Pacific Symposium on Biocomputing 6:251-262 (2001)

Analyzing sensory systems with the information distortion function

Alexander G Dimitrov and John P Miller
Center for Computational Biology
Montana State University
Bozeman, MT 59715-3505
{*alex,jpm*}*@nervana.montana.edu*

The nature and information content of neural signals have been discussed extensively in the neuroscience community. They are important ingredients in many theories on neural function, yet there is still no agreement on the details of neural coding. There have been various suggestions about how information is encoded in neural spike trains: by the number of spikes, by temporal correlations, through single spikes, or by spike patterns in one, or across many neurons. The latter scheme is most general and encompasses many others. We present an algorithm which can recover a coarse representation of a pattern coding scheme, through quantization to a reproduction set of smaller size. Among many possible quantizations, we choose one which preserves as much of the informativeness of the original stimulus/response relation as possible, through the use of an information-based distortion function. This method allows us to study coarse but highly informative models of a coding scheme, and then to refine them when more data becomes available. We shall describe a model in which full recovery is possible and present example for cases with partial recovery.

1 Introduction

When discussing neural systems, one of the questions we are interested in is how the activity of a set of neurons represents the input to these neurons from other cells or from the environment. The nature and information content of neural signals have been discussed extensively in the neuroscience community. They are important ingredients in many theories on neural function, yet there is still no agreement on the details of neural coding. There have been various hypotheses about how information is encoded in neural spike trains: by the number of spikes, by temporal correlations between spikes, through single spikes, or by complete temporal patterns of spikes from a single neuron or groups of neurons [10]. The latter scheme is most general and encompasses many others.

The search for pattern codes requires exponentially more data than the search for mean rate or correlation codes [9]. We will describe a method that allows us to uncover as much of the details of a coding scheme as is supported by the available data, by quantizing the set of responses to a smaller reproduction set of finite size. To assess the quality of the quantization we use an information-based distortion measure. The quantization is optimized to have minimal distortion for a fixed reproduction size. This method allows us to study coarse models of coding schemes which can be refined as more data becomes available.

2 Quantizing neural responses

With communication systems, the usual use of information theory is to design a coding scheme given the structure of the communication channel. Our application differs since now we are analyzing an already implemented neural coding scheme (the cricket cercal system in particular [2,5]). Our goal is to uncover the structure of the the scheme from observations of the stimulus and response properties of a neural system.

2.1 Quantizing the response

The basic concepts of information theory are the entropy $H(X)$ and the mutual information I(X,Y) of random sources (X, Y) [12,1]. The information quantities H and I depend only on the underlying probability function and not on the structure of the event space. This allows us to estimate them in cases where more traditional statistical measures (e.g., variance, correlations, etc.) simply do not exist. There is a drawback though, since now we must either model the necessary probabilities or use large amounts of data to estimate them nonparametrically. As pointed out by Johnson et.al. [9], the amount of data needed to support coding schemes which contain long sequences (length T) across multiple neurons (N) grows exponentially with T and N. It is conceivable that for some systems the required data recording time may well exceed the expected lifespan of the system.

To resolve this issue we need to sacrifice some detail in the description of the coding scheme in order to obtain robust estimates of a coarser description. This can be achieved through quantization [1,7] of the neural representation Y into a coarser representation in a smaller event space Y_N. Y_N is referred to as the *reproduction* of Y. Most of the results in this section are valid for the general case of continuous, ergodic random variables [7]. The formulation for the most general case requires special attention to details though. For clarity of the presentation here we shall assume that all random variables are finite and discrete.

Quantizers are maps from one probability space to another. They can be deterministic (functions) or stochastic (given through a conditional probability) [11]. We shall consider the most general case of a stochastic quantizer $q(y_N|y)$ – the probability of a response y belonging to an abstract class y_N. A deterministic quantizer $f : Y \to Y_N$ is a special case in which q takes values 0 or 1 only. In both cases, stimulus, response and reproduction form a Markov chain $X \to Y \to Y_N$. In information theory the quality of a quantization is characterized by a distortion function [6]. We shall look for a minimum distortion quantization using an information distortion function and discuss its relationship to the codebook estimation problem.

