Biol. Cybern. 54, 289–300 (1986)

Biological Cybernetics
© Springer-Verlag 1986

Maximum-Entropy Approximations of Stochastic Nonlinear Transductions: An Extension of the Wiener Theory

J. D. Victor[1] and P. Johannesma[2]

[1] The Rockefeller University and Cornell University Medical College, New York City, NY 10021, USA
[2] Department of Medical Physics and Biophysics, University of Nijmegen, Nijmegen, The Netherlands

Abstract. We consider the description of a nonlinear stochastic transduction in terms of its input/output distribution. We construct a sequence of approximating maximum-entropy estimates from a finite set of input/output observations. This procedure extends the Wiener theory to the analysis of nonlinear stochastic transducers and to the analysis of transducers with multiple outputs but an inaccessible input.

Introduction

The Wiener orthogonal functional series (Wiener 1958) provides a canonical way to characterize a nonlinear operator. The application of the Wiener procedure to biological transducers has enabled a detailed, quantitative analysis of some nonlinear neural transductions (Marmarelis and Naka 1972; Naka et al. 1975; Sakuranaga and Naka 1985a–c). The success of the orthogonal series approach has motivated two kinds of generalizations: generalizations of the test (input) signal to signals other than Gaussian white noise (Krausz 1975; Marmarelis 1977; Victor and Knight 1979; Yasui 1979) and generalizations to transducers with multiple inputs (Yasui et al. 1979).

Here we are concerned with two other avenues of generalization: applications in which the variability of the system's response, as well as its average or most likely response, is of interest, and applications in which there is no clearly-defined or accessible input, but only multiple outputs that may be observed simultaneously. Both of these extensions are important for the analysis of interactions in the central nervous system, in which variability is a prominent feature and the physiological input may be inaccessible or only partially under the experimenter's control. We approach these problems by reformulating the Wiener series in a way which lends itself to these extensions. The reformulation rests on a maximum-entropy interpretation of the Wiener series.

The investigator studying an unknown transducer aims to use a set of input-output observations to form a description of the transducer. An exhaustive list of all input-output pairs, along with their relative probabilities, is a complete (although perhaps awkward) description of the input-output behavior of the system. It is highly impractical if not impossible to determine this distribution experimentally. However, one might hope to develop a canonical sequence of approximations to this distribution, in which the first few members of the sequence can be determined experimentally. Measurement of the first few members of the sequence may serve to distinguish among models of the transducer under study. Even if the later members of the sequence are not practical to measure, the existence of the sequence might prove useful in theoretical investigations. For a deterministic transducer, this is the role played by the Wiener series and related orthogonal functional series.

In the Wiener approach, each member of the approximating sequence is a transducer of a given finite order of nonlinearity whose output most closely approximates the output of the transducer under study. The criterion of approximation is the mean-squared difference between the output of the approximating transducer and the transducer under study, when the test signal is a Gaussian white noise. Each approximating transducer is deterministic, and the transducer under study is assumed to be deterministic.

In the present approach, we construct a sequence of stochastic transducers. The input/output distributions of the approximating transducers are maximum-entropy distributions, subject to the constraint that particular functionals of the input/output distribution match exactly those of the transducer under study. Later members of the approximating sequence are constrained by more functionals than earlier members, and thus give a more accurate picture of the input/output distribution of the transducer under study.

The sequence of constraining functionals determines the nature of the approximating sequence of distributions. (For an inappropriate choice of functionals, the approximating sequence may not even be defined.) One natural choice of functionals is the set of moments and cross-correlations of the input and output. We will see that for a particular choice of moments and cross-correlations, this approach reduces to the Wiener approach if the transducer under study is deterministic.

Real transducers have noise. An application of the Wiener procedure to a stochastic transducer amounts to assuming that the noise is additive and uncorrelated with the input. The Wiener series converges to a deterministic transducer whose output to a given input is the average output of the stochastic transducer to that input. The difference between the mean-squared output of the stochastic transducer and the mean-squared output of the complete Wiener model is a measure of the variability of the response. More generally, cross-correlations that are of second- and higher-order in the output provide additional information about the stochastic features of the transducer which is not present in the Wiener kernels.

We first discuss a rationale for a maximum-entropy approach. Next, we construct constrained maximum-entropy multivariate distributions by the method of Lagrange multipliers. We provide conditions sufficient for the existence of a sequence of maximum-entropy distributions and discuss the sense in which this sequence converges. This analysis provides guidelines for using this approach for the analysis of stochastic nonlinear transducers. Then, we will show how the Wiener theory may be viewed as an instance of this method for deterministic transducers with a particular choice of constraints. Finally, we present two examples: one showing how one may distinguish between additive and more complex sources of variability in the output of a transducer, and a second example illustrating the analysis of a simple nonlinear stochastic system with two observable outputs but an inaccessible input.

Maximum-Entropy Distributions

We consider a stationary nonlinear transducer T. We assume that T is causal and of finite memory. Time is discretized into intervals Δt. An input to the transducer is denoted by a vector $\mathbf{x}$; and output of the transducer is denoted by a vector $\mathbf{y}$. The components x_i of $\mathbf{x}$ and y_i of $\mathbf{y}$ denote, respectively, the size of the input and output at the time $i\Delta t$.

In an experiment, one makes a series of observations of the output signal $\mathbf{y}$ while manipulating the input signal $\mathbf{x}$. The hypotheses of stationarity and finite memory imply that this experiment is equivalent to making a sequence of observations of the output y_0 at time 0 to independent examples of the input signal. This view will be most convenient for our analysis.

The hypothesis of causality implies that the output $y_0$at time 0 depends only on the input signal at previous times, x_j for $j \leqq 0$. Therefore, each input-output observation may be thought of as a vector

$$\mathbf{z} = \langle x_0, x_{-1}, x_{-2}, \ldots, y_0, y_{-1}, y_{-2}, \ldots \rangle \tag{1}$$

in an input-output space. For a deterministic transducer T, the $\mathbf{x}$-components of $\mathbf{z}$ determine the $\mathbf{y}$-components uniquely. For a stochastic transducer, the $\mathbf{x}$-components determine a probability distribution for the $\mathbf{y}$-components. In either case, the characterization of the transducer is equivalent to characterizing the joint distribution of the $\mathbf{x}$-components and the $\mathbf{y}$-components. We denote the measure corresponding to this distribution by du.

We have posed the problem in a context which is inspired by the Wiener identification procedure: the input is assumed to be controllable by the experimenter, and the output is assumed to be related in a causal fashion to the input. The present formulation in terms of a joint distribution of the two signals $\mathbf{x}$ and $\mathbf{y}$ remains well-defined when the input and output can only be observed. In this circumstance, there is no formal distinction between input and output; we may think of the transducer under study as a system with two ports, $\mathbf{x}$ and $\mathbf{y}$. In such a system, causality is no longer assured unless one has prior knowledge of the internal structure of the system (since there is no a priori notion of input). The only role that causality plays in the approximation theory to be developed below is to stipulate that certain cross-correlations vanish. Thus, the only modification that is needed to treat the no-input, multi-output problem is to allow these cross-correlations to be experimental observations. The basic approximation procedure remains unchanged.

Why Maximum Entropy? Maximum entropy formalisms have been invoked in a variety of settings to select a specific probability distribution out of an ensemble of possible distributions (Jaynes 1979, 1982). We present an intuitive justification for this approach in the setting in which the distribution sought is the input-output relationship of a stochastic transducer.

A stochastic transducer is completely described by knowledge of the probabilities of every possible input/output vector $\mathbf{z}$. Although the exact distribution is required to specify the transducer exhaustively, usually one is more interested in particular functionals of the distribution, such as the mean, the variance, higher moments, and cross-correlations.

