Information Distortion and Neural Coding

Tomáš Gedeon† Albert E. Parker†
Alexander G. Dimitrov‡*
†Department of Mathematical Sciences and
‡Center for Computational Biology
Montana State University
Bozeman MT 59717

October 12, 2001

Abstract

Our main interest is the question of how neural ensemble activity represents sensory stimuli. In this paper we discuss a new approach to characterizing neural coding schemes. It attempts to describe the specific stimulus parameters encoded in the neural ensemble activity and at the same time determines the nature of the neural symbols with which that information is encoded. This recently developed approach for the analysis of neural coding [7, 9] minimizes an intrinsic information-theoretic cost function (the information distortion) to produce a simple approximation of a coding scheme, which can be refined as more data becomes available. We study this optimization problem. The admissible region is a direct product of simplices. We show that the optimal solution always occurs at a vertex of the admissible region. This allows us to reformulate the problem as a maximization problem on the set of vertices and develop an algorithm, which, under mild conditions, always finds a local extremum. We compare the performance of this new algorithm to standard optimization schemes on synthetic cases and on physiological recordings from the cricket cercal sensory system.

1 Introduction

One of the steps toward understanding the neural basis of an animal's behavior is characterizing the code with which its nervous system represents information. All computations underlying an animal's behavioral decisions are carried out within the context of this code.

Deciphering the neural code of a sensory system means determining the correspondence between neural activity patterns and sensory stimuli. This task can be reduced further to three related problems: determining the specific stimulus parameters encoded in the neural ensemble activity, determining the nature of the neural symbols with which that information

*This research partially supported by NSF-DGE grant 9972824 (AEP), NIH grant MH12159 (AGD) and by NSF Grant MRI 9871191 and NIH Grant R01 MH57179 to John P. Miller.

is encoded, and finally, quantifying the correspondence between these stimulus parameters and neural symbols. If we model the coding problem as a correspondence between the elements of an input set X and an output set Y, these three tasks are: finding the spaces X and Y and the correspondence between them.

Common approaches to this problem include stimulus reconstruction [27] and the use of impoverished stimulus sets to characterize stimulus/response properties [15]. However, these methods often introduce multiple assumptions that may affect the character of the obtained solution. Some of these approaches start with an assumption about the relevant structures of the space Y (e.g., a single spike in the first-order stimulus reconstruction method, or the mean spike rate over a defined interval [29]) and proceed by calculating the expected stimulus features that are correlated with these codewords. Other approaches make an assumption about the relevant stimulus features (the space X), such as moving bars and gratings when investigating parts of the visual cortex, and proceed to study the patterns of spikes that follow the presentation of these features.

Our goal in this paper is to present and extend a new analytical approach that minimizes the number of assumptions we impose on the input and output spaces. The goal of this approach is to allow a quantitative determination of the type of information encoded in neural activity patterns and, at the same time, identify the code with which this information is represented. Any neural code must satisfy at least two conflicting demands. On the one hand, the organism must recognize the same natural object as identical in repeated exposures. On this level the response of the organism needs to be *deterministic*. On the other hand, the neural code must deal with uncertainty introduced by both external and internal noise sources. Therefore the neural responses are by necessity *stochastic* on a fine scale. In this respect the functional issues that confront the early stages of any biological sensory system are similar to the issues encountered by communication engineers in their work of transmitting messages across noisy media. In this paper we show how tools from information theory can be used to characterize the neural coding scheme of a simple sensory system.

We model the input/output relationship present in a biological sensory system as an *optimal information channel* [30], see Figure 1A. Although this model is stochastic, in this context a coding scheme consists of classes of stimulus/response pairs which form a structure akin to a dictionary: each class consists of a stimulus set and a response set, which are synonymous. The classes themselves are almost independent, with few intersecting members. The number of distinguishable classes is related to the mutual information between stimulus and response [6, 30].