2.2 A distortion measure based on mutual information

In engineering applications, the distortion function is usually chosen in a moderately random fashion [1,6], and is the one that introduces structures in the original space, to be preserved by the quantization. We can avoid this arbitrariness since we expect that the neural system is already reflecting pertinent structures of the sensory stimuli and we would like to preserve this in the reproduction. Thus our choice of distortion function is determined by the informativeness of the quantization. The mutual information $I(X;Y)$ tells us how many different states on the average can be distinguished in X by observing Y. If we quantize Y to Y_N (a reproduction with N elements), we can estimate $I(X;Y_N)$ - the mutual information between X and the reproduction Y_N. Our information preservation criterion will then require that we choose a quantizer that preserves as much of the mutual information as possible, i.e., the quantizer $q(y_N|y)$ which minimizes the difference

$$D_I(Y;Y_N) = I(X;Y) - I(X;Y_N) \tag{1}$$

(note that $D_I \geq 0$). We use the functional D_I as a measure of the average distortion of the quality of a quantization. It can be interpreted as an *information distortion measure*, hence the symbol D_I. The only term that depends on the quantization is $I(X;Y_N)$ so we can reformulate the problem as the maximization of the effective functional $D_{eff} = I(X;Y_N)$.

The average distortion can be rewritten as the expectation of a pointwise distortion function of a rather interesting form. Using the definition of the mutual information and the Markov relation $X \rightarrow Y \rightarrow Y_N$ between the spaces, we can express D_I as the expectation

$$D_I = E_{p(y,y_N)} d(y, y_N) \tag{2}$$

where

$$d(y, y_N) \equiv KL(q(x|y)||q(x|y_N)) \tag{3}$$

is the Kullback-Leibler directed divergence of the input stimulus conditioned on a response y relative to the stimulus conditioned on a reproduction y_N. Intuitively, this measures the similarity between the stimulus partition induced by the quantization to the one induced by the sensory system.

2.3 Implementations

Using a quantization (deterministic or stochastic) of the output space [7] allows us to control the exponential growth of required data. With this approach we estimate a quantity which is known to be a lower bound of the actual mutual information. We obtain a biased estimate but control the precision with which

it can be estimated. We fix the coarseness of the quantization (the size of the reproduction, N) and look for a quantization that minimizes the information distortion measure $D_I = I(X;Y) - I(X;Y_N)$ described previously.

Constrained maximum entropy optimization.

The problem of optimal quantization has been formulated for a large class of distortion functions [11] as a maximum entropy problem [8]. We cannot use this analysis directly, since in our case the distortion function depends explicitly on the quantizer. The reasoning behind the maximum entropy formulation is that, among all quantizers that satisfy a given set of constraints, the maximum entropy quantizer does not implicitly introduce further restrictions in the problem. We pose the minimum distortion problem as a maximum quantization entropy problem with a distortion constraint:

$$\begin{aligned} \max_{q(y_N|y)} H(Y_N|Y) \qquad & \text{constrained by} \\ D_I(q(y_N|y)) \leq D_o \qquad & \text{and} \\ \sum_{y_N} q(y_N|y) = 1 \qquad & \forall y \in Y \end{aligned} \tag{4}$$

This is an ordinary constrained optimization problem that can be solved numerically with standard optimization tools. The cost function $H(Y_N|Y)$ is concave in $q(y_N|y)$, and the probability constraints $\sum_{y_N} q(y_N|y) = 1$ are linear in $q(y_N|y)$ [1]. The constraint D_I is also concave in $q(y_N|y)$, which make the whole problem one of concave maximization.