It is impossible to calculate or even estimate the distribution of input/output vectors $\mathbf{z}$ from a finite set of observations without making explicit assumptions. In part the appropriate assumptions are motivated by characteristics of the particular transduction under study and the ultimate goal of the investigation. On the other hand, the use of a maximum entropy estimate has the advantage that it formalizes an assumption of ignorance about aspects of the distribution other than the functionals explicitly used as constraints.

One naive approach to estimating the input/output distribution is to assume that the observed values of $\mathbf{z}$ are an unbiased and universal sampling of the true distribution. An estimate of any functional of the true distribution can be obtained by applying the functional to the finite sampling. This is equivalent to the assumption that independent additional samples of the true distribution may be obtained by resampling the finite sample. Although this estimate for the true distribution is unbiased, it is not an intuitively reasonable estimate: it will be very ragged because of the sampling error inherent in a finite set of observed values.

Typically, one does not place much reliance on the particular values of single observations but rather on the values of functionals of the distribution as a whole (such as mean, variance, etc.). Let us select certain such functionals of the empirical distribution as significant, and measure these functionals on the empirical (finitely-sampled) distribution. We use this finite set of descriptors to generate an estimate of the true distribution.

In general, there will be many distributions for which the values of the functionals will agree with the values of the functionals on the empirical distribution. Although the values of the functionals of interest are identical on all of these candidate distributions, the values of functionals that have not been measured (such as higher moments) will vary.

Consider a large but finite sampling of N observations derived from one of these candidate distributions p. We discretize the samplings into narrow bins $\mathbf{z}_i$ of width Δz. Thus, the expected number of observations in the bin $\mathbf{z}_i$ is $Np(\mathbf{z}_i)\Delta z$. The number of ways that the N observations can be distributed into the bins in this fashion is given by a multinomial expression whose numerator is $N!$ and whose denominator is the product of the factorials of the number of observations in each bin. This is a kind of partition function, of the sort encountered in thermodynamics. As N grows without bound, the logarithm of the partition function asymptotically grows proportionally with N. This proportionality constant is akin to a thermodynamic entropy, and we may think of it as the entropy of the distribution, $E[p]$. Using Stirling's approximation, the entropy is given by

$$E[p] = -\int p(\mathbf{z}) \ln p(\mathbf{z}) d\mathbf{z}. \tag{2}$$

We seek to maximize the entropy subject to the condition that the values of the functionals of interest on the distribution p agree with the values of the functionals on the observed distribution. Distributions that are concentrated in only a few bins have lower entropy than a distribution that is dispersed evenly among many bins. Maximizing the entropy is thus a kind of smoothing of the distribution. Thus, the proposed procedure embodies the assumption that distributions in which the populations of the bins are uneven are a priori less likely than those in which the populations of the bins are more uniform.

This maximization procedure also has a thermodynamic interpretation. One can think of each observation of a distribution as a particle, and one can think of the bins as corresponding to states of the particles. We assume that the energy associated with each state is equal, or, equivalently, that the temperature of the system is infinitely high so that only combinatorial factors will influence state occupancy. The initial finite set of observations, when replicated a large number of times, corresponds to a system in which all of the particles are concentrated in a few states. Imagine that the particles interact according to a dynamical law which allows many-body interactions and requires only that the functionals are conserved quantities. The initial distribution will typically be unstable under this dynamical law; the maximum-entropy distribution, if it exists, will be stable under this dynamical law. In other words, the maximum-entropy distribution is the most evenly-populated distribution consistent with the experimentally-measured functionals.

It might appear that the maximum-entropy constraint could be expressed simply in terms of cumulants. If the constraints consist of means, variances, and simple (first-order) cross-correlations, the maximum-entropy distribution is the unique multivariate Gaussian distribution with matching statistics. The higher cumulants of the Gaussian are zero. However, for constraints that include statistics of higher order, the maximum-entropy constraint is not equivalent to assuming that higher cumulants are zero.

The Constrained Maximum-Entropy Distribution. We will use the method of Lagrange multipliers to maximize the entropy $E[p]$ (2) subject to a set of functionals. Let us denote the constraining functionals by B_k $(k=0, 1, \ldots, K)$. The functionals are assumed to be linear in the sense that there is a weighting function $B_k(\mathbf{z})$ such that the value $B_k[p]$ of B_k applied to a distribution p is given by

$$B_k[p] = \int p(\mathbf{z}) B_k(\mathbf{z}) d\mathbf{z}. \tag{3}$$

Moments and cross-correlations of all orders are linear functionals in this sense; they correspond to a weighting function $B_k(\mathbf{z})$ equal to a product of appropriate components of $\mathbf{z}$. We will denote by b_k the value of the constraint B_k applied to the empirical data set. For convenience, we will take $B_0=1$ and $b_0=1$, so that the zeroth constraint is that the distribution p is normalized to unity.

Let $p_{\max}(\mathbf{z})$ denote a maximum-entropy distribution subject to the constraints

$$B_k[p_{\max}]=b_k, \quad k=1,\ldots,K. \tag{4}$$

Distributions in the neighborhood of $p_{\max}(\mathbf{z})$ can be expressed in the form

$$p_{\text{near}}(s,q,\mathbf{z})=p_{\max}(\mathbf{z})+sq(\mathbf{z}), \tag{5}$$

where s is a small real number. We think of $p_{\text{near}}(s,q,\mathbf{z})$ as a path in the space of possible distributions that passes through $p_{\max}(\mathbf{z})$ at $s=0$. The "direction" of this path at $s=0$ is determined by $q(\mathbf{z})$.

According to the method of Lagrange multipliers as applied to function spaces (Crowder and McCuskey 1964, pp. 499–503) a condition that $p_{\max}(\mathbf{z})$ is an extreme point of the entropy is that at $s=0$,

$$\frac{d}{ds}\left\{E[p_{\text{near}}(s,q,\mathbf{z})]+\sum_{k=0}^{K}L_kB_k[p_{\text{near}}(s,q,\mathbf{z})]\right\}=0. \tag{6}$$

By substituting (2), (3), and (5) into (6), the extremization condition becomes

$$\int\left[-1-\ln p_{\max}(\mathbf{z})+\sum_{k=0}^{K}L_kB_k(\mathbf{z})\right]q(\mathbf{z})d\mathbf{z}=0. \tag{7}$$

Since the extremization condition (7) must hold for any small perturbation $q(z)$, it follows that the integrand itself must be identically zero. Therefore, a necessary condition that entropy is extremized is

$$p_{\max}(\mathbf{z})=\exp\left[-1+\sum_{k=0}^{K}L_kB_k(\mathbf{z})\right]. \tag{8}$$

The Lagrange multipliers L_k must be determined from (4) and (8). In general, this results in a system of $K+1$ transcendental equations. However, in the special case that the constraints are only first- and second-order moments and first-order cross-correlations, a simplification occurs. According to (8), the distribution $p_{\max}(\mathbf{z})$ is the exponential of a quadratic form in $\mathbf{z}$ – a correlated multivariate Gaussian distribution. In this case, the solution of the Eqs. (4) reduces to the diagonalization of a quadratic form.

Equation (8) only guarantees that the distribution $p_{\max}(\mathbf{z})$ lies at a critical point, but does not guarantee that this critical point is a maximum point. We next show that any distribution that satisfies (4) and (8) must be a *maximum* point. We evaluate the second derivative of the entropy along any path $p_{\text{near}}(s,q,\mathbf{z})$ through a distribution $p(\mathbf{z})$:

$$\frac{d^2}{ds^2}E[p_{\text{near}}(s,q,\mathbf{z})]=-\int q^2(\mathbf{z})(p(\mathbf{z}))^{-1}d\mathbf{z}. \tag{9}$$

The right-hand side of this expression is always less than zero. In particular, the entropy of distributions on any path through a critical distribution $p_{\max}(\mathbf{z})$ must have a maximum at $p_{\max}(\mathbf{z})$.