We look for high quality approximations of such a coding scheme. To do this, we quantize the neural responses to a small reproduction set. This quantization is optimized to minimize an information-based distortion function. Fixing the size of the reproduction produces an approximation of the coding scheme described above. The approximation can be refined by increasing the size of the reproduction. For the model described above, there is a critical size, beyond which further refinements do not significantly decrease the distortion. Given sufficient data, we choose the optimal quantization at this size to represent the coding scheme. If the amount of data is not enough, we stop at a coarser reproduction, determined by the uncertainty to our estimate of the cost function.

The paper is organized as follows. In section 2 we describe the basic tools from informa-

tion theory, derive an information distortion function and formulate the coding problem as an optimization problem of finding a quantizer which minimizes the information distortion function.

In section 3 we present three algorithms to solve this optimization problem. Two of them use an annealing approach. The third one is based on a reformulation of the problem as a combinatorial search over the set of vertices of the admissible region. The proof of validity of this reformulation, as well as the proof of the convergence of the algorithm to a local optimum, are postponed to section 5.

In section 4 we apply the algorithms to synthetic data to asses their performance, and then to physiological recordings of the cricket cercal sensory system. We discuss their relative merits and problems.

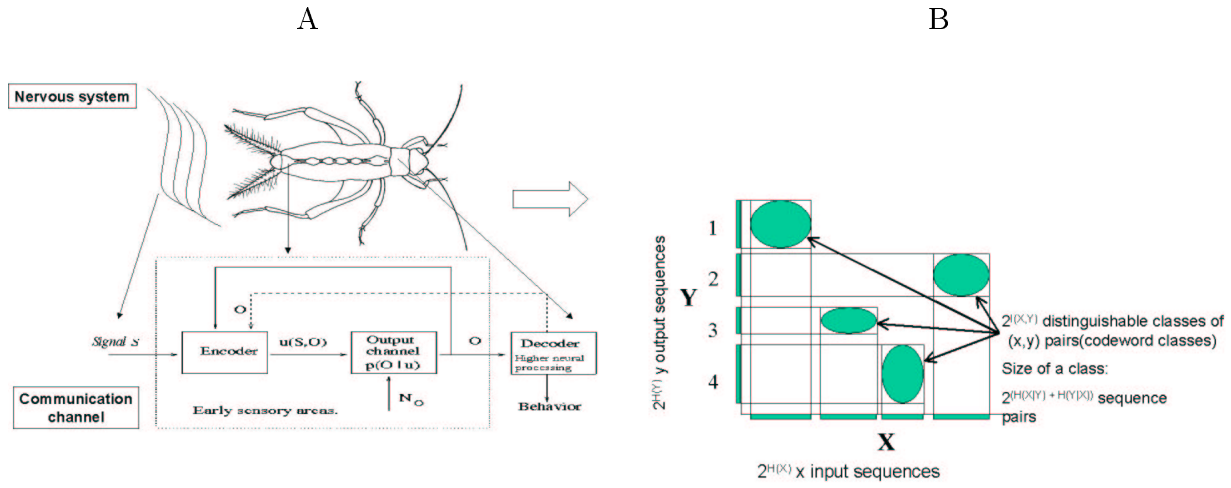

Figure 1: (A) The cricket cercal sensory system modeled as a communication channel. (B) The structure of a communication system. There are about $2^{H(X)}$ stimulus (x) sequences, $2^{H(Y)}$ response (y) sequences but only about $2^{I(X;Y)}$ distinguishable equivalence classes y_N of (x, y) pairs.

2 Introduction to Information Theory

2.1 Basic Concepts

The basic object in information theory is an *information source* or a random variable X. A source X is a mathematical model for a physical system that produces a succession of symbols $\{x_1, x_2, \ldots, x_n\}$ in a manner which is unknown to us and is treated as random [6, 14].