The problem with this formulation is that it relies on knowing D_I, which depends on the mutual information between X and Y. We can easily avoid the need for that by using the effective distortion $D_{eff} \equiv I(X;Y_N)$. In this case, the optimization problem is

$$\begin{aligned} \max_{q(y_N|y)} H(Y_N|Y) \quad & \text{constrained by} \\ D_{eff} \equiv I(q(y_N|y)) \geq I_o \qquad & \text{and} \\ \sum_{y_N} q(y_N|y) = 1 \qquad & \forall y \in Y \end{aligned} \tag{5}$$

The solution to the optimization problem (5) depends on a single parameter I_o, which can be interpreted as the informativeness of the quantization. If $I_o \leq 0$, the distortion constraint is always satisfied and we obtain only the unconstrained maximum entropy solution $q(y_N|y) = 1/N$ for all pairs (y, y_N). For $I_o \geq 0$ the distortion constraint becomes active and the uniform quantizer

is no longer a solution to the optimization problem. Because of the convexity of the problem, the optimal solution will lie on the boundary of the constraint and thus carry $I(X;Y_N) = I_o$ bits of information.

Maximum cost optimization.

A standard approach to constrained optimization problems is through the use of Lagrange multipliers. The system (4) can be solved as the unconstrained optimization of

$$\max_{q(y_N|y)} \Big(H(Y_N|Y) - \beta D_{eff}(q(y_N|y)) + \sum_y \lambda_y \sum_{y_N} q(y_N|y) \Big).$$

The solution depends on the parameters $(\beta, \{\lambda_y\})$ which can be found from the constraints

$$D_{eff}(q(y_N|y)) \geq I_o$$
$$\sum_{y_N} q(y_N|y) = 1 \ \forall y \in Y$$

Since β is a function of I_o, which is a free parameter, we can as well discard I_o and reformulate the optimization problem as finding the maximum of the cost function

$$\max_{q(y_N|y)} F(q(y_N|y)) \equiv \max_{q(y_N|y)} \Big(H(Y_N|Y) + \beta D_{eff}(q(y_N|y)) \Big) \qquad (6)$$
$$\text{constrained by}$$
$$\textstyle\sum_{y_N} q(y_N|y) = 1 \ \forall y \in Y.$$

An implicit solution for the optimal quantizer.

Further analysis of the problem uses the simplicity of the linear constraint in (6). Extrema of F can be found by setting its derivatives with respect to the quantizer $q(y_N|y)$ to zero. In the subsequent steps we shall explicitly use the assumption that all spaces are finite and discrete. The expressions are thus in a form convenient for programming on a computer. The results for continuous random variables can easily be adapted from this using analogous methods from the calculus of variations. We use Latin indices (i, j, k) to denote members in the original spaces X, Y and Greek indices (μ, ν, η) for elements of the reproduction Y_N. With this in mind we solve the Lagrange multiplier problem

$$q(y_\nu|y_k) = \frac{e^{\beta\left(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\right)}}{\sum_\nu e^{\beta\left(\frac{(\nabla D_{eff})_{\nu k}}{p(y_k)}\right)}}. \qquad (7)$$

In practice, the expression (7) can be iterated for a fixed value of β to obtain a solution for the optimization problem, starting from a particular initial state. For small β, before the first bifurcation [11], the obvious initial condition is the uniform solution $q(y_N|y) = 1/N$. The solution for one β can be used as the initial condition for a subsequent β because solutions are continuous with respect to the quantizer.

3 Neural Information Processing

We shall present a general model of a neural system which can be fully recovered by the quantization procedure described earlier. Sensory stimuli and their neural representation can be quite complex. Information theory suggests a way for dealing with this complexity by extracting the essential parts of the signal while maintaining most of its information content. The method of choice is through *typical sequences*. We can afford just a brief sketch of the model here. A more complete presentation can be found in an earlier paper [3].

3.1 Jointly Typical Sequences

When analyzing information channels, we deal with two sets of random sequences – input and output. In this case it is necessary to consider the combined behavior of the pair (X, Y). In the current formalism this is achieved by using *jointly typical sequences*: x^n and y^n are independently typical and (x^n, y^n) is also typical in the product space [1]. All the elements of the *jointly typical set* are nearly equiprobable, the set has probability close to 1, and the number of elements is about $2^{nH(X,Y)}$.

