find the limit of $P\{X_1 \le x | a < n^{-1}\Sigma h(X_i) < b\}$ through an integration of $P\{X_1 \le x | n^{-1}\Sigma h(X_i) = t\}$ over $t \in (a, b)$. In order to do so, however, one would have to know the limiting distribution of $n^{-1}\Sigma h(X_i)$ on the interval (a, b), and furthermore one would have to verify the interchange of limits and integration as $n \to \infty$. Thus, again, the result in Theorem 2 is insufficient to provide the solution.

In this section, a direct approach is taken toward the identification of $\lim_{n\to\infty} P\{X_1 \le x | a < n^{-1}\Sigma h(X_i) < b\}$. In contrast to the seemingly arbitrary way in which tilting was introduced in the previous sections, this operation will now appear quite naturally in a much different context.

Let the function h: $\mathbb{R} \to \mathbb{R}$ be a bounded Borel measurable function, and let A and B denote the (essential) infimum and supremum of $h(X_1)$, respectively. Since the more general case of unbounded and vector-valued h-functions is discussed in Lanford's work, we shall limit ourselves to a simple case of bounded scalar h functions. We have the following result.

Theorem 3 (Lanford, 1973): Let $X_1, X_2, \cdots$ be i.i.d. random variables and let h: $\mathbb{R} \to \mathbb{R}$ be a bounded Borel measurable function. For $\text{ess.}\inf h(X_1) < a < b < \text{ess.}\sup h(X_1)$ define the distribution function $F_\lambda(x)$ by

$$F_\lambda(x) = \left(\int_{-\infty}^{x} e^{\lambda h(x)} P\{X_1 \in dx\}\right)/c(\lambda), \qquad (21)$$

where $c(\lambda) = Ee^{\lambda h(X)}$ and λ is chosen so that

$$\int_{-\infty}^{+\infty} h(x)\,dF_\lambda(x) = \begin{cases} b, & b < Eh(X_1) \\ Eh(X_1), & a \le Eh(X_1) \le b \\ a, & a > Eh(X_1). \end{cases}$$

Then, as $n \to \infty$, and for all continuity points x of $F_\lambda(x)$,

$$P\left\{X_1 \le x | a < \frac{1}{n}\sum_{i=1}^{n} h(X_i) < b\right\} \to F_\lambda(x). \qquad (22)$$

Thus Theorem 3 implies that if $a < Eh(X_1) < b$, then $\lambda = 0$ and $F_\lambda(x) = F(x) = P\{X_1 \le x\}$. That is, the conditioning on the interval (a, b) has an asymptotically negligible effect. This statement can be directly verified, since by the law of large numbers we have

$$P\left\{a < \frac{1}{n}\sum_{i=1}^{n} h(X_i) < b\right\} \to 1$$

as $n \to \infty$. Since furthermore $P(B|A) = P\{B \cap A)/P(A) \to P(B)$ as $P(A) \to 1$, the theorem follows.

However, if $Eh(X_1) \notin (a, b)$, then this reasoning is not applicable. Theorem 3 asserts that the conditional distribution of X_1 still converges and identifies the limiting distribution $F_\lambda(x)$ as belonging to the exponential family associated with $F(x)$. The distribution F_λ is the closest to F in the Kullback–Leibler sense, of all distributions $F^*(x)$ absolutely continuous with respect to $F(x)$ ($F^* \ll F$) and agreeing with the "asymptotic evidence" $a < (1/n)\Sigma h(X_1) < b$. More precisely, F_λ maximizes the F-relative entropy, or minimizes the Kullback–Leibler number

$$K(F^*, F) = \int \log \frac{dF^*(x)}{dF(x)}\,dF^*(x) \qquad (23)$$

over all distributions $F^* \ll F$ for which $\int h(x)\,dF^*(x) \in [a, b]$.

We will now outline a proof of Theorem 3. The arguments presented here are extracted from Lanford's work [3] and its extension by Bahadur and Zabell [8], and we refer to this work for details. The proof of Theorem 3 rests essentially on an extension of the asymptotic theory of tail probabilities to the probabilities of arbitrary open convex sets. The exponential decay of these probabilities is established in the following lemma.