The basic concepts of information theory are *entropy* and *mutual information*. In information theory, entropy is described as a measure of the uncertainty, or of the self information, of a random variable, and is defined as

$$H = -E_x \log p(x).$$

Next we define the *conditional* and *joint* entropy respectively as

$$\begin{aligned} H(Y|X) &= -E_{x,y} \log p(y|x) \\ H(X,Y) &= -E_{x,y} \log p(x,y). \end{aligned}$$

The notion of *mutual information* $I(X;Y)$ is introduced as a measure of the degree of dependence between a pair of random variables (X,Y):

$$I(X;Y) = \log E_{x,y} \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

Both entropy and mutual information are special cases of a more general quantity – the *Kullback-Leibler directed divergence* or *relative entropy* [21] between two probability measures on the same event space:

$$KL(p\|q) = E_p \log \left(\frac{p(x)}{q(x)} \right). \tag{2}$$

The Kullback-Leibler divergence is always nonnegative and it is zero if and only if $p(x) = q(x)$ a.e. However, it is not symmetric and so it is not a proper distance on a set of probability measures. In spite of this it provides a sense of how different two probability measures are.

The information quantities H, I and KL depend only on the underlying probability distributions and not on the structure of X and Y. This allows us to evaluate them in cases where more traditional statistical measures (e.g. variance, correlation, etc.) do not exist.

Why are entropy and mutual information valid measures to use when analyzing an information channel between X and Y? Let $\{y_1, y_2, \ldots, y_n\}$ be i.i.d. observations from an information source Y. Then the Strong Law of Large Numbers provides theoretical justification for making inference about population parameters (e.g. response parameters) from data collected experimentally. In particular, the Shannon Entropy Theorem [6] in this case assures that the entropy (and hence the mutual information) calculated from data taken experimentally converges to the true population entropy as the amount of data available increases. In the case of physiological recordings from a biological sensory system, $\{y_1, y_2, \ldots, y_n\}$ are not usually i.i.d.. For example, in the data that we present in this paper, we take a single, "long" recording of a neural response and break it up into observations of length 10 ms. Inference made about population parameters from data collected this way is justified if we can assume that Y is ergodic. Now we may appeal to the Ergodic Theorem [4] and the Shannon-McMillan-Breiman Theorem [6] to justify the use of our information theoretic quantities.

Each element of the output space Y can be modeled as a sequence of zeroes and ones, where 1 indicates the presence and 0 indicates the absence of a spike in a particular time interval. Hence $Y = Z^k$ (the *k-th* extension of Z) can be thought of as the set of all sequences of length k of symbols from $Z := \{0,1\}$. There is a limited number of distinct messages which can be transmitted with sequences of length k from the source Z. These are the *typical sequences* of Z [6].

Definition: The *typical set* A_ϵ^k with respect to probability density $p(z)$ on Z is the set of sequences $(z_1, z_2, \ldots, z_k) \in Z^k$ for which

$$2^{-k(H(Z)+\epsilon)} \leq p(z_1, z_2, \ldots, z_k) \leq 2^{-k(H(Z)-\epsilon)}.$$

The typical set has the following properties:

Theorem 1 (Properties of typical sequences.) *If Z is ergodic, then*

1. *$Pr\{A_\epsilon^k\} > 1 - \epsilon$ for k sufficiently large*
2. *$(1-\epsilon)2^{k(H(Z)-\epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(Z)+\epsilon)}$ for k sufficiently large. Here $|A|$ is the number of elements in set A.*

When considering information channels, we deal with two sources, W and Z. We consider the behavior of the pair (W, Z).
Definition: The set A_ϵ^k of *jointly typical* sequences $\{(w^k, z^k)\}$ with respect to the joint distribution $p(w, z)$ on $W \times Z$ is the set

$$\begin{aligned} A_\epsilon^k \;=\; & \left\{(w^k, z^k) \in W^k \times Z^k : \right. \\ & 2^{-k(H(W)+\epsilon)} \leq p(w^k) \leq 2^{-k(H(W)-\epsilon)}, \\ & 2^{-k(H(Z)+\epsilon)} \leq p(z^k) \leq 2^{-k(H(Z)-\epsilon)}, \\ & \left. 2^{-k(H(W,Z)+\epsilon)} \leq p(w^k, z^k) \leq 2^{-k(H(W,Z)-\epsilon)} \right\}, \end{aligned}$$