Not all pairs of typical x^n and typical y^n are also jointly typical. The probability that a randomly chosen pair is jointly typical is about $2^{-nI(X;Y)}$. Hence, for a fixed y^n, we can consider about $2^{nI(X;Y)}$ such pairs before we are likely to come across a jointly typical pair. This suggests there are about $2^{nI(X;Y)}$ distinguishable messages in X^n that can be communicated through Y^n (figure 1).

3.2 Decoding with jointly typical sequences

The jointly typical pairs (x^n, y^n) can be used as *codewords*. Since there are $2^{nI(X;Y)}$ distinguishable signals and $2^{nH(X,Y)}$ codewords, some of the codewords represent the same signals. The redundancy helps to combat noise or is due to projections to a lower dimensional subspace. Sets of codewords representing the same signal form equivalence classes, which we call *codeword classes*. We shall equate the distinguishable signals with the codeword classes. Within each class, a stimulus in X^n invokes a corresponding jointly typical response in Y^n with high probability (about $1 - 2^{-nI(X;Y)}$).

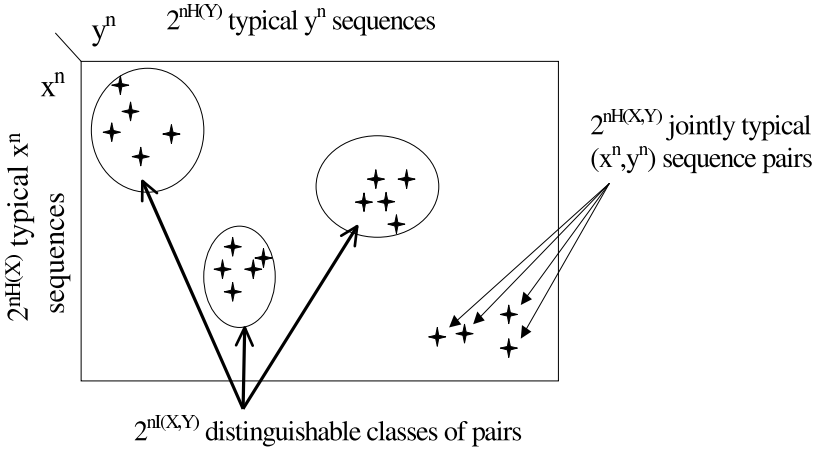

Figure 1: The structure of the jointly typical set. There are about $2^{nH(X)}$ typical x sequences, $2^{nH(Y)}$ typical y sequences but only $2^{nH(X,Y)}$ jointly typical sequences. This suggests there are about $2^{nI(X;Y)}$ distinguishable *equivalence classes* C_i of (x^n, y^n) pairs. The number of output sequences in each class is about $|C_i| \approx 2^{nH(Y|X)}$.

We define the *codebook* of this system as the map $\mathcal{F} : x^n \to y^n$. The codebook is stochastic on individual elements, so it is better represented through association probabilities $q(y^n|x^n)$. When considered on codeword classes though, the map is *almost bijective*, that is, with probability close to 1 elements of Y^n are assigned to an elements of X^n in the same codeword class. We shall decode an output y^n as (any of) the inputs that belong to the same codeword class. Similarly, we shall consider the representation of an input x^n to be any of the outputs in the same codeword class.

