Ergodicity and Fourth-Order Spectral Moments

Bernard Picinbono, *Fellow, IEEE*

Abstract—Relationships between ergodicity and structures of fourth-order spectral moments are investigated. In particular it is shown that second-order ergodicity of a random process is directly related to the distribution of these moments on the normal manifolds of the frequency domain. This result is illustrated by various examples.

Index Terms—Ergodicity, normal processes, shot noise, stationary and normal manifolds, trispectrum.

I. INTRODUCTION AND NOTATIONS

Higher order statistics, or, more precisely, statistics of order higher than two, are used in various areas of signal processing and information sciences. The purpose of this correspondence is to show that they can also be useful for testing ergodic properties of some stochastic processes (SP). From a strictly mathematical point of view this has been known for a long time [1], [2]. However, these papers are written in a very abstract form, and the engineering community has considered the same problem again with a quite different approach, and especially in the case of harmonic processes [3], [4]. This correspondence has an intermediary perspective: it is not especially mathematically oriented and not limited to harmonic processes.

Let $x(t)$ be a continuous-time real and strictly stationary SP. In order to estimate its mean value m it is appropriate to use its causal time average $y_T(t)$ given by

$$y_T(t) = \frac{1}{T} \int_{t-T}^t x(\theta) d\theta. \quad (1)$$

It is said that $x(t)$ is *weakly first-order* (WFO) ergodic if $y_T(t)$ tends to m in the quadratic mean sense when T tends to infinity. The characterization of this kind of ergodicity requires only the knowledge of second-order properties of $x(t)$. This means that conditions for ergodicity can be deduced from the behavior of the correlation function

$$\gamma(\tau) = E[x(t)x(t-\tau)] - m^2$$

or from the structure of its Fourier transform $\Gamma(f)$, the power spectrum of $x(t)$. In order to simplify the presentation we assume that this spectrum can be expressed as

$$\Gamma(f) = \Gamma_b(f) + \sum_{i=1}^N \gamma_i \delta(f - f_i) \quad (2)$$

where $\Gamma_b(f)$ is nonnegative, bounded, and integrable. Physically, this means that the spectrum of $x(t)$ contains a continuous part, with a density $\Gamma_b(f)$, and a finite number of spectral lines at frequencies f_i . In this case it is known that $x(t)$ is WFO ergodic if and only if there is no spectral line at the frequency zero [5].

It is obvious that WFO ergodicity does not imply that the time average of $x(t)x(t-s)$ converges to the ensemble average $E[x(t)x(t-s)]$ for any s , which characterizes weak second-order (WSO) ergodicity. For this, it is necessary to use the correlation function of $x(t)x(t-s)$,

Manuscript received May 22, 1996; revised September 1, 1996.

The author is with the Laboratoire des Signaux et Systèmes, Supélec, Plateau de Moulon, 91190 Gif-sur-Yvette, France.

Publisher Item Identifier S 0018-9448(97)03451-2.

which is a fourth-order moment of $x(t)$. The most general form of such a moment is

$$m_4(\mathbf{t}) = m_4(t_1, t_2, t_3, t_4) = E[x(t_1)x(t_2)x(t_3)x(t_4)] \quad (3)$$

and we shall also extensively use its Fourier transform

$$M_4(\mathbf{f}) = M_4(f_1, f_2, f_3, f_4)$$

called fourth-order spectral moment. These functions are discussed in detail in [5, p. 236]. In particular, it is shown that if $x(t)$ is stationary, the spectral moment $M_4(\mathbf{f})$ is equal to zero except when $\mathbf{f}^T \mathbf{r} = 0$, where $\mathbf{r}$ is the vector $[1, 1, \dots, 1]^T$. This allows us to write $M_4(\mathbf{f}) = \Gamma_3(\mathbf{f})\delta(\mathbf{f}^T \mathbf{r})$, where $\Gamma_3(\mathbf{f})$ is called the *moment trispectrum* of $x(t)$. Therefore, $M_4(\mathbf{f})$ and $\Gamma_3(\mathbf{f})$ contain the same information and in the following discussion it will be more convenient to use $M_4(\mathbf{f})$ instead of $\Gamma_3(\mathbf{f})$. The equation

$$\mathbf{f}^T \mathbf{r} = f_1 + f_2 + f_3 + f_4 = 0$$

defines the *stationary manifold* in the frequency domain. For our discussion it is necessary to recall the definition of the *normal manifolds* (NM).