Theorem 2 (Properties of jointly typical sequences.) *Let (W^k, Z^k) be a pair of ergodic sources. Then*

1. *$Pr(A_\epsilon^k) > 1 - \epsilon$.*
2. *$(1-\epsilon)2^{k(H(W,Z)-\epsilon)} \leq |A_\epsilon^k| \leq 2^{k(H(W,Z)+\epsilon)}$ for n sufficiently large.*
3. *If $(\tilde{W}^k, \tilde{Z}^k)$ are a pair of random variables with joint probability $p(w^k, z^k) = p(w^k)p(z^k)$ (i.e. $\tilde{W}^k$ and $\tilde{Z}^k$ are independent with the same marginal distributions as W^k and Z^k), then for sufficiently large k,*

$$(1-\epsilon)2^{-k(I(W;Z)+3\epsilon)} \leq Pr\left((\tilde{W}^k, \tilde{Z}^k) \in A_\epsilon^k\right) \leq 2^{-k(I(W;Z)-3\epsilon)}.$$

Property 3 suggests that the set of jointly typical sequences can be divided into $2^{kI(W,Z)}$ disjoint sets, such that projections of these sets to W^k as well as to Z^k are almost disjoint. This justifies figure 1.B for spaces $X = W^k$ and $Y = Z^k$. A complete proof of these statements can be found in [6].

A random variable Y can be related to another random variable Y_N through the process of *quantization* (lossy compression) [6, 14]. Y_N is referred to as the *reproduction* of Y. The process is defined by a map q from the probability space Y to Y_N, called a *quantizer*. In general, quantizers can be stochastic: q assigns to $y \in Y$ the probability that the response y belongs to an abstract class y_N. A deterministic quantizer is a special case in which q takes the values of 0 or 1 only. It can be shown [14] that the mutual information $I(X;Y)$ is the least upper bound of $I(X;Y_N)$ over all possible reproductions Y_N of Y. Hence, the original mutual information can be approximated with arbitrary precision using carefully chosen reproduction spaces.

2.2 Neural systems as information channels

Communication channels characterize a relation between two random variables: an input X and an output Y. When mapping this structure to neural systems, the output space is usually the set of activities of a group of neurons. The input space can be sensory stimuli from the environment or the set of activities of another group of neurons. We would like to recover the correspondence between stimuli and responses, which we call a *coding scheme* [32].

The early stages of neural sensory processing encode information about sensory stimuli into a representation that is common to the whole nervous system. We will consider this encoding process within a probabilistic framework [1, 20, 27]: *The input signal* X is produced by a source with a probability $p(x)$. This may be a sensory stimulus or the activity of a set of neurons. *The output signal* Y is produced with probability $p(y)$. This is the temporal pattern of activity across a set of cells. *The encoder* $p(y|x)$ is a stochastic mapping from X to Y. From the point of view of information theory, the designation of spaces X and Y as an input and output space is arbitrary. Thus we can choose to characterize the same information channel as a source Y with probability $p(y)$ and a *decoder* stochastic mapping $p(x|y)$ from Y to X. In the context of a sensory system, one can measure the set of responses Y quite easily, as it consists of various spike patterns. One can then analyze the estimate of the stimulus $\hat{p}(x) = \sum_y p(x|y)p(y)$. This methodology was used in [27] under the name of *stimulus reconstruction*.