3.3 Resolving the decoding problem

The minimal information distortion quantization can help us resolve the neural decoding problem. We quantize the neural response by fixing the size of the reproduction to N. This bounds our estimate of D_{eff} to be no more than $\log N$ bits. In the ideal case, $\max D_{eff} \equiv \max I(X; Y_N) \approx \log N$ but in general it will be lower. Due to the Markov relation, $I(X; Y_N) \leq I(X; Y)$ as well. Since $\log N$ increases with N and $I(X; Y)$ is a constant, these two independent bounds intersect for some $N = N_c$ at which point adding more elements to Y_N does not improve the distortion measure. If $I(X, Y_N)$ increases with N until $N = N_c$ and then levels off, we can identify the correct N_c by looking at the behavior of the expected distortion (or, equivalently, $D_{eff} \equiv I(X; Y_N)$) as a function of N, given sufficient data. The elements of Y_N are then representa-

tives of the original equivalence classes which we wanted to find. The quantizer $q(y_N|y)$ gives the probability of a response y belonging to an equivalence class y_N. As mentioned in [11], the optimal quantizer for low distortions (high β) is deterministic (or effectively deterministic, in case of duplicate classes) and so we recover an almost complete reproduction of the model.

If there is not enough data to support a complete recovery, the algorithm has to stop earlier. The criterion we use in such a case is that the estimate of D_{eff} does not change with N *within its error bounds* (obtained analytically or by statistical re-estimation methods like bootstrap, or jack-knife). Then $N < N_c$ and the quantized mutual information is at most $\log N$. We can recover at most N classes and some of the original classes will be combined. The quantizer may also not be deterministic due to lack of enough data to resolve uncertainties. Thus we can recover a somewhat impoverished picture of the actual input/output relationship which can be refined as more data becomes available.

4 Results

We shall discuss the application of the method described so far to a few test cases of synthetic data. Applying it to physiological data from a sensory system involves additional difficulties associated with the estimates of D_I for complex input stimuli, which are dealt with elsewhere [4,5].

4.1 Random Clusters

We present the analysis of data drawn from the probability shown in figure 2a. This model was chosen to resemble the picture of decoding with jointly typical sequences (figure 1). The mutual information between the two sequences is about 1.8 bits, which is comparable to the mutual information of single neurons in the cricket cercal system [2]. In this case we assume the original relation between X and Y is known (the joint probability $p(x, y)$ is used explicitly).

The results can be seen in figure 2b. The algorithm recovers an incomplete representation when two classes are forced (b.1). This is improved for the 3 class version (b.2). The next refinement (b.3) separates all the classes correctly and recovers most of the mutual information. Further refinements (b.4) fail to split the classes and are effectively identical to b.3 (note that classes 4 and 5 in b.4 are almost evenly populated and the class membership there is close to a uniform 1/2). The quantized mutual information (c) increases with the number of classes until it recovers about 90% of the original mutual information ($N = 4$) at which point it levels off.

The behavior of D_{eff} as a function of the annealing parameter β is shown in figure 3. One can observe the bifurcations of the optimal solution (1 through 5) and the corresponding transitions of the effective distortion. The abrupt

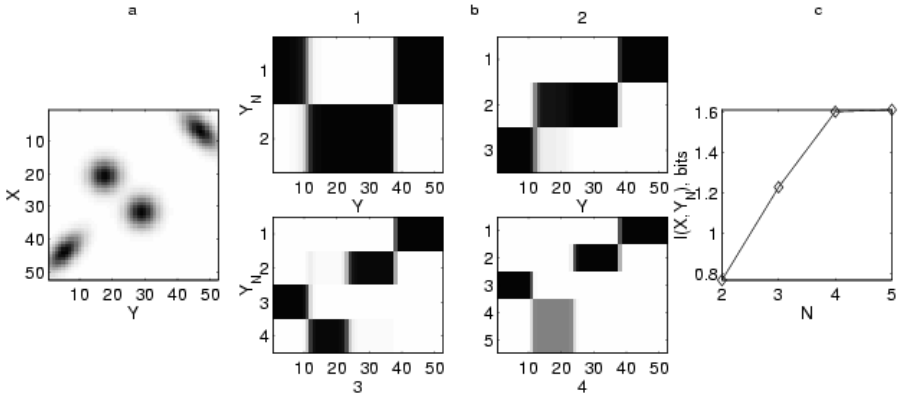

Figure 2: A joint probability for the relation between two random variables X and Y with 52 elements each (a) with optimal quantizers $q(y_N|y)$ (b) for different number of classes. The behavior of the mutual information with increasing N can be seen in (c).