As a consequence, there can be *at most* one solution for the system of (4) and (8). If two solutions were to exist, then there would be two local maxima of the entropy. To see that this is impossible, consider the behavior of the entropy along a straight path between two putative solutions. Because the distributions at the endpoints of the paths satisfy the constraints (3) and because these constraints are linear, any distribution along this path also satisfies the constraints. If the entropy had a maximum at the two endpoints, then there must be a minimum at some intermediate point along the path. At such a point, the second derivative of the entropy would have to be positive. This contradicts (9), and excludes the possibility of more than one maximum. Thus, there is at most one solution to (4) and (8), and this solution must correspond to a global maximum.

An Existence Result. We have shown that there is at most one maximum-entropy distribution corresponding to a particular set of constraints B_k. The possibility remains that the system of (8) and (4) have no solution. If we make two additional assumptions, we can demonstrate simply that a solution to these equations always exists.

Condition I: The number of components of $\mathbf{z}$ is finite, and the domain of values for each component z_i has been discretized. This technical assumption means that there are only a finite number of degrees of freedom for the probability distributions $p(\mathbf{z})$.

Condition II: There exists a distribution p_{int} that satisfies the constraints $B_k[p_{\text{int}}]=b_k$ and for which $p_{\text{int}}(\mathbf{z})>0$. This assumption will hold, for example, if the constraining values b_k are obtained from an empirical distribution in which no bins are empty.

These two conditions guarantee the existence of a maximizing distribution which is interior, and this distribution must (by the theory of Lagrange multipliers) simultaneously satisfy (4) and (8).

The first condition implies that the space S of all distributions p is a compact space. Only a subset of this space satisfies the constraints (4); we call this subset S_{const}. Since the functionals B_k are linear, S_{const} is compact as well. The second condition guarantees that the interior of S_{const} is not empty, because it contains p_{int}. The entropy is a continuous function on S (and therefore is continuous on its compact subset S_{const}). Therefore, the entropy must attain its maximum somewhere in S_{const}.

To show that this maximizing distribution must be in the *interior* of S_{const}, it is necessary to examine the behavior of the entropy near the boundary of S_{const}. We call any distribution in which no bins are empty an *interior* distribution; distributions in which at least one bin is empty is called a *boundary* distribution. The entropy of any distribution p_{bound} on the boundary of S_{const} is lower than that of any interior distribution $p_{\text{bound}}(\mathbf{z})+sq(\mathbf{z})$ which is sufficiently close:

$$\begin{aligned} &E[p_{\text{bound}}+sq]-E[p_{\text{bound}}] \\ &\quad=\int[-sq(\mathbf{z})\ln(p_{\text{bound}}(\mathbf{z})+sq(\mathbf{z})) \\ &\qquad -p_{\text{bound}}(\mathbf{z})\ln(1+sq(\mathbf{z})/p_{\text{bound}}(\mathbf{z}))]dz\,. \qquad (10)\end{aligned}$$

Under the assumption of a finite number of bins for $\mathbf{z}$, this integral becomes a discrete sum. Since $p_{\text{bound}}(\mathbf{z})+sq(\mathbf{z})$ is interior, then $sq(\mathbf{z})$ must be positive at all vectors $\mathbf{z}$ where $p_{\text{bound}}(\mathbf{z})=0$. The boundary terms $sq(\mathbf{z})\ln sq(\mathbf{z})$ dominate the sum since all other terms are bounded by a constant multiple of s. It follows that the right-hand side of (10) always becomes positive as s approaches zero from above. Thus, no boundary distribution $p_{\text{bound}}(\mathbf{z})$ can possibly be a local maximum of the entropy. The maximum entropy distribution, which must be attained somewhere in S_{const}, cannot be attained on its boundary and therefore must be attained on its interior.

Stability of the Discretization Process. The discretization required by condition I is likely to be required by any numerical implementation of this procedure. The following argument provides some insight into whether this discretization is likely to introduce instability.

Let us consider probability distributions on two discretizations of the domain, one of bin width Δz and one of finer bin width $\Delta z/M$. We assume that the coarser of the two discretizations is sufficiently fine so that the constraining functionals B_k are constant on each bin. We further assume that each of the larger bins is the disjoint union of exactly M of the smaller bins.

Under these conditions, there are natural correspondences between probability distributions constructed on the two discretizations. Any probability distribution p' on the finer mesh defines a probability distribution $\text{lump}(p')$ on the coarser mesh, whose density on the bin $\mathbf{z}_i$ is the average of the densities $p'(\mathbf{z}'_j)$ over all bins $\mathbf{z}'_j$ contained in $\mathbf{z}_i$. Similarly, any probability distribution p on the coarser mesh defines a probability distribution $\text{split}(p)$ on the finer mesh, whose density at the bin $\mathbf{z}'_j$ is equal to $p(\mathbf{z}_i)$, where $\mathbf{z}_i$ is the bin which contains $\mathbf{z}'_j$. Since we have assumed that the constraining functionals are constant on each bin, it follows that if p' and p satisfy the constraints (3), then so do $\text{lump}(p')$ and $\text{split}(p)$.

Elementary properties of the entropy (2) imply that $E[p']\leqq E[\text{lump}(p')]$, with equality achieved when $p'=\text{split}(\text{lump}(p'))$. That is, the finer distribution p' carries at least as much information as the coarser distribution $\text{lump}(p')$, and the amount of information in the two distributions is equal only if p' is the evenly-split subdivision of $\text{lump}(p')$.

Now let $p_{\max}$ be the constrained maximum-entropy distribution on the coarser discretization. It follows that $\text{split}(p_{\max})$ must be the constrained maximum-entropy distribution on the finer discretization. This is because $\text{split}(p_{\max})$ achieves the maximum entropy of all distributions p' with $\text{lump}(p')=p_{\max}$, and the existence of any distribution q' with a greater entropy would imply the existence of $\text{lump}(q')$, whose entropy would have to exceed that of $p_{\max}$.

This argument has shown that once the discretization of the domain is so fine that the constraining functionals are constant on each bin, further refinement produces no change in the maximum-entropy distribution. In a typical application, the constraining functionals will never be exactly constant over a bin of the discretization. To the extent that the constraining functionals can be closely approximated by step functions, the above argument holds in an approximate sense. Note, however, that if the domain of the probability distributions are infinite and uniform approximation of the constraining functionals by step functions cannot be guaranteed, instabilities may occur as the range of the numerical integration grows without bound. Under these circumstances, stability can be recovered by requiring the constraining functionals B_k to be uniformly-approximable by step functions (see the last two paragraphs of this section).

A Notion of Convergence. Let us postulate that the empirical distribution is an exhaustive sampling of a true input/output distribution $p(\mathbf{z})$ of the transducer under study. There is a useful sense in which the approximating distributions converge to $p(\mathbf{z})$. In general, two distributions may be thought of as similar if they yield similar values when integrated against test functions $f(\mathbf{z})$. We will show that if the test function f can be approximated by the weighting functions $B_k(\mathbf{z})$, then the sequence of approximating distributions will approach $p(\mathbf{z})$ as tested by f.