Assume now that $X = W^k$ and $Y = Z^k$ for some sources W and Z. We can identify a candidate space Z as a space consisting of the symbols 1 (for a spike) and 0, standing for no spike in a particular window of time. However, usually we do not understand too well the set of stimuli to which a physiological preparation is attuned to. Therefore we do not know the input space W too well. We approach this problem by using statistical modeling.

Although the information channel model is stochastic, an almost deterministic relation emerges naturally on the level of clusters of stimulus/response pairs. As we have mentioned in the preliminaries, the jointly typical sequences in $(X, Y) = (W^k, Z^k)$ form an almost bijective relation, see Figure 1B. We call each jointly typical class a codeword class. We see that with probability close to 1 elements of Y are assigned to elements of X in the same codeword class. We shall decode an output y as (any of) the inputs that belong to the same codeword class. Similarly, we shall consider the representation of an input x to be any of the outputs in the same codeword class.

2.3 Recovering a neural coding scheme

We now discuss a recently developed approach to finding a neural coding scheme through quantization of the neural response Y into a coarser representation in a smaller event space Y_N [9]. An important reason for using quantization for this purpose is the goal of using available data in the most efficient way. As pointed out in [17], the amount of data needed to support non-parametric estimates of coding schemes which contain long sequences of length T across N neurons grows exponentially with T and N. For some systems the required data recording time may well exceed the expected lifespan of the system. To resolve this issue we choose to sacrifice some detail in the description of the coding scheme in order to obtain

robust estimates of a coarser description.

A quantization [6, 14] in this context is a stochastic map $q(y_N|y)$ of the neural representation Y into a coarser representation in a smaller event space Y_N. The random variables $X \to Y \to Y_N$ form a Markov chain. We characterize the quality of a quantization by a distortion function [6] and look for a minimum distortion quantization. The resulting relation between stimulus and reproduction, $q(y_N|x) = \sum_y q(y_N|y)p(y|x)$, will be an approximation of the neural coding scheme. By increasing the size of the reproduction, N, we can refine the approximation as much as the available data will allow.

2.3.1 The distortion function

A quantization $q(y_N|y)$ produces a new random variable (a reproduction space) Y_N with associated probabilities $p(y_N)$. At the same time, in our case a quantization induces probabilities $p(x|y_N)$ which allow us to obtain a reconstruction of the input $\hat{p}(x) = \sum_N p(x|y_N)p(y_N)$ related to the quantized observations $p(y_N)$. We view the distribution $p(x|y_N)$ as an approximation of the neural decoder $p(x|y)$. We require that this approximation is the best possible under the constraint that the number of classes N is fixed. The quality of a quantization is characterized by a distortion function [6]. In engineering applications, the distortion function $D(\cdot,\cdot)$ is usually chosen in a fairly arbitrary fashion [6, 13], typically the Euclidean squared distance [28]. We want to avoid this arbitrariness. The natural measure of closeness between two distributions is the Kullback-Leibler divergence KL. For each fixed $y \in Y$ and $y_N \in Y_N$, $p(x|y)$ and $p(x|y_N)$ are a pair of distributions on the space X. As a *pointwise distortion function* we take $d(y, y_N) = KL(p(x|y_N)||p(x|y))$. Unlike the pointwise distortion functions usually investigated in information theory [6, 28], this one depends on the quantizer $q(y_N|y)$ through $p(x|y_N)$. We define our *distortion function* as the expected Kullback-Leibler divergence over all pairs (y, y_N)

$$D_I(Y, Y_N) = D_I(q(y_N|y)) := E_{y,y_N} KL(p(x|y_N)||p(x|y)).$$

We derive an alternate expression for D_I. Starting from the definition

$$\begin{aligned}
D_I &= \sum_{y,y_N} p(y, y_N)\ KL\big(p(x|y)\|p(x|y_N)\big) \\
&= \sum_{y,y_N} p(y, y_N) \sum_x p(x|y) \log \frac{p(x|y)}{p(x|y_N)} \\
&= \sum_{x,y,y_N} p(x, y, y_N) \Big(\log p(x|y) - \log p(x|y_N)\Big) \qquad (3) \\
&= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} - \sum_{x,y_N} p(x, y_N) \log \frac{p(x, y_N)}{p(x)p(y_N)} \qquad (4) \\
&= I(X;Y) - I(X;Y_N)
\end{aligned}$$