Suppose that $x(t)$ is a *normal* SP defined by its covariance function $\gamma(\tau)$ or its spectrum $\Gamma(f)$. The fourth-order moments of $x(t)$ can be expressed in terms of $\gamma(\tau)$ by a well-known formula (see [5, p. 278]). By Fourier transformation of this expression we obtain

$$M_4(\mathbf{f}) = \sum_N \Gamma(f_i)\Gamma(f_k)\delta(f_i + f_j)\delta(f_k + f_l). \quad (4)$$

The letter N (Normal) means that this sum contains three terms defining the three NM's, subsets of the stationary manifold, and defined by

$$\begin{aligned} f_1 + f_2 &= 0, & f_3 + f_4 &= 0 \\ f_1 + f_3 &= 0, & f_2 + f_4 &= 0 \\ f_1 + f_4 &= 0, & f_2 + f_3 &= 0. \end{aligned} \quad (5)$$

It is clear that (4) is unique up to an exchange of either f_i and f_j or f_k and f_l , due to the delta term and the symmetry of $\Gamma(f_i)$. Furthermore, (4) shows that on the NM's the density is not arbitrary but appears as a product of spectral densities $\Gamma(f_i)\Gamma(f_k)$. This form of density is called a *normal density*. This can be summarized by saying that the fourth-order spectral moment of a normal SP is uniquely distributed in the NM's with a normal density. For the following discussion it is worth pointing out that there are submanifolds of the NM's that can play a certain role. For example, to the first group of equations of (5) we can add the condition $f_1 + f_3 = 0$. This gives the manifold defined by $f_1 = -f_2 = -f_3 = f_4$.

II. CONDITIONS FOR SECOND-ORDER ERGODICITY

In order to apply the result indicated above, it is necessary to calculate the power spectrum of $y(t; s) = x(t)x(t-s)$. For this purpose, and to simplify the discussion, let us introduce some general assumptions. Suppose that $x(t)$ is an SP with zero mean value and a power spectrum denoted by $\Gamma(f)$. We assume that its fourth-order spectral moment can be written as

$$\begin{aligned} M_4(\mathbf{f}) &= \sum_N A(f_i, f_k)\delta(f_i + f_j)\delta(f_k + f_l) \\ &\quad + B(f_1, f_2, f_3)\delta(f_1 + f_2 + f_3 + f_4). \end{aligned} \quad (6)$$

The first three terms of this expression correspond to a distribution on the NM's with density $A(f_i, f_k)$ and the last term is a distribution

on the stationary manifold with density $B(\mathbf{f})$. In the normal case we have $B(\mathbf{f}) = 0$ and $A(f_i, f_k) = \Gamma(f_i)\Gamma(f_k)$. It is clear that $B(\mathbf{f})$ is the contribution to the fourth-order spectral moment that does not possess the normal manifolds as support. The relation (6) is verified by all the models of SP's analyzed below. Furthermore, there is no example of SP discussed in the literature for which the terms $A(f_i, f_k) = 0$, and, therefore, the structure (6) seems quite general.

Weak second-order ergodicity is related to the structure of the power spectrum of $y(t; s) = x(t)x(t-s)$. Its covariance function is

$$\begin{aligned}\gamma_y(\tau; s) &= E[y(t; s)y(t-\tau; s)] - \gamma^2(s) \\ &= m_4(t, t-s, t-\tau, t-s-\tau) - \gamma^2(s)\end{aligned}\quad (7)$$

where $\gamma(\cdot)$ is the covariance function of $x(t)$, and m_4 its fourth-order moment. The power spectrum of $y(t; s)$ is the Fourier transform of $\gamma_y(t; s)$, or

$$\Gamma_y(f; s) = \int \gamma_y(\tau; s) \exp(-2\pi j f \tau) d\tau. \quad (8)$$