Lemma 2: Let $Y_1, Y_1, \cdots$ be i.i.d. bounded random variables taking values in $\mathbb{R}^k$. Let J be a finite union of open convex sets of $\mathbb{R}^k$. Then

1) $S(Y; J) = \lim n^{-1} \log P\{n^{-1}\Sigma_{j=1}^{n} Y_j \in J\}$ exists (possibly infinite);
2) with $s(Y; x) = \inf_J\{S(Y; J): x \in J, J \text{ open convex}\}$ we have $S(Y; J) = \sup_{x \in J} s(Y; x)$.

The set function $\sup_{x\in J} s(Y; x)$ is known as the Lanford entropy of J.

Let μ denote the measure on $\mathbb{R}^k$ induced by Y_1, and define the function σ: $\mathbb{R}^k \to \mathbb{R}$ by

$$\sigma(y) = -\inf\left\{K(\nu; \mu): \int_{\mathbb{R}^k} t\nu(dt) = y, \nu \ll \mu\right\}, \qquad (24)$$

where, as before, $K(\nu; \mu) = \int \log(d\nu/d\mu)\,d\nu$ is the Kullback–Leibler number between ν and μ. Thus $-\sigma(y)$ is the minimum Kullback–Leibler number between the measure μ and any measure $\nu \ll \mu$ that has expectation y. It is well known that $\sigma(y) \le 0$ and $\sigma(EY_1) = 0$.

The function $\sigma(y)$ is useful in that it allows us to compute the Lanford point entropy $s(x)$ as a Kullback–Leibler number. Letting C denote the convex closure of the support of Y_1, we have

Lemma 3 (Bahadur and Zabell, [8]): For every x not on the boundary of C we have $\sigma(x) = s(Y; x)$. If Y_1 is one-dimensional then $\sigma(x) = s(Y; x)$ for all x.

An important fact is that for a bounded random variable Y_1, the infimum in (24) is attained by the measure

$$\nu(dx) = e^{\lambda' x}\mu(dx)/C(\lambda), \tag{25}$$

where $c(\lambda) = E\exp(\lambda' Y_1)$ and the vector λ is chosen such that

$$\left(\int_{\mathbb{R}^k} x e^{\lambda' x}\mu(dx)\right)/c(\lambda) = y.$$

Thus

$$\sigma(y) = \log c(\lambda) - \lambda' y, \tag{26}$$

and it can be shown that $\sigma(y)$ is a strictly concave function of y.

We next outline the proof of Theorem 3. Let $f: \mathbb{R} \to \mathbb{R}$ be a bounded continuous function, and consider the two-dimensional bounded random variables $Y_i = (f(X_i), h(X_i))$, $i = 1, 2, \cdots$. Let $J \subset \mathbb{R}^2$ be given by the open rectangle $\mathbb{R} \times (a, b)$, and letting $\overline{Y}_n = n^{-1}\Sigma_{i=1}^n Y_i$, consider the conditional expectation

$$E\left(n^{-1}\sum_{i=1}^n f(X_j)|\overline{Y}_n \in J\right) = (\mu_n(J))^{-1}\int_J x\,d\mu_n(x, y), \tag{27}$$

where μ_n is the measure induced on $\mathbb{R}^2$ by $\overline{Y}_n$. According to Lemma 2, $\lim n^{-1}\log\mu_n(J) = S(Y; J)$ exists and is given by $\sup_{x \in J} s(Y; J)$. By the concavity of $s(Y; x)$ and the boundedness of $\overline{Y}_n$, the supremum above is attained at some finite point $y_0 = (f_0, h_0)$ in the closure of J. With $\epsilon > 0$ let J' denote the set $(-\infty, f_0 - \epsilon) \times (a, b) \cup (f_0 + \epsilon, \infty) \times (a, b)$, a finite union of open convex sets. Again, $S(Y; J') = \lim n^{-1}\log\mu_n(J') = \lim n^{-1}\log P\{a < n^{-1}\sum h(X_i) < b, |f_0 - n^{-1}\sum f(X_j)| > \epsilon\}$ exists. However, by the strict concavity of s on J, we have $S(Y; J') < S(Y; J)$, and thus for all $\epsilon > 0$, $\mu_n(J')/\mu_n(J) \to 0$ as $n \to \infty$. Since f is bounded it follows that

$$E\left(n^{-1}\sum_{i=1}^n f(X_j)|\overline{Y}_n \in J\right) \to f_0. \tag{28}$$