Assume that a test function $f(\mathbf{z})$ can be uniformly-approximated by a sequence of linear combinations of the weighting functions $B_k(\mathbf{z})$. Then for any arbitrarily-small tolerance ε, there is a finite linear combination of weighting functions which approximates $f(\mathbf{z})$ to within ε:

$$f_\varepsilon(\mathbf{z})=\sum_{k=0}^{K(\varepsilon)} c_k B_k(\mathbf{z})\,; \qquad |f_\varepsilon(\mathbf{z})-f(\mathbf{z})|<\varepsilon\,. \qquad (11)$$

Consider the approximating distribution $p_{K(\varepsilon)}(\mathbf{z})$ which is formed using all of the constraints that enter into this approximating sum. By the triangle inequality,

$$\begin{aligned} &|\int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}-\int f(\mathbf{z})p_{K(\varepsilon)}(\mathbf{z})d\mathbf{z}| \\ &\quad \leqq |\int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}-\int f_\varepsilon(\mathbf{z})p(\mathbf{z})d\mathbf{z}| \\ &\qquad +|\int f_\varepsilon(\mathbf{z})p(\mathbf{z})d\mathbf{z}-\int f_\varepsilon(\mathbf{z})p_{K(\varepsilon)}(\mathbf{z})d\mathbf{z}|. \end{aligned} \tag{12}$$

The first term must be less than ε because of the approximation (11). The second term must be zero because $p_{K(\varepsilon)}(\mathbf{z})$ is constrained to match $p(\mathbf{z})$ when integrated against each of the terms in the approximation (11). Thus, $p_{K(\varepsilon)}$ approximates p to within a tolerance ε when tested by the function f.

For example, in a finite domain, every piecewise continuous function $f(\mathbf{z})$ can be uniformly-approximated by a sequence of polynomials. Each polynomial is in turn a sum of monomials, and a monomial is a weighting function which corresponds to a moment or a cross-correlation. Therefore, in a finite domain, the moments and cross-correlations form a set of constraints which results in a sequence of approximating distributions which converge to the actual distribution when tested by integration against any continuous function.

Practical Considerations. Conditions I and II above are sufficient but not necessary conditions for the existence of a unique maximizing distribution. Later we will show that for a particular choice of constraints [the input/output cross-correlations of Lee and Schetzen (1965)], one may demonstrate that a maximizing distribution exists without postulating either Condition I or Condition II. Additionally, our Example 2 below demonstrates the existence of maximizing distributions not derived from the Wiener series when Condition II does not hold.

On the other hand, it is possible to choose constraints in such a way that on an unbounded domain for $\mathbf{z}$, no maximizing distribution exists. This pathology appears even in the approximation of one-dimensional distributions on the real line by maximum-entropy distributions based on moments. The maximum-entropy distribution, if it exists, is the exponential of a polynomial of order equal to the highest moment specified (8). But if this polynomial is of odd degree, the integrals that appear in the Lagrange multiplier conditions (4) diverge unless the distribution is restricted to a semi-infinite interval. Thus, the highest moment specified must be of even order. For example, if only the mean but not the variance of a distribution is specified, no maximum-entropy distribution will exist on the real line: the greater the spread of the distribution, the greater the entropy. If the mean and variance are specified, the maximum-entropy distribution is a Gaussian of matching mean and variance. If mean, variance, and skewness are specified, the maximum-entropy distribution also does not exist: a nearly-Gaussian distribution with a small asymmetry on its tail can be constructed with skewness to match the constraint. As this asymmetry is decreased in size and is moved further and further from the center of the distribution, the entropy of the distribution approaches but does not attain the entropy of a Gaussian. [For distributions restricted to the non-negative reals, this particular pathology does not appear. For example, the maximum-entropy distribution with only the mean constrained is the exponential distribution (Montroll and Shlesinger 1983).]

If the highest moment specified is of even order and the domain of $\mathbf{z}$ is finite, we have seen above that a maximum-entropy distribution will always exist if any interior distribution exists. However, the polynomial exponent of this distribution may have a positive leading term. While this does not prevent performance of the integrations (3) on a finite domain, it does prevent extension of this distribution to an infinite domain. The simplest example of this pathology is the approximation of a symmetric distribution by a maximum-entropy distribution of specified variance and kurtosis. A positive leading term in the exponent will be required if the kurtosis is sufficiently large. In this case, no maximum-entropy distribution exists.

These pathologies cannot occur with real data collected on a finite domain if the domain of the distribution is restricted to that of the data. Alternatively, a simple maneuver will allow a distribution calculated on a finite domain to be extended to the infinite domain. The exponent in (8) must be bounded from above for large $\mathbf{z}$. One way to accomplish this is to use weighting functions $B_k(\mathbf{z})$ that are zero for sufficiently large $\mathbf{z}$. For example, if moments are used as constraints, moments should be calculated after discarding outliers. We use this approach in Example 1 below. If the data do not have extensive "tails," then it appears that such caution is not necessary. This is the case in Example 2 below.

Relation to the Wiener Orthogonal Series

We begin by defining the familiar objects of the Wiener white-noise theory (Marmarelis and Marmarelis 1978) in a manner suitable for the present analysis. The univariate Gaussian weighting of variance V will be denoted

$$\mathrm{Gau}(x, V)=(2\pi)^{-1/2}V^{-1}\exp(-x^2/2V). \tag{13}$$

The univariate Hermite polynomial of order n with respect to $\mathrm{Gau}(x, 1)$ ($\equiv\mathrm{Gau}(x)$) will be denoted by

$h_m(x)$. We use the normalization

$$\int h_m(x)h_{m'}(x)\,\mathrm{Gau}(x)dx = \begin{cases} 0, & m \neq m', \\ m!, & m = m'. \end{cases} \tag{14}$$

It is convenient to use vector indices **m**, **n**, etc., with the following conventions: All entries $m_0, m_1, m_2, \ldots$ must be nonnegative integers. Only a finite number of entries may be nonzero (unless otherwise noted). The *total order* of an index **m**, to be written $s(\mathbf{m})$, is the sum of its entries. The symbol $\mathbf{m}!$ denotes $m_0! \cdot m_1! \cdot m_2! \cdot \ldots$. Moments and generalized input/output cross-correlations are parametrized by two vector indices **m** and **n** and will be denoted $M_{\mathbf{m};\mathbf{n}}$. We define du to be the measure on the input/output space that corresponds to the input/output relation of the transducer under study. The moments and cross-correlations are averages of products of input and output weighted by the measure du:

$$M_{\mathbf{m};\mathbf{n}} = \int x_0^{m_0} x_{-1}^{m_1} \cdot \ldots \cdot y_0^{n_0} y_{-1}^{n_1} \cdot \ldots \cdot du. \tag{15}$$

The moments $M_{\mathbf{m};\mathbf{0}}$ are moments of the input alone; the moments $M_{\mathbf{0};\mathbf{n}}$ are moments of the output alone. The mean squared output, to be denoted V, is equal to $M_{\mathbf{0};\langle 2,0,0,\ldots\rangle}$. The cross-correlations of the output at time zero and the inputs at multiple previous times play a special role in the Lee and Schetzen (1965) method of evaluating Wiener kernels; they will be denoted by

$$c_{\mathbf{m}} = M_{\mathbf{m};\langle 1,0,0,\ldots\rangle}. \tag{16}$$

The integration in (15) is over an infinite-dimensional space. Concepts from measure theory in infinite-dimensional product spaces (Halmos 1950, pp. 154–160) are required to give this equation (and other similar equations to follow) a rigorous interpretation. A *cylinder set* is a direct product of open intervals in which a finite number of the intervals are bounded, and the rest of the intervals consist of the entire real axis. A measure specifies a weighting on cylinder sets. The integration in (15) is a limit of weighted sums over cylinder sets in which the coordinates corresponding to the bounded intervals are the coordinates that appear in the integrand.