Expressing m_4 of (7) in terms of its Fourier transform $M_4(\mathbf{f})$ yields

$$\begin{aligned}m_4(t, t-s, t-\tau, t-s-\tau) \\ = \iiint M_4(\mathbf{f}) \exp(2\pi j \mathbf{f}^T \mathbf{r} t) \\ \cdot \exp\{-2\pi j[(f_3 + f_4)\tau + (f_2 + f_4)s]\} d\mathbf{f}_4\end{aligned}\quad (9)$$

where $d\mathbf{f}_4$ means $df_1 df_2 df_3 df_4$. The stationary manifold $\mathbf{f}^T \mathbf{r} = 0$ appears immediately in this expression. Indeed, because of the assumption of stationarity, the function m_4 in (9) does not depend on t . This is ensured by the relation $\mathbf{f}^T \mathbf{r} = 0$, and therefore the first exponential term of the integral is equal to one. By inserting this relation into (8), performing the integration with respect to τ , which introduces the term $\delta(f + f_3 + f_4)$, and afterwards the integration in f_4 , we obtain

$$\begin{aligned}\Gamma_y(f; s) &= \iiint M_4(f_1, f_2, f_3, -f-f_3) \\ &\cdot \exp[-2\pi j(f_2 - f - f_3)s] df_1 df_2 df_3 - \gamma^2(s)\delta(f).\end{aligned}\quad (10)$$

Let us now calculate the contribution to $\Gamma_y(f; s)$ of the various terms appearing in $M_4(\mathbf{f})$. The last term of (6) introduces the term $\delta(\mathbf{f}^T \mathbf{r})$. Used in (10) this gives $\delta(f_1 + f_2 - f)$. As a consequence, the contribution of the density $B(\mathbf{f})$ on the stationary manifold introduces the term

$$T_B = \iint B(f_1, f - f_1, f_3) \exp[2\pi j(f_1 + f_3)s] df_1 df_3. \quad (11)$$

Let us now calculate the contribution T_1 of the first NM appearing in (6). This introduces the term $A(f_1, f_3)\delta(f_1 + f_2)\delta(f_3 - f - f_3)$ in the integral (10). As a result, this term is

$$T_1 = \delta(f) \iint A(f_1, f_3) \exp[2\pi j(f_1 + f_3)s] df_1 df_3. \quad (12)$$

By following the same procedure we find that the two terms due to the last two manifolds of (5) are equal to

$$\begin{aligned}T_2 &= \int A(f_1, f - f_1) df_1 \\ T_3 &= \int A(f_1, f - f_1) \exp[2\pi j(2f_1 - f)s] df_1.\end{aligned}\quad (13)$$

Finally, we obtain

$$\Gamma_y(f; s) = \delta(f) \left[\iint A(f_1, f_3) \exp[2\pi j(f_1 + f_3)s] df_1 df_3 - \gamma^2(s) \right]$$

$$\begin{aligned}&+ \int A(f_1, f - f_1) \{1 + \exp[2\pi j(2f_1 - f)s]\} df_1 \\ &+ \iint B(f_1, f - f_1, f_3) \\ &\cdot \exp[2\pi j(f_1 + f_3)s] df_1 df_3.\end{aligned}\quad (14)$$

In order to test WSO ergodicity by applying the result indicated in the Introduction, we have to verify whether or not there is a spectral line at the frequency zero. If the coefficient of $\delta(f)$ in (14) is equal to zero and if the other terms are bounded, there is no spectral line, and $x(t)$ is WSO ergodic. On the other hand, lack of ergodicity can come from several origins. This can appear if the coefficient of $\delta(f)$ in (14) is not equal to zero and if the other terms are bounded. In these cases, it appears that the spectral line, and therefore the ergodicity property, is entirely due to the structure of the terms $A(f_1, f_3)$ on the NM's. However, there are more complicated situations, the best example being the case where $x(t)$ is normal, as analyzed in the next section.

The previous analysis was only devoted to second-order ergodicity and the corresponding calculations require the use of spectral moments up to the fourth order. It is clear that similar calculations can be extended to ergodicity of an order higher than two. The principles are the same but the detailed expressions are more complicated. It is obvious that testing weak ergodicity of order n requires the use of spectral moments of order $2n$. However, it appears that the role of the NM's remains the same.