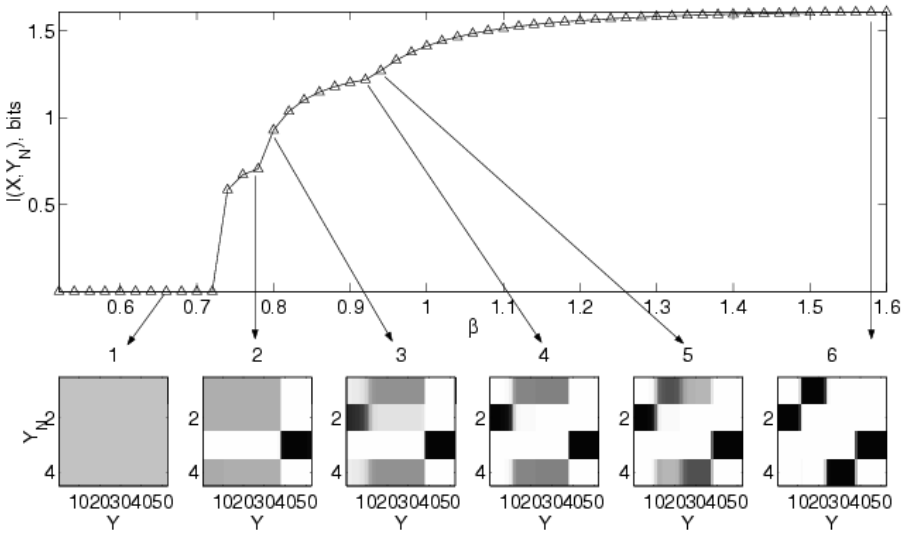

Figure 3: Behavior of D_{eff} (top) and the optimal quantizer $q(y_N|y)$ (bottom) as a function of the annealing parameter β.

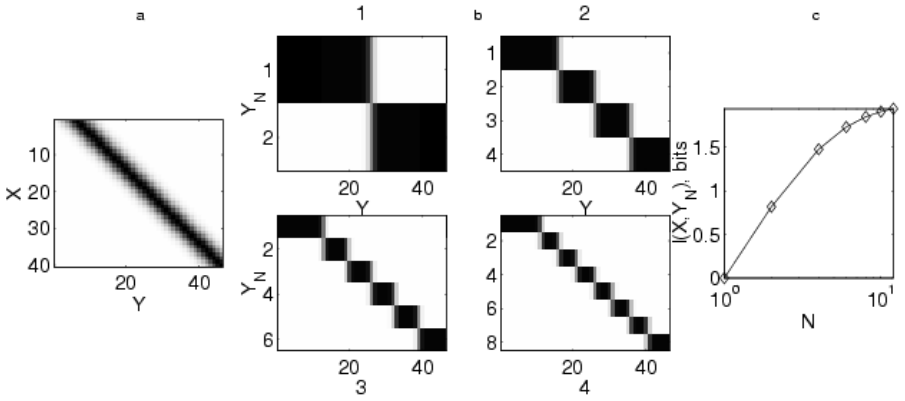

Figure 4: A joint probability for a linear relation between two random variables X and Y (a) with optimal quantization (b) for different number of classes. The behavior of the mutual information with increasing N can be seen in (c).

transitions ($1 \to 2, 2 \to 3$) are similar to the ones described in[11] for an arbitrary distortion function. We also observe transitions ($4 \to 5$) which appear to be smooth in D_{eff} even though the solution for the optimal quantizer undergoes a bifurcation.

A random permutation of the rows and columns of the joint probability in figure 2a has the same channel structure. The quantization is identical to the case presented in figure 2 after applying the inverse permutation and fully recovers the permuted classes (the quantization is contravariant with respect to the action of the permutation group).