If the measure du is smooth (i.e., corresponds to a density), then the integral (15) can be recast as an ordinary multidimensional integral. Since only a finite number of entries in the vector indices **m** and **n** are nonzero, we may assume that $m_i = 0$ and $n_i = 0$ for $i > I$.

If the measure du has a density p, then the weight of an infinitesimally-narrow cylinder set

$$(x_0, x_0 + dx_0) \cdot \ldots \cdot (x_{-I}, x_{-I} + dx_{-I})$$
$$\times (y_0, y_0 + dy_0) \cdot \ldots \cdot (y_{-I}, y_{-I} + dy_{-I})$$

containing a point **z** approaches the product

$$p(x_0, \ldots, x_{-I}; y_0, \ldots, y_{-I})dx_0 \ldots dx_{-I}dy_0 \ldots dy_{-I}.$$

Here, p is the probability density function corresponding to the measure du. The above expression is the probability that the first $I+1$ **x**-components of **z** and the first $I+1$ **y**-components of **z** all lie within the volume element centered at $(x_0, \ldots, x_{-I}; y_0, \ldots, y_{-I})$. The integral expression (15) for the generalized moment now becomes an ordinary integral over $2(I+1)$ coordinates:

$$\begin{aligned} M_{\mathbf{m};\mathbf{n}} = \int & x_0^{m_0} \cdot \ldots \cdot x_{-I}^{m_I} y_0^{n_0} \cdot \ldots \cdot y_{-I}^{n_I} \\ & \times p(x_0, \ldots, x_{-I}; y_0, \ldots, y_{-I}) \\ & \times dx_0 \ldots dx_{-I}dy_0 \ldots dy_{-I}. \end{aligned} \tag{17}$$

These concepts enable us to define in a rigorous way ensemble averages with respect to Gaussian white noise. An ensemble of inputs drawn from a Gaussian white noise of unit variance corresponds to a measure $dW(\mathbf{x})$ on the space of possible inputs **x**. This measure is the formal product

$$dW(\mathbf{x}) = \mathrm{Gau}(x_0)\,\mathrm{Gau}(x_{-1})\,\mathrm{Gau}(x_{-2}) \ldots d\mathbf{x}.$$

That is, each input value x_{-i} is distributed independently with weighting $\mathrm{Gau}(x_{-i})dx_{-i}$. An integral over all input signals reduces to an integral over a finite-dimensional multivariate Gaussian distribution provided that only a finite number of the input values x_{-i} enter into the integrand. If an infinite number of values x_{-i} enter into the integrand, no such reduction is possible; a rigorous interpretation of the integral requires a limiting process and assumptions such as "finite memory" and "finite bandwidth."

The multivariate Hermite polynomials $h_{\mathbf{m}}(\mathbf{x})$ are defined in terms of their univariate analogs by

$$h_{\mathbf{m}}(\mathbf{x}) = h_{m_0}(x_0) \cdot h_{m_1}(x_{-1}) \cdot \ldots. \tag{18}$$

Let the input signal **x** be drawn from an ensemble according to a Gaussian white measure $dW(\mathbf{x})$. Thus, each input value x_i is independently distributed with weight $\mathrm{Gau}(x_i)$. The multivariate Hermite polynomials are orthogonal with respect to this input:

$$\int h_{\mathbf{m}}(\mathbf{x})h_{\mathbf{m}'}(\mathbf{x})dW(\mathbf{x}) = \begin{cases} 0, & \mathbf{m} \neq \mathbf{m}', \\ \mathbf{m}!, & \mathbf{m} = \mathbf{m}'. \end{cases} \tag{19}$$

This equation follows from the basic orthogonality relation (14) and the interpretation of the integral (18) over the measure $dW(\mathbf{x})$ as a finite-dimensional integral over cylinder sets.

The Wiener Series for a Deterministic Transducer. The output of a deterministic transducer T at time 0 in response to the input **x** may be expressed as a sum of orthogonal functionals of ascending order:

$$y_0 = \sum_{r=0}^{\infty} \sum_{s(\mathbf{m})=r} g_{\mathbf{m}} h_{\mathbf{m}}(\mathbf{x}). \tag{20}$$

The coefficients $g_{\mathbf{m}}$ can be determined by cross-correlation of lagged products of the input and the output y_0 (Lee and Schetzen 1965). The general relationship follows from the orthogonality condition (19):

$$g_{\mathbf{m}} = (1/\mathbf{m}!) \int y_0 h_{\mathbf{m}}(\mathbf{x}) du . \tag{21}$$

This integral is over the joint distribution of the output at time zero, y_0, and all possible input vectors $\mathbf{x}$. For deterministic transducers, this reduces to a simpler form: given any input vector $\mathbf{x}$, there is only one possible output value $y_0 = T[\mathbf{x}]$. Thus, the integral over the output values collapses:

$$g_{\mathbf{m}} = (1/\mathbf{m}!) \int T[\mathbf{x}] h_{\mathbf{m}}(\mathbf{x}) dW(\mathbf{x}) . \tag{22}$$

If no entry in $\mathbf{m}$ is greater than 1, this reduces to

$$g_{\mathbf{m}} = c_{\mathbf{m}} . \tag{23}$$

If $\mathbf{m}$ has entries greater than 1, then $g_{\mathbf{m}}$ is equal to $(1/\mathbf{m}!)c_{\mathbf{m}}$ plus correction terms. The correction terms consist of cross-correlations $c_{\mathbf{m}'}$ of total order $s(\mathbf{m}')$ strictly less than the total order $s(\mathbf{m})$. Thus, a complete set of cross-correlations of order up to and including a particular total order determines the coefficients $g_{\mathbf{m}}$ of order up to and including that order. Moments of order greater than 1 in the output never enter into the calculation of the Wiener series.

A Stochastic View of the Wiener Approximations. Let us consider the jth Wiener approximation to a transducer T, given by a truncation of the series (20) at order j. This truncated series describes the input/output relationship of an approximating transducer, which we will call T_j.

Cross-correlations c_m (which are first order in the output) will be identical for the transducer T and its approximant T_j, provided that the total order $s(\mathbf{m})$ in the input is j or less. However, moments and cross-correlations $M_{\mathbf{m},\mathbf{n}}$ that are of total order $s(\mathbf{n})$ greater than 1 in the output will (in general) be different for T and T_j. In particular, the mean squared output of T_j is given by

$$\begin{aligned} V(T_j) &= \left[\sum_{s(\mathbf{m}) \leqq j} g_{\mathbf{m}} h_{\mathbf{m}}(\mathbf{x}) \right]^2 dW(\mathbf{x}) \\ &= \sum_{s(\mathbf{m}) \leqq j} \mathbf{m}! (g_{\mathbf{m}})^2 , \end{aligned} \tag{24}$$

where the orthogonality relations (19) have been used.

The mean squared output $V(T)$ of the full transduction T in general exceeds that of the approximating transducer, $V(T_j)$. This difference, often called the mean squared error, is often used as an indicator of the goodness of fit of the finite approximation T_j. The mean squared error of the jth approximating transducer is given by

$$\begin{aligned} \mathrm{MSE}_j &= V(T) - V(T_j) \\ &= \sum_{s(\mathbf{m}) > j} \mathbf{m}! (g_{\mathbf{m}})^2 . \end{aligned} \tag{25}$$

The deterministic interpretation is that the residual mean squared error reflects inaccuracies of the predicted response to the ensemble of white noise inputs. There is also a stochastic interpretation of the mean squared error: the transducer's *average* response to a particular sample of white-noise input is given exactly by T_j, but there is a Gaussian noise added to the output whose variance is equal to the mean squared error MSE_j.