III. EXAMPLES

A. Normal Case

Suppose that $x(t)$ is a normal SP. This implies that the last term of (14) is zero. Suppose first that there is no spectral line in the spectrum of $x(t)$. This means that the second term of (2) is zero. In this case, the coefficient of $\delta(f)$ in (14) is zero and the second term is bounded. As a consequence, $x(t)$ is WFO and WSO ergodic. Suppose now that there is one spectral line in the spectrum of $x(t)$ at a nonzero frequency. As $x(t)$ is real, its power spectrum is symmetric, and our assumption allows us to write (2) in the form

$$\Gamma(f) = \Gamma_b(f) + \gamma[\delta(f - f_i) + \delta(f + f_i)].$$

It is clear that these spectral lines do not change the fact that $x(t)$ is WFO ergodic. Furthermore, the coefficient of $\delta(f)$ in (14) is still equal to zero. This does not mean that there is no spectral line at the frequency zero. Indeed, by coupling the terms $\delta(f - f_i)$ and $\delta(f + f_i)$, the second integral of (14) yields the term $2\gamma^2[1 + \cos(4\pi f_i s)]\delta(\nu)$, which introduces a spectral line at the frequency zero. Therefore, even if $x(t)$ is WFO ergodic, it is not WSO ergodic. So we see that the spectral line at the frequency zero comes not from the first term of (14) but from its second term, because $A(f_1, f_3)$ is not bounded. In conclusion, ergodicity of normal SP's requires that there is no spectral line at all in the spectrum of $x(t)$. This is in accordance with a known result, shown by a completely different method (see [6, p. 157]): a normal process is ergodic if and only if its spectral distribution is continuous everywhere. As the power spectrum is the derivative of the spectral distribution, this is equivalent to saying that there is no spectral line in the spectrum.

B. Spherically Invariant Processes

A process $x(t)$ is said to be spherically invariant if it can be written as $x(t) = Au(t)$, where $u(t)$ is a normal SP and A a random variable independent of $u(t)$. For simplicity we assume that the mean values of A and $u(t)$ are zero. Furthermore, we assume that the variance of A^2 is not zero, which means that A is not a random variable taking only two values $+a$ and $-a$. The results from these assumptions are

that $x(t)$ cannot be normal and, therefore, its fourth-order spectral moment has no reason to take the form (4). We shall, however, see that this spectral moment satisfies (6), with specific values of A and B . For this, note that the power spectrum of $x(t)$ is $m_2\Gamma_u(f)$, where m_2 is the variance of A and $\Gamma_u(f)$ the power spectrum of $u(t)$. We assume that this spectrum does not contain a spectral line. As $u(t)$ is normal, its fourth-order spectral moment satisfies (4) where $\Gamma(\cdot)$ is replaced by $\Gamma_u(\cdot)$. As a result, the fourth-order spectral moment of $x(t)$ is given by (6) where the term $B(f_1, f_2, f_3)$ is zero and

$$A(f_i, f_k) = m_4\Gamma(f_i)\Gamma(f_k).$$

This implies that the last two terms of (14) are bounded. The first term can be expressed as $\delta(f)(m_4 - m_2^2)\gamma_u^2(s)$, where $m_4 = E(A^4)$ and $\gamma_u(\cdot)$ is the correlation function of $u(t)$. The conclusion is straightforward: the SP $x(t)$ is WSO ergodic if and only if $m_4 - m_2^2 = 0$. As this expression is the variance of A^2 , this is equivalent to saying that A^2 is not random. This implies either that A is not random or that A is a random variable taking only two values $+a$ and $-a$, so that $A^2 = a^2$. If A is not random, $Au(t)$ is normal. It is the same if $A = \pm a$, because $u(t)$ and $-u(t)$ have the same probability distribution. As a conclusion, the only spherically invariant SP that are WSO are the normal processes. This property has also been obtained by completely different methods in [7]. Note finally that $x(t)$ is an example of SP for which $M_4(\mathbf{f})$ is uniquely distributed on the normal manifolds, which means that the term $B(\mathbf{f})$ is zero, and the nonnormal character is due to the fact that the density in the NM's is not normal.