By symmetry it is clear that $E(f(X_i)|\overline{Y}_n \in J) = E(f(X_j)|\overline{Y}_n \in J)$, for all i, j, and hence we obtain

$$E\left(f(X_1)|\overline{Y}_n \in J\right) \to f_0. \tag{29}$$

To complete the proof, we have to identify the point (f_0, h_0) of S where s attains its maximum. Since s is a strictly concave function attaining the global maximum $s = 0$ at $(Ef(X_1), Eh(X_1))$, we have

$$h_0 = \begin{cases} b, & b \le Eh(X_1), \\ Eh(X_1), & a < Eh(X_1) < b, \\ a, & a \ge Eh(X_1). \end{cases} \tag{30}$$

In order to identify f_0, consider the probability distribution $dF_\lambda(x) = e^{\lambda h(x)}dF(x)/c(\lambda)$, which through $f(X_1)$ and $h(X_1)$, induces the measure

$$\mu_\lambda\{dx, dy\} = e^{\lambda y}\mu\{dx, dy\}/c(\lambda) \tag{31}$$

on $\mathbb{R}^2$. Choosing λ such that $h_0 = \int y\,d\mu_\lambda(x, y)$ and letting $\hat{f}_0 = \int x\,d\mu_\lambda(x, y)$, we have according to (26) that

$$\begin{aligned} s\left(Y; (\hat{f}_0, h_0)\right) &= \log c(\lambda) - (0, \lambda)(\hat{f}_0, h_0)^t \\ &= s\left(h(X_1); h_0\right). \end{aligned} \tag{32}$$

However, since for all x $s(Y; (x, h_0)) \le s(h(X_1); h_0)$, it follows that the (unique) value f_0 of x maximizing $s(Y; (x, h_0))$ is precisely equal to $\hat{f}_0$. Thus

$$f_0 = \hat{f}_0 \int x\,d\mu_\lambda(x, y) = \int f(t)e^{\lambda h(t)}\,dF(t)/c(\lambda) \tag{33}$$

Since f is an arbitrary bounded continuous function, the proof of Theorem 3 follows.

VI. Conclusion

Suppose that we have observed a large number of similarly drawn (i.i.d.) random variables $X_1, X_2, \cdots, X_n$. An observer suggests that the function $\boldsymbol{h}(x) \in \mathbb{R}^d$ contains all the "useful" information in x. We observe the empirical average $\bar{\boldsymbol{h}}_n = (1/n)\Sigma_{i=1}^n \boldsymbol{h}(X_i)$ and are disconcerted to find that $\bar{\boldsymbol{h}}_n \neq E\boldsymbol{h}(X_1)$. We did not observe what we expected to observe. Obviously the conditional distribution of X_1 will be affected. The results of this paper show that the conditional distribution f of X_1 is given asymptotically by the closest distribution f (in the Kullback–Leibler sense) to the initial distribution g over all distributions f satisfying the observed constraint

$$\int f(x)\boldsymbol{h}(x)\,dx = \bar{\boldsymbol{h}}_n. \tag{34}$$

We can then say the following.

1) The conditional distribution is the maximum entropy distribution (relative to the initial distribution).

2) The conditional distribution f is the most difficult to distinguish from the initial distribution over all f satisfying (34). This interpretation follows from the well-known result that the Kullback–Leibler number $K(f; g)$ is the exponent in the probability of error in the two-hypothesis test f versus g when the probability of error under g is fixed. Thus low values of K result in slow convergence of the probability of error.

3) The explicit form of the limiting conditional density $f(x_1|\bar{\boldsymbol{h}}_n = \boldsymbol{\alpha})$ is given explicitly, under mild conditions, by

$$f\left(x|\bar{\boldsymbol{h}}_n = \boldsymbol{\alpha}\right) = ce^{\lambda' \boldsymbol{h}(x)}g(x).$$

Thus f is the normalized product of a maximum entropy density $e^{\lambda' \boldsymbol{h}(x)}$ and the initial density g.

We may ask why maximum entropy distributions arise naturally in physics. The answer is that physicists have identified good observation functions $\boldsymbol{h}(x)$. Moreover independence of the observations holds asymptotically; and the sample size n is large. In effect the physicist has identified

$(1/n)\Sigma_{i=1}^{n}\boldsymbol{h}(X_i)$ as a sufficient statistic. In short $\bar{\boldsymbol{h}}_n$ summarizes all that is physical about the sample; and, given $\bar{\boldsymbol{h}}_n$, the sample is maximally random and conveys no further physical information.