These possibilities cannot be distinguished by measurements of the cross-correlations $c_{\mathbf{m}}$ of total order $s(\mathbf{m})$ up to and including j, but can be distinguished by measurements of higher-order cross-correlations. For a transducer which in fact is T_j plus additive noise, the higher-order cross-correlations will be zero and the MSE will not be improved by the inclusion of higher-order terms in the Wiener expansion. On the other hand, if the transducer is deterministic, inclusion of all terms in the Wiener series (20) will (in principle) result in a MSE of zero.

It is unlikely that a biological transducer is either purely deterministic or is characterized by purely additive Gaussian white noise superimposed on an otherwise deterministic transduction. Most likely, there is correlation structure in the variability of the response, and response variability may well depend in a complex manner on the input. This information is contained in moments and cross-correlations $M_{\mathbf{m},\mathbf{n}}$ of total order $s(\mathbf{n})$ in the *output* of 2 or more. Such statistics, therefore, reflect the internal structure of the physiological system.

These moments are not incorporated into the Wiener series, which is an orthogonal series approximation of the average output. However, the information contained in these moments can be incorporated into a description of the transducer by means of the maximum-entropy approach. In the next subsection, we will show that the maximum-entropy approach reduces to the Wiener series when the variance of the output is included in the constraints, but all other moments of order greater than one in the output are neglected.

The Wiener Approximants are Maximum-Entropy Distributions. The stochastic interpretation of T_j can be made explicit by displaying its corresponding input/output probability distribution, $p_j(\mathbf{x}, y_0)$:

$$p_j(\mathbf{x}, y_0) dy_0 d\mathbf{x} = \mathrm{Gau}(y_0 - T_j[\mathbf{x}], \mathrm{MSE}_j) dy_0 dW(\mathbf{x}) . \tag{26}$$

To see that this probability distribution is a stochastic interpretation of T_j, it is sufficient to calculate the Wiener kernels $g_{\mathbf{m}}(p_j)$ of the stochastic transduction it describes. A cross-correlation formula (21) is most convenient for this purpose:

$$\begin{aligned} g_{\mathbf{m}}(p_j) &= (1/\mathbf{m}!) \int y_0 \mathbf{h}_{\mathbf{m}}(\mathbf{x}) p_j(\mathbf{x}, y_0) dy_0 d\mathbf{x} \\ &= (1/\mathbf{m}!) \int \Big[v + \sum_{s(\mathbf{m}') \leqq j} g_{\mathbf{m}'} \mathbf{h}_{\mathbf{m}'}(\mathbf{x}) \Big] \\ &\quad \times \mathbf{h}_{\mathbf{m}}(\mathbf{x}) \, \mathrm{Gau}(v, \mathrm{MSE}_j) dv dW(\mathbf{x}) \\ &= \begin{cases} g_{\mathbf{m}}, & s(\mathbf{m}) \leqq j, \\ 0, & s(\mathbf{m}) > j. \end{cases} \end{aligned} \tag{27}$$

The second equality is a consequence of the substitution of v for the deviation of y_0 from its average value given the input $\mathbf{x}$:

$$v = y_0 - T_j[\mathbf{x}] = y_0 - \sum_{s(\mathbf{m}') \leqq j} g_{\mathbf{m}'} \mathbf{h}_{\mathbf{m}'}(\mathbf{x}).$$

The final equality in (27) follows from the orthogonality relations (19).

Thus, the Wiener kernels of p_j are equal to those of T_j: kernels of order j or less are identical to the kernels of T; kernels of order greater than j are zero. However, the mean squared output $V(p_j)$ of this stochastic transduction is greater than that of T_j; a Gaussian of variance MSE_j has been superimposed. This is exactly what is required to make the mean squared output $V(p_j)$ of the stochastic transducer equal to $V(T)$ rather than to $V(T_j)$:

$$\begin{aligned} V(p_j) &= \int (y_0)^2 p_j(\mathbf{x}, y_0) dy_0 d\mathbf{x} \\ &= \int \Big[v + \sum_{s(\mathbf{m}) \leqq j} g_{\mathbf{m}} \mathbf{h}_{\mathbf{m}}(\mathbf{x}) \Big]^2 \mathrm{Gau}(v, \mathrm{MSE}_j) dv dW(\mathbf{x}) \\ &= \int v^2 \, \mathrm{Gau}(v, \mathrm{MSE}_j) dv dW(\mathbf{x}) + \sum_{s(\mathbf{m}) \leqq j} \mathbf{m}! \, (g_{\mathbf{m}})^2 \\ &= \mathrm{MSE}_j + V(T_j) \\ &= V(T). \end{aligned} \tag{28}$$

Now consider a maximum-entropy estimate for T, constrained to match the cross-correlations $c_{\mathbf{m}}(T)$ of order j or less and to match the variance $V(T)$. Since the input is known to be Gaussian white noise, moments of all orders in the input $M_{\mathbf{m};0}$ are also constrained to conform to the statistics of Gaussian white noise. The weighting functions $B_k(\mathbf{x}, y_0)$ for these constraints fall into three groups: (i) monomials of all orders involving components of $\mathbf{x}$ alone, corresponding to the known statistics of the input, (ii) monomials of total order no more than j in $\mathbf{x}$ and first order in y_0, corresponding to the cross-correlations, and (iii) y_0^2, corresponding to the mean squared output.

Equations (27) and (28) demonstrate that the distribution $p_j(\mathbf{x}, y_0)$ matches the transducer T when tested with the required functionals. Furthermore, the form of the distribution is manifestly of the form (8) demanded by the Lagrange multiplier theory. It follows from the results in the previous section that $p_j(\mathbf{x}, y_0)$ is the unique maximum-entropy distribution whose input/output statistics [(i), (ii), and (iii)] match those of the transducer T.

This argument extends without difficulty to a distribution $p_j(\mathbf{x}, \mathbf{y})$ defined on the past history of the output as well as its present value. The maximum-entropy distribution corresponds to the solution of the above Lagrange multiplier problem, to which additional constraints have been added. The additional weighting functions and their corresponding constraints are: (ii′) monomials of total order no more than j in $\mathbf{x}$ and first order in $\mathbf{y}$, corresponding to cross-correlations first-order the input at previous times, and (iii′) y_i^2, corresponding to mean squared output at previous times. Because of the hypothesis of stationarity, these moments and cross-correlations have values identical to their values at time zero. The formal expression for the density of the maximum-entropy distribution is

$$\begin{aligned} & p_j(\mathbf{x}, \mathbf{y}) d\mathbf{y} d\mathbf{x} \\ & \quad = \prod_{i=0}^{\infty} \mathrm{Gau}(y_i - T_j[S_i[\mathbf{x}]], \mathrm{MSE}_j) d\mathbf{y} dW(\mathbf{x}), \end{aligned} \tag{29}$$

where $S_i[\mathbf{x}]$ denotes a shift of $\mathbf{x}$ backwards in time by an amount $i\Delta t$:

$$S_i[\mathbf{x}] = \langle x_{-i}, x_{-i-1}, x_{-i-2}, \dots \rangle.$$

Integrations with respect to this measure are carried out using the notion of cylinder sets, as described in connection with (17).

Examples

We present two examples that illustrate many of the ideas discussed above. The first example focuses on the how one can distinguish additive noise from other kinds of sources of variability. The second example shows how maximum-entropy estimates can be used to characterize a system in which there is no accessible input, but in which there are two stochastically-coupled outputs that are observed simultaneously. In both cases, the important points are evident without having to consider dynamics. Therefore, in these examples, we will restrict our attention to static (memoryless) stochastic systems.