C. Shot Noise of Poisson or Poisson Compound Processes

The shot noise of a point process is defined by $s(t) = \sum h(t - t_i)$ where the time instants $\{t_i\}$ are the random points of the process (see [5, p. 187]). After subtracting the mean value we obtain the zero mean SP $x(t) = s(t) - E[s(t)]$ that can be written as

$$x(t) = \int h(t - \theta) dM(\theta)$$

where $E[dM(\theta)] = 0$ (see [5, p. 326]). When the time instants $\{t_i\}$ are the random points of a Poisson process of density λ , $M(t)$ is an SP with independent increments satisfying $E[dM^k(\theta)] = \lambda d\theta$. As a result, the fourth-order spectral moments of these increments are distributed on the stationary manifold with the density λ and on the normal manifolds with the density λ^2 . As a consequence the spectral properties of $x(t)$ are characterized by

$$\begin{aligned} \Gamma_x(f) &= \lambda|G(f)|^2, \\ A(f_i, f_k) &= \lambda^2|G(f_i)|^2|G(f_k)|^2 \\ B(f_1, f_2, f_3) &= \lambda G(f_1)G(f_3)G(f_3)G(-f_1 - f_2 - f_3) \end{aligned} \quad (15)$$

where $G(f)$ is the frequency response of the filter, or the Fourier transform of its impulse function $h(t)$. We see that the first term of (14) is equal to zero. Therefore, if $|G(f)|$ is bounded, which is a very general assumption for the frequency response of a filter, $x(t)$ is WSO ergodic. This is no longer true for a compound Poisson process in which the density λ becomes a positive random variable Λ . In fact, we must replace in the above equations λ and λ^2 by $E(\Lambda)$ and $E(\Lambda^2)$, respectively. In order to suppress the spectral line at the frequency zero we must have $E(\Lambda^2) = E^2(\Lambda)$, which implies that Λ is no longer random. This result is quite similar to that obtained for spherically invariant SP's, sometimes called compound normal processes. The difference between the two processes appears with the term B of (6) which is equal to zero for spherically invariant SP's. However, this term does not play any role in WSO ergodicity which is associated with the form of the density on the NM's. These results have a simple physical interpretation. An SP containing in its

definition a random variable, which is essentially time-independent, is in general not ergodic because time averages cannot be equal to ensemble averages.

D. Processes with Independent Spectral Components

Let $x(t)$ be a harmonizable SP with the spectral decomposition

$$x(t) = \int dX(f) \exp(2\pi jft). \quad (17)$$

As $x(t)$ is real we have $dX(f) = dX^*(-f)$ and we can use only positive frequencies. We assume that $E[dX(f)] = 0$, which implies that the mean value of $x(t)$ is zero. As it is stationary, the second-order properties of the spectral increments are characterized by

$$E[dX(f_1) dX(f_2)] = \delta(f_1 + f_2)\Gamma(f_1) df_1 df_2 \quad (18)$$

where $\Gamma(f)$ is the power spectrum. This means that the increments $dX(f_1)$ and $dX(f_2)$ are uncorrelated if $f_1 \neq -f_2$. We assume now that they are *independent*. In order to simplify the discussion let also assume that there is no spectral line in the spectrum. This means that the last term of (2) is zero. The fourth-order moment

$$E[dX(f_1) dX(f_2) dX(f_3) dX(f_4)] = M_4(\mathbf{f}) d\mathbf{f}_4$$

is zero outside the stationary manifold. We shall first show that it is only distributed on the NM's, which means that the term $B(\cdot)$ in (6) is zero. For this, note that vectors $\mathbf{f}$ belonging to the stationary manifold $f_1 + f_2 + f_3 + f_4 = 0$ and not to the NM's must satisfy simultaneously

$$f_1 + f_2 \neq 0, f_1 + f_3 \neq 0, f_1 + f_4 \neq 0$$

because if one of these equations is violated, $\mathbf{f}$ belong to one of the NM's. Indeed, suppose, for example, that $f_1 + f_2 = 0$. As we have $f_1 + f_2 + f_3 + f_4 = 0$, this gives $f_3 + f_4 = 0$, which characterizes the first NM defined by (5). This implies that outside the NM's it is impossible to have $f_i = f_1 + f_2 + f_3, 1 \leq i \leq 3$. However, on the stationary manifold the spectral moment takes the form

$$E[dX(f_1) dX(f_2) dX(f_3) dX^*(f_1 + f_2 + f_3)]$$

and the previous result shows that $dX^*(f_1 + f_2 + f_3)$ cannot be coupled with another increment. As these increments are zero-mean and independent, the fourth-order spectral moment is zero outside the NM's.