In (3) we used the Markov property $p(x, y, y_N) = p(x|y)p(y, y_N)$ and (4) is justified by using the identities $p(x, y) = \sum_{y_N} p(x, y, y_N)$, $p(x, y_N) = \sum_y p(x, y, y_N)$ and the Bayes property $p(x, y)/p(y) = p(x|y)$. This shows that the information distortion can be written as

$$D_I = I(X;Y) - I(X;Y_N).$$

This function can be interpreted as an *information distortion measure*, hence the symbol D_I. The only term in D_I that depends on the quantization is $I(X;Y_N)$, so we can replace D_I with the effective distortion

$$D_{eff} := I(X;Y_N)$$

in our optimization schemes. Our goal is to find a quantization $q(y_N|y)$ that minimizes the information distortion measure D_I or, equivalently, maximize D_{eff}, for a fixed reproduction size N.

2.3.2 Finding the codebook

Following examples from rate distortion theory [6, 28], the problem of optimal quantization was formulated [9] as a maximum entropy problem [16]. The reason is that, among all quantizers that satisfy a given set of constraints, the maximum entropy quantizer does not implicitly introduce additional constraints in the problem. In this framework, the minimum distortion problem is posed as a maximum quantization entropy problem with a distortion constraint:

$$\begin{aligned}
&\max_{q(y_N|y)} H(Y_N|Y) \quad \text{constrained by} \\
&D_I(q(y_N|y)) \le D_0 \quad \text{and} \\
&\sum_{y_N} q(y_N|y) = 1 \quad \text{and} \quad q(y_N|y) \ge 0 \quad \forall y \in Y.
\end{aligned}$$

Since the only part of D_I which depends on the quantizer is $D_{eff} = I(X;Y_N)$, this is equivalent to

$$\begin{aligned}
&\max_{q(y_N|y)} H(Y_N|Y) \quad \text{constrained by} \\
&D_{eff}(q(y_N|y)) \ge I_0 \quad \text{and} \\
&\sum_{y_N} q(y_N|y) = 1 \quad \text{and} \quad q(y_N|y) \ge 0 \quad \forall y \in Y.
\end{aligned} \tag{5}$$

The goal is to find the maximal entropy solution for a maximal possible value of D_{eff}.

The conditional entropy $H(Y_N|Y)$ and the function D_{eff}, can be written explicitly in terms of $q(y_N|y)$

$$\begin{aligned}
H(Y_N \mid Y) &= E_{y,y_N} \log q(y_N|y) \\
&= \sum_{y,y_N} p(y) q(y_N|y) \log \left(q(y_N|y) \right)
\end{aligned}$$

and

$$\begin{aligned}
D_{eff} = I(X;Y_N) &= \log E_{x,y_N} \frac{p(x,y_N)}{p(x)p(y_N)} \\
&= \sum_{x,y,y_N} q(y_N|y) p(x,y) \log \left(\frac{\sum_y q(y_N|y) p(x,y)}{p(x) \sum_y p(y) q(y_N|y)} \right).
\end{aligned} \tag{6}$$