Example 1: A Linear System with Noise

In the first example, we show how the maximum-entropy estimates distinguish between purely additive noise and more complex sources of variability. Consider a transducer whose output y equals its instantaneous input x to which a Gaussian white noise w scaled by a factor a has been added:

$$y = x + aw. \tag{30}$$

We assume that the Gaussian white noise w is uncorrelated with the input. Let us test this transducer with an input x drawn from a Gaussian white noise of unit variance. Under this circumstance, the input/output probability distribution is given exactly by

$$p_{\mathrm{add}}(x, y) = \mathrm{Gau}(x)\,\mathrm{Gau}(y - x, a^2). \tag{31}$$

This distribution is shown in Fig. 1a for $a = 1$.

Consider a second transducer in which the noise is multiplicative:

$$y = x(1 + aw). \tag{32}$$

When tested with an input x drawn from a Gaussian white noise of unit variance, the input/output probability distribution of this transducer is given exactly by

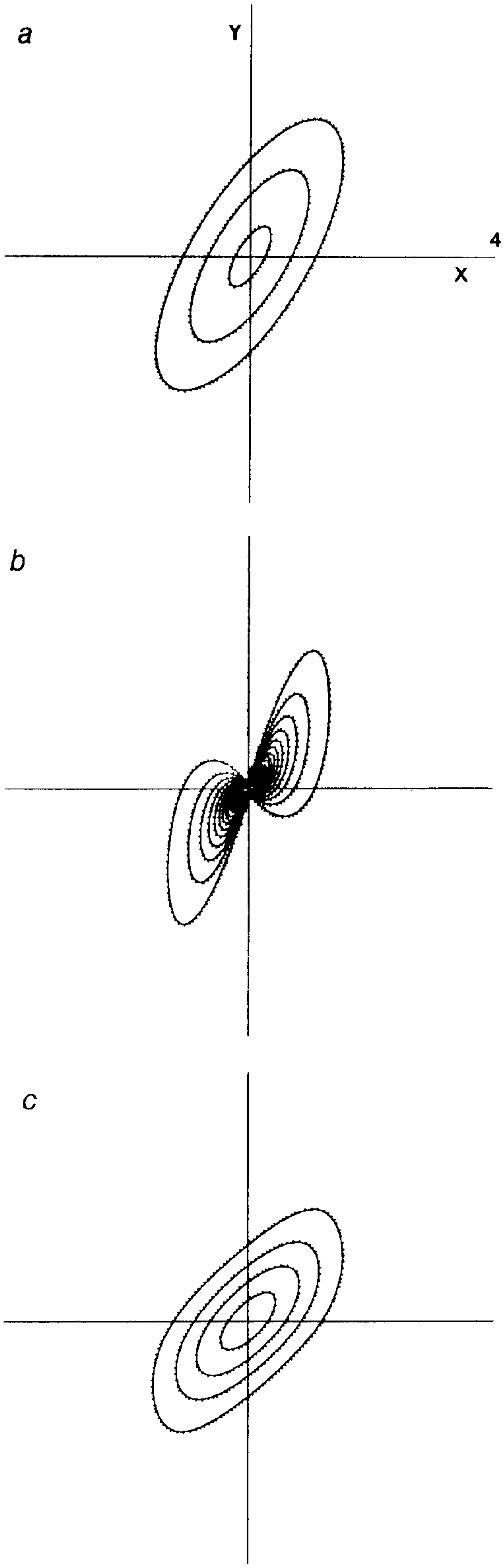
$$p_{\mathrm{mult}}(x, y) = \mathrm{Gau}(x)\,\mathrm{Gau}(y - x, a^2 x^2), \tag{33}$$

which is shown in Fig. 1b for $a = 1$. There is a singularity at the origin; $p_{\mathrm{mult}}(0, y) = \mathrm{Gau}(0)\delta(y)$.

Clearly the two transducers are quite different in character. However, for both transducers, the average output is equal to the input. Thus, the Wiener series for both transducers are identical, and terminate after the first term: $y = x$. Both transducers also have the same output variance, $1 + a^2$. Thus, the MSE of the first (and higher-order) Wiener approximations will all be a^2.

The transducers are readily distinguished by examining statistics whose total order in input and output is four or more and whose order in the output alone is two or more. (Cross-correlations whose total order is odd must be zero, and cross-correlations which are at most first-order in the output are the same for both transducers.) For the transducer (30) with additive noise, the fourth-order cross-correlations and moments are:

$$\begin{aligned}
&M_{0,4}[p_{\mathrm{add}}] = 3 + 6a^2 + 3a^4, \\
&M_{1,3}[p_{\mathrm{add}}] = 3 + 3a^2, \\
&M_{2,2}[p_{\mathrm{add}}] = 3 + a^2, \\
&M_{3,1}[p_{\mathrm{add}}] = 3, \\
&M_{4,0}[p_{\mathrm{add}}] = 3.
\end{aligned} \tag{34}$$



◀

Fig. 1. **a** The input/output probability distribution p_{add} (31) of a transduction with additive noise. **b** The input/output probability distribution p_{mult} (33) of a transduction with multiplicative noise. There is a singularity at the origin. **c** The maximum entropy estimate for p_{mult} using moments and cross-correlations up to order four. In all contour maps, each contour line represents a probability density of 0.05

For the transducer (32) with multiplicative noise, the fourth-order cross-correlations and moments are:

$$\begin{aligned}
&M_{0,4}[p_{\text{mult}}]=3+18a^2+9a^4\,,\\
&M_{1,3}[p_{\text{mult}}]=3+9a^2\,,\\
&M_{2,2}[p_{\text{mult}}]=3+3a^2\,,\qquad\qquad (35)\\
&M_{3,1}[p_{\text{mult}}]=3\,,\\
&M_{4,0}[p_{\text{mult}}]=3\,.
\end{aligned}$$

In the case of additive noise, the maximum-entropy estimate constructed from moments and cross-correlations of total order 2 is exactly the true distribution (31), and hence further constraints from higher-order cross-correlations do not alter the estimate. In contrast, the maximum-entropy estimate for the multiplicative-noise distribution p_{mult} is identical to that for p_{add} when only second-order statistics are considered, but is different when fourth-order statistics are considered.

To demonstrate this, a maximum-entropy estimate for p_{mult} was calculated by solving (numerically) (4) and (8) using the cross-correlations of (35). In this calculation, it is necessary to limit the range of input and output values in order to obtain a solution, because the tails of the distribution (33) are larger than those of a Gaussian of equivalent variance; for this reason, only data within four standard deviations of the mean were retained. The resulting approximating distribution, illustrated in Fig. 1c, shows some of the features of the actual distribution p_{mult}, in that its dispersion in the y-direction is smaller than that of the Gaussian (Fig. 1a) near $x=0$, and larger than that of the Gaussian for large x. However, the exact distribution p_{mult} has a singularity near the origin, which is not suggested by the approximate distribution of Fig. 1c.

Thus, the maximum-entropy estimate based on cross-correlations and moments of total order two succeeds in distinguishing additive from multiplicative noise, but gives a relatively poor estimate of the true distribution. This is most evident at the origin, where the true distribution has a singularity but any maximum-entropy estimate based on cross-correlations must be smooth [cf. (8)]. The way in which the maximum-entropy estimates err suggests that the procedure may be made more efficient by basing it on functionals other than cross-correlations which have a more rapid variation in their weights near the origin. In particular, the exact expression (33) for p_{mult} is of the maximum-entropy form for a set of constraints consisting of moments of all orders in the input (which are known, since the input is Gaussian white noise) and functionals with weights y/x and $(y/x)^2$, rather than cross-correlations.