The remaining task is to calculate the term $A(f_i, f_k)$ of (6). For this suppose that $E[|dX(f)|^4] = C(f) df$. On the first NM of (5) this term appears when $f_1 = f_3$ or $f_1 = -f_3$. The same result appears for the other NM's. So by using (18) we deduce that

$$A(f_i, f_k) = \Gamma(f_i)\Gamma(f_k) + C(f_i)[\delta(f_i + f_k) + \delta(f_i - f_k)]. \quad (19)$$

Inserting this expression in the first term of (14) shows that ergodicity of $x(t)$ is entirely due to the term $C(f)$ that must now be analyzed.

For this let us recall a classical result: the only real SP's with independent increments that are continuous with probability one are the Wiener or the Poisson processes. This can be extended to complex SP's like the spectral function $X(f)$ appearing in (17). For this it suffices to use the real and imaginary parts $M(f)$ and $N(f)$ of $X(f)$. It is known that if $x(t)$ is normal, $M(f)$ and $N(f)$ are two independent Wiener processes with the same covariance function. As a result, $dX(f)$ is a normal circular random variable with variance $\Gamma(f) df$. This implies that $E[|dX(f)|^4] = 2[\Gamma(f) df]^2$ and this term can be neglected due to its dependence on $[df]^2$. Therefore, $C(f) = 0$, and $x(t)$ is WSO ergodic. This is in accordance with the result obtained in Section III-A.

Suppose now that $M(f)$ is a Poisson process with zero mean value obtained by the same procedure as in Section III-C. We have seen in this section that $E[dM(f)^4] = 2\lambda(f) df$. This implies that $C(f)$ is not zero and, therefore, cannot be WSO ergodic.

This can be summarized by saying that the only SP's with independent spectral components that are also WSO ergodic are normal processes without spectral lines. This result was obtained by an entirely different procedure in [8].

REFERENCES

- [1] Y. G. Sinai, "On higher order spectral measures of ergodic stationary processes," *Theory of Probab. and its Appl.*, vol. 8, pp. 429–436, 1963.
- [2] A. N. Shiryaev, "On conditions for ergodicity of stationary processes in terms of higher order moments," *Theory Probab. Appl.*, vol. 8, pp. 436–439, 1963.
- [3] A. Ferrari and G. Alengrin, "High-order ergodicity of a complex harmonic process," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1228–1229, 1994.
- [4] H. Parthasarathy, S. Prasad and S. D. Joshi, "Conditions for third-order stationarity and ergodicity of a harmonic random process," *IEEE Trans. Signal Processing*, vol. 42, pp. 222–225, 1994.
- [5] B. Picinbono, *Random Signals and Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [6] H. Cramer and M. Leadbetter, *Stationary and Related Stochastic Processes*. New York: Wiley, 1967.
- [7] A. M. Vershik, "Some characteristic properties of Gaussian stochastic processes," *Theory of Probab. and its Appl.*, vol. 9, pp. 353–356, 1964.
- [8] G. Maruyama, "Infinitely divisible processes," *Theory of Probab. and its Appl.*, vol. 15, pp. 1–22, 1970.

A General Minimax Result for Relative Entropy

David Haussler, *Member, IEEE*

Abstract—Suppose Nature picks a probability measure P_θ on a complete separable metric space X at random from a measurable set $\mathcal{P}_\Theta = \{P_\theta: \theta \in \Theta\}$. Then, without knowing θ , a statistician picks a measure Q on X . Finally, the statistician suffers a loss $D(P_\theta||Q)$, the relative entropy between P_θ and Q . We show that the minimax and maximin values of this game are always equal, and there is always a minimax strategy in the closure of the set of all Bayes strategies. This generalizes previous results of Gallager, and Davisson and Leon-Garcia.

Index Terms—Bayes risk, channel capacity, computational learning theory, density estimation, minimax redundancy, minimax risk, minimax theorem, Kullback–Leibler divergence, relative entropy, source coding.