The optimal quantizer $q^*(y_N|y)$ induces a coding scheme from $X \to Y_N$, $p^*(y_N|x) = \sum_y q^*(y_N|y)p(y|x)$, which is the most informative approximation of the original relation $p(x|y)$ for a fixed size N of the reproduction Y_N. Increasing N produces a refinement of the approximation, which is more informative (has lower distortion) and thus preserves more of the original mutual information $I(X;Y)$. Quantizing to a reproduction of fixed size N bounds the estimate of D_{eff} to be no more than $\log_2 N$ bits. In the ideal case, $\max D_{eff} \equiv \max I(X;Y_N) \approx \log_2 N$ but in general it will be lower. $I(X;Y_N)$ is also bounded from above by $I(X;Y_N) \leq I(X;Y)$. Since $\log_2 N$ 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. Since we in general don't know $I(X;Y)$, we empirically choose N_c at which the rate of change of D_I with N sharply decreases. If there is not enough data to support so fine a quantization, the algorithm has to stop earlier. The criterion we use in such cases is that the estimate of D_{eff} does not change with N within its error bounds (obtained analytically or by statistical re-sampling methods like bootstrap, or jack-knife). Then $N < N_c$ and the value of D_{eff} is at most $\log_2 N$. We can recover at most N classes and some distinct classes will be combined. Thus this method allows us to study coarse but highly informative models of a coding scheme, to automatically refine them when more data becomes available, and to identify a natural reproduction size N_c.

3 Optimization schemes

In this section, we investigate and compare three different approaches to solving the optimization problem (5). Two of them use a reformulation of (5) which solve the system by starting in the interior of the feasible region and using the method of *annealing* to find extrema. Technically they are related to the problem of lossy compression with minimal distortion, arising in Rate Distortion Theory [6, 13]. The third method is based on the observation (Theorem 4) that an optimal solution of (5) lies generically at a vertex of the feasible region. As a consequence of this fact, in Theorem 5 we formulate an equivalent problem to (5) and an algorithm, called *vertex search* (9), to solve it. This algorithm finds an optimal solution of (5) under mild conditions (Theorem 19).

When searching for the extrema of a general optimization problem, there is no known theory indicating whether using continuous, gradient-type algorithms is cheaper than searching over a finite, large set which contains the extrema. We compare these methods in section 4 on synthetic data.

Before describing the algorithms we define the feasible region Δ, determined by the linear constraints of the optimization problem (5)

$$\Delta := \{q(y_N|y) \mid \sum_{y_N} q(y_N|y) = 1 \ \forall y \in Y \quad \text{and} \quad q(y_N|y) \geq 0\}.$$

Observe that Δ is a product of simplices $\Delta := \Pi_y \Delta_y$, where

$$\Delta_y := \{q(y_N|y) \mid \sum_{N} q(y_N|y) = 1\}.$$

In all subsequent discussion we assume that the random variable Y is discrete and $|Y| = s$.

3.1 Annealing

Using the method of Lagrange multipliers we can reformulate the optimization problem (5) as finding the maximum of the cost function

$$\begin{aligned} \max_{q(y_N|y)} F(q(y_N|y)) \quad &\equiv \quad \max_{q(y_N|y)} \Big(H(Y_N|Y) + \beta D_{eff}(q(y_N|y)) \Big) \\ \text{constrained by} \quad & \quad q(y_N|y) \in \Delta. \end{aligned} \tag{7}$$

This construction removes the nonlinear constraint from the problem and replaces it with a parametric search in $\beta = \beta(I_0)$. We now compare the two formulations. We start with an observation that D_{eff}, as a continuous function on compact domain Δ, has a maximal value I^*. Therefore, for values of the parameter $I_0 > I^*$ problem (5) has no solution. On the other hand, problem (7) has a solution for all values of β, since F is a continuous function on compact set Δ. We have the following result

Lemma 3 *Let q^* be a solution of (5) with $I_0 = I^*$. Let $q(\beta)$ be a solution of problem (7) as a function of the annealing parameter β. Then*

$$lim_{\beta \to \infty} D_{eff}(q(\beta)) \to I^*.$$

Proof. As $\beta \to \infty$ the solution $q(\beta)$ converges to the solution of the problem

$$max\ D_{eff}.$$

The maximum of D_{eff} is I^*. □

Since for $\beta = 0$ the optimal solution is the uniform solution $q(y_N|y) = 1/N$ [28], we need to track the optimal solution from $\beta = 0$ to $\beta = \infty$. We increment β in small steps and use the optimal solution at one value of β as the initial condition for a subsequent β. To do this we must solve (7) at a fixed value of β. We have implemented two algorithms to solve this problem: an Augmented Lagrangian algorithm and an implicit solution algorithm.