This observation suggests a general approach to tailoring the maximum-entropy approach to the system under study: the first step might consist of using cross-correlations and moments of total order two or less; this will provide an estimate which is a Gaussian. Subsequent functionals might be chosen to have large variations in their weights in the regions in which the empiric distribution deviates from this best-fitting Gaussian. In this regard, logarithmic functions have been proposed as being generally useful constraints for a maximum-entropy formalism in which the distributions are highly skewed (Montroll and Shlesinger 1983).

Example 2:
Two Nonlinearly-Coupled Stochastic Processes

We illustrate the application of the maximum-entropy method to a system in which there are two outputs x and y that can be observed, but no accessible input. We will examine the maximum-entropy estimates for two related model systems.

In the first system, x and y are random variables subject to the constraint $x^2+y^2=R^2$. We may think of x and y as the scaled cosine and sine of an unobservable angle distributed randomly around the circle. For $R=2^{1/2}$, x and y have unit variance. The joint distribution of x and y is most conveniently expressed in terms of $r^2\equiv x^2+y^2$:

$$p_{\text{rim}}(x,y)=\delta(r^2-2)/\pi\,. \qquad (36)$$

In the second system, x and y are random variables constrained to lie within a given circle: $x^2+y^2\leqq R'^2$. For $R'=2$, x and y have unit variance.

$$p_{\text{disk}}(x,y)=\begin{matrix}1/4\pi\,, & r\leqq 2\,,\\ 0\,, & r>2\,.\end{matrix} \qquad (37)$$

The two distributions p_{rim} and p_{disk} both have circular symmetry. Therefore cross-correlations of odd order either in x or in y will be zero. In particular, the cross-correlation $M_{1,1}$ is zero for both distributions. We have scaled the distributions so that x and y have unit variance in both cases. Thus, the second-order statistics of both distributions agree with that of a product of two Gaussians of unit variance, which therefore must be the second-order maximum-entropy estimate.

As in Example 1, statistics of higher order serve to discriminate between p_{rim} and p_{disk}. The maximum-entropy estimates for these two distributions were calculated using constraints consisting of all moments and cross-correlations of total order not exceeding two, four, six, and eight. The radial profiles of these distributions are shown in Fig. 2. Although the maximum-entropy estimates of order two are identical

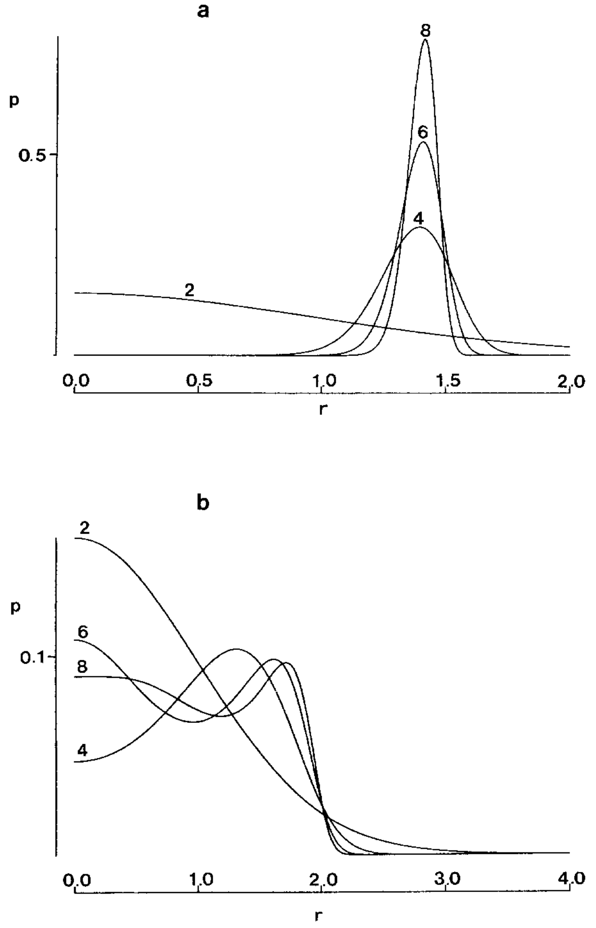


Fig. 2. **a** The radial dependence of a sequence of maximum entropy estimates for p_{rim} (36) of two random variables constrained to lie on the rim of a circle of radius $2^{1/2}$. **b** The radial dependence of a sequence of maximum entropy estimates for p_{disk} (37) of two random variables constrained to lie within a circle of radius 2. In both parts, the maximum entropy estimates are generated using moments and cross-correlations up to orders two, four, six, and eight. The maximum entropy estimates of order two are identical in the two cases

for both distributions, the order-four estimates are very different and give rather good qualitative pictures of the actual distributions. Higher-order estimates improve the approximation further, but the largest improvement is from order two to order four.

Acknowledgements. JV is supported in part by grants EY188, EY2349, EY5466, and NS877 from the US National Institutes of Health, The Hartford Foundation, The Harry Frank Guggenheim Foundation, and the McKnight Foundation. We thank Professor Ken-Ichi Naka and the National Institute of Basic Biology, Okazaki, Japan, for providing the opportunity for the discussions that led to this work.

References

Crowder HK, McCuskey SW (1964) Topics in higher analysis. The Macmillan Company, New York

Halmos P (1950) Measure theory. Van Nostrand Reinhold, New York

Jaynes ET (1979) Where do we stand on maximum entropy? In: Levine RD, Tribus M (eds) The maximum entropy formalism. The MIT Press, Cambridge, pp 15–118

Jaynes ET (1982) On the rationale of maximum-entropy methods. Proc IEEE 70:939–952

Krausz HI (1975) Identification of nonlinear systems using random impulse trains. Biol Cybern 19:217–230

Lee YN, Schetzen M (1965) Measurement of the kernels of a nonlinear system by cross-correlation. Int J Control 2:237–254

Marmarelis VZ (1977) A family of quasi-white random signals and its optimum use in biological system identification. I. Theory. Biol.Cybern 27:49–56

Marmarelis PZ, Marmarelis VZ (1978) Analysis of physiological systems: The white-noise approach. Plenum, New York

Marmarelis PZ, Naka K-I (1972) White noise analysis of a neuron chain: an application of the Wiener theory. Science 175:1276–1278

Montroll EW, Shlesinger MF (1983) Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: a tale of tails. J Stat Phys 32:209–229

Naka KI, Marmarelis PZ, Chan R (1975) Morphological and functional identification of catfish retinal neurons. III. Functional identification. J Neurophysiol 38:92–131

Sakuranaga M, Naka K-I (1985a) Signal transmission in the catfish retina. I. Transmission in the outer retina. J Neurophysiol 53:373–389

Sakuranaga M, Naka K-I (1985b) Signal transmission in the catfish retina. II. Transmission to type-*N* cell. J Neurophysiol 53:390–410

Sakuranaga M, Naka K-I (1985c) Signal transmission in the catfish retina. III. Transmission to type-*C* cell. J Neurophysiol 53:411–428

Victor JD, Knight BW (1979) Nonlinear analysis with an arbitrary stimulus ensemble. Q Appl Math 37:113–136

Wiener N (1958) Nonlinear problems in random theory. Wiley, New York

Yasui S (1979) Stochastic functional Fourier series, Volterra series, and nonlinear systems analysis. IEEE Trans AC-24:230–242

Yasui S, Davis W, Naka K-I (1979) Spatio-temporal receptive field measurement of retinal neurons by random pattern stimulation and cross correlation. IEEE Trans BME-26:263–272

Received: February 2, 1986

J. D. Victor
The Rockefeller University
1230 York Avenue
New York, NY 10021
USA