References

[1] A. M. Kagan, Yu. V. Linnik, and C. R. Rao, *Characterization Problems in Mathematical Statistics.* New York: Wiley, 1973.

[2] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, no. 3, Sept. 1968.

[3] O. E. Lanford, "Entropy and equilibrium states in classical statistical mechanics," in *Statistical Mechanics and Mathematical Problems.* Berlin: Springer-Verlag, 1973.

[4] S. L. Zabell, "A limit theorem for conditional expectations with applications to probability theory and statistical mechanics," Ph.D. dissertation, Harvard University, Cambridge, MA, Aug. 1974.

[5] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493-507, 1952.

[6] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[7] K. L. Chung, *A Course in Probability Theory.* New York: Academic, 1974.

[8] R. R. Bahadur and S. L. Zabell, "Large deviations of the sample mean in general vector spaces," *Ann. Probability*, vol. 7, no. 4, pp. 587-621, Aug. 1979

[9] P. Bartfai, "On a conditional limit theorem," *Coll. Math. Soc. J. Bolyai*, vol. 9, European Meeting of Statisticians, Budapest, pp. 85-91, 1972.

[10] I. Vincze, "On the maximum probability principle in statistical physics," *Coll. Math. Soc. J. Bolyai*, vol. 9, European Meeting of Statisticians, Budapest, pp. 869-893, 1972.

[11] C. G. Darwin and R. H. Fowler, "On the partition of energy," *Phil. Mag.*, vol. 44, pp. 450-479, 1922.

[12] E. T. Jaynes, "Probability theory in science and engineering," Lecture notes issued by the Socony-Mobil Research Labs, Dallas, TX, pp. 137-138, 1958.

[13] O. A. Vasicek, "A conditional law of large numbers," *Ann. Probability*, vol. 8, no. 1, pp. 142-147, 1980.

Error Performance of Differentially Coherent Detection of Binary DPSK Data Transmission on the Hard-Limiting Satellite Channel

JHONG S. LEE, MEMBER, IEEE, ROBERT H. FRENCH, MEMBER, IEEE, AND YOON K. HONG, MEMBER, IEEE

***Abstract*—The error performance of differentially coherent detection of a binary differential phase-shift keying (DPSK) system operating over a hard-limiting satellite channel is derived. The main objective is to show the extent of error rate degradation of a DPSK system when a power imbalance exists between the two symbol pulses that are used in a bit decision interval. Consideration is also given to the DPSK error rate performance for the special case of *uncorrelated* uplink and *correlated* downlink noises at the sampling instants in adjacent time slots. Error probabilities are given as functions of uplink signal-to-noise ratio (SNR) and downlink SNR with different levels of SNR imbalance and different downlink SNR and uplink SNR as parameters, respectively. Our numerical results show that 1) as long as the symbols are equiprobable, the error probability is not dependent upon the downlink noise correlation, regardless of whether there is a power imbalance; 2) error performance is definitely affected by the power imbalance for all cases of symbol distributions; and 3) the error probability does depend upon downlink noise correlation for all levels of power imbalance if the symbol probabilities are not equal.**

I. Introduction

In some applications, a modem employing differentially encoded phase-shift-keyed (DPSK) signal transmission with differentially coherent demodulation is an appropriate choice when the circuit simplicity and the accompanying cost effectiveness are the overriding considerations in a data communication system. A typical application might be an expendable buoy with a cost constraint that requires a modem capability for data transmission and command-signal reception.

In this paper we consider such a DPSK system operating over the hard-limiting satellite channel. The main purpose of the paper is to show the extent of error-performance degradation of the DPSK system under the following two assumptions: 1) there exists a power imbalance between two symbol pulses that are used in a bit decision interval, and 2) the sample values of the downlink additive noise at the sampling instants in adjacent time slots are correlated whereas the sampled version of the uplink noises are independent. The motivation for considering the effects of these assumptions on the DPSK error performance is based

Manuscript received October 18, 1979; revised August 18, 1980. This work was supported by the Probability and Statistics Program of the Office of Naval Research under Contract N00014-77-C-0056. A portion of this paper was presented at the Tenth Annual Pittsburgh Conference on Modeling and Simulation, Pittsburgh, PA, April 25-27, 1979.

The authors are with J. S. Lee Associates, Inc., 2001 Jefferson Davis Highway, Arlington, VA 22202.

0018-9448/81/0700-0489$00.75 ©1981 IEEE