I. INTRODUCTION

Consider a sequential estimation game in which a statistician is given n independent observations $Y_1, \dots, Y_n$ distributed according to an unknown distribution $\tilde{P}_\theta$ chosen at random by Nature from the set $\{\tilde{P}_\theta: \theta \in \Theta\}$ according to a known prior distribution μ on Θ . For each time t between 1 and n , the statistician must produce an estimate $\hat{P}_t$ for the unknown distribution $\tilde{P}_\theta$, based on the previous $t - 1$ observations $Y_1, \dots, Y_{t-1}$. At the end of this time period, the

statistician suffers a *cumulative relative entropy loss* $\sum_{t=1}^n D(\tilde{P}_\theta||\hat{P}_t)$, which measures the quality of the sequential estimates made. Variants of this game have been studied by several authors (see, e.g., [3], [4], [9], [14], [15]). If the observations are restricted to a finite alphabet, then the actions of the statistician can be interpreted as adaptive source coding for an unknown source. For fixed θ , the average loss suffered is the redundancy [6, 3]. When we also average over the random choice of θ according to the prior μ , we get the *risk* for this game, which is then average redundancy. This risk is called *cumulative relative entropy risk* in statistics.

The product of the sequence of conditional distributions $\{\hat{P}_t(Y_t|Y_1, \dots, Y_{t-1}): 1 \leq t \leq n\}$ forms a joint distribution Q on the n -fold product of the space of observations. Let us call this n -fold product space X . In this way, the actions of the statistician can be interpreted as choosing a joint distribution Q on the space X . Let $P_\theta = \tilde{P}_\theta^n$ be the "true" joint distribution for the observations. Using the chain rule for relative entropy, it is seen that the risk for this game reduces to $\int D(P_\theta||Q) d\mu(\theta)$ [3]. This leads us to a simpler, more general game: Nature picks a prior μ on Θ and then picks a probability measure P_θ on a space X at random (according to μ) from a set $\{P_\theta: \theta \in \Theta\}$. Then, knowing μ but not knowing θ , a statistician picks a measure Q on X . Finally, the statistician suffers a loss $D(P_\theta||Q)$.

We show that the minimax and maximin values of this game are always equal, and there is always a minimax strategy in the closure of the set of all Bayes strategies. This generalizes results of Gallager [11], and Davisson and Leon-Garcia [7], which were restricted to the case when the observations are chosen from a finite set of symbols. The proof of the general result closely follows that in [10, Theorem 2, p. 85], which is based on earlier results of Le Cam [2], with one fairly simple extension to handle the case when $\{P_\theta: \theta \in \Theta\}$ is not uniformly tight (Lemma 4 below).

In the source coding interpretation of this game, the minimax value is the capacity of the channel from Θ to X [4], [7]. A similar interpretation applies in computational learning theory, where the cumulative relative entropy risk is interpreted as the average additional loss suffered by an adaptive algorithm that predicts each observation before it arrives, based on the previous observations, as compared to an algorithm that makes predictions knowing the true distribution [12], [13]. Here, to get this interpretation, we assume that the observation at time t is predicted by the "predictive" probability distribution $\hat{P}_t$, formed by the adaptive algorithm using the previous $t - 1$ observations, and that when this t th observation arrives, the loss is the negative logarithm of its probability under $\hat{P}_t$. The game has interpretations in other fields as well. For example, in mathematical finance and gambling theory, the cumulative relative entropy risk measures the expected reduction in the logarithm of compounded wealth due to lack of knowledge of the true distribution [1], [3].

II. PRELIMINARY DEFINITIONS

We first briefly review some basic facts about probability measures on complete separable metric spaces; proofs of these can be found in, e.g., [8]. Let (X, ρ) be a complete separable metric space and let $\mathcal{A}(X)$ be the set of all probability measures defined on the σ -field generated by the open sets of X (i.e., the Borel subsets of X). For any real-valued function f on X

$$\|f\|_\infty = \sup_x |f(x)|$$

$$\|f\|_L = \sup_{x \neq y} |f(x) - f(y)|/\rho(x, y)$$

Manuscript received June 7, 1995; revised December 12, 1996. This work was supported by NSF under Grant IRI-9123692.

The author is with Computer and Information Sciences, University of California at Santa Cruz, Santa Cruz, CA 95064 USA
Publisher Item Identifier S 0018-9448(97)03452-4.