4.2 Linear encoding

We also applied the algorithm to a case which, unlike the previous cases, does not have clearly defined clusters. This model tries to simulate the process of a physical measurement where X is the physical system and Y is the measurement. In this example we model a linear relation between X and Y and Gaussian measurement noise, that is

$$Y = kX + \eta$$

where $\eta \in \mathcal{N}(0, \sigma)$ is drawn from a normal distribution with zero mean and variance σ^2. The particular relation we used (figure 4a) contains about 2 bits of mutual information.

The results can be seen in figure 4. The algorithm recovers a series of representations (b.1 – b.4), where each is a refinement of the previous one. The reproduction classes were permuted to roughly follow the original linear

relation. There isn't a natural stopping point and so the quantized mutual information $I(X;Y_N)$ approaches a constant. This is in contrast to the previous two cases where $I(X;Y_N)$ abruptly stopped changing after some N.

5 Conclusions

We presented a method for recovering the structure of a neural coding scheme from observations. We choose to recover an impoverished description of the coding scheme by quantizing the responses to a reproduction set of a few variables. To assess the quality of the reproduction, we defined the information distortion $D_I = I(X;Y) - I(X;Y_N)$ which measures how much information is lost in the quantization process. For a fixed reproduction size N we pose the optimization problem of finding the quantization with smallest distortion, as the one which preserves most of the information present in the original relation between X and Y. Refining the reproduction by increasing N was shown to decrease the distortion. We showed empirically on a set of synthetic problems that, if the original relation contains almost disjoint clusters, a sufficiently fine optimal quantization recovers them completely. If the quantization is too coarse, then some of the clusters will be combined, but in such a way that a large fraction of the original information is still preserved.

It is interesting to note that, although we had neural systems in mind while developing the information distortion method, the ensuing analysis is in no way limited to nervous systems. Indeed, the constraints on the two signals we analyze are so general that they can represent *almost any* pair of interacting physical systems. In this case, finding a minimal information distortion reproduction allows us to recover certain aspects of the interaction between the two physical systems, which may improve considerably any subsequent analysis performed on them. It is also possible to analyze parts of the structure of a single physical system Y, if X is a system with known properties (e.g., a signal generator, controlled by a researcher) and is used to perturb Y. These cases point to the exciting possibility of obtaining a more automated approach for succinct descriptions of arbitrary physical systems through the use of minimal information distortion quantizers.

Acknowledgments

This research was supported in part by NIH grants MH12159 (AGD) and MH57179 (JPM). The authors are indebted to Penio Penev (Rockefeller U), Tomáš Gedeon, Zane Aldworth and Kay Kirkpatrick (Montana State U) for numerous discussions which helped shape this paper.

References

1. T. Cover and J. Thomas. *Elements of Information Theory.* Wiley Series in Communication, New York, 1991.
2. A. G. Dimitrov and J. P. Miller. Natural time scales for neural encoding. *Neurocomputing*, 32-33:1027–1034, 2000.
3. A. G. Dimitrov and J. P. Miller. Neural coding and decoding: joint typicality and quantization. *Network: Computation in Neural Systems*, 2000. *submitted.*
4. A. G. Dimitrov, J. P. Miller, and Z. Aldworth. The fine structure of neural codes. In *preparation*, 2000.
5. A. G. Dimitrov, J. P. Miller, and Z. Aldworth. Non-uniform quantization of neural spike sequences through an information distortion measure. In J. Bower, editor, *Computational Neuroscience: Trends in Research.* 2000 *(to appear).*
6. A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.
7. R. M. Gray. *Entropy and Information Theory.* Springer-Verlag, 1990.
8. E. T. Jaynes. On the rationale of maximum-entropy methods. *Proc. IEEE*, 70:939–952, 1982.
9. D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of the neural code. *J. Comp. Neurosci.*, 2000. *under review.*
10. F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the neural code.* The MIT Press, 1997.
11. K. Rose. Deteministic annealing for clustering, compression, classification, regerssion, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.
12. C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:623–656, 1948.