3.1.1 Augmented Lagrangian

The Augmented Lagrangian algorithm is similar to other penalty methods in that the constraints to the problem are subtracted from F to create a new cost function to maximize

$$P(q, \mu) := F(q) - \frac{1}{2\mu} \sum_y (c_y(q))^2$$

where $c_y(q) := 1 - \sum_{y_N} q(y_N|y)$, is the constraint imposed for every $y \in Y$. The more infeasible the constraints $c_y(q)$ (when $1 - \sum_{y_N} q(y_N|y) >> 0$), the harsher the penalty in P.

The Augmented Lagrangian, however, avoids the ill-conditioning of other penalty methods (as $\mu \to \infty$) by introducing explicit approximations of the Lagrange multipliers into the cost function at each optimization iteration. These approximations are constructed in such a way so that the solution to this algorithm satisfies the Karush-Kuhn-Tucker (KKT) conditions [25].

We use the Augmented Lagrangian, constructed specifically to deal with the equality constraints in (5)

$$\mathcal{L}_A(q, \lambda, \mu) = \mathbf{F}(q) - \sum_y \lambda_y c_y(q) - \frac{1}{2\mu} \sum_y c_y(q)^2$$

and use a projected line search at each Augmented Lagrangian iteration to deal with the constraint $q(y_N|y) \geq 0$.

A Newton Conjugate Gradient method [25] is used to efficiently find a search direction for each linesearch. Once the active sets are identified, the theory assures us that this algorithm procures a stationary point (where $\nabla_q F = 0$) [19].

3.1.2 Implicit solution algorithm

This algorithm is based on the observation that extrema of F can be found by setting its derivatives with respect to the quantizer $q(y_N|y)$ to zero [9]. Solving this system produces the implicit equation (∇D_{eff} depends on $q(y_N|y)$)

$$q(y_N|y) = \frac{e^{\beta \frac{\nabla D_{eff}}{p(y)}}}{\sum_{y_N} e^{\beta \frac{\nabla D_{eff}}{p(y)}}}. \tag{8}$$

Here ∇D_{eff} denotes the gradient of D_{eff} with respect to the quantizer. For a fixed value of β we use a fixed point iteration

$$q_{n+1} := f(q_n),$$

where f is the right hand side of expression (8), to find a solution for the optimization problem.

3.2 Vertex search algorithm

Applying standard results from information theory, it is easy to show [9] that the function $D_{eff} = I(X; Y_n)$ is a convex function of the quantizer $q(y_N|y)$. Since the domain Δ is a product of simplices and therefore convex, we can prove the following Theorem

Theorem 4 *Let E be the set of vertices of Δ. Then*

$$\max_E D_{eff} \geq \max_\Delta D_{eff}.$$

Proof. Section 5.

This result allows us to reformulate problem (5) as follows

Theorem 5 *The optimal solution of the problem (5) with maximal possible value of D_{eff} can be found by the following algorithm:*

1. *Find a vertex $e \in E$ such that*

$$D_{eff}(e) := \max_E D_{eff}$$

2. *Assume e is a strict maximum of D_{eff} on the set E, i.e., for all neighboring vertices e_i we have $D_{eff}(e_i) < D_{eff}(e)$. Then e is an optimal solution of (5) with maximal possible value of D_{eff}.*

3. *Assume that $e = e_1$ is not a strict maximum. Then there are neighboring vertices $e_1, \ldots, e_k$ such that $D^* := D_{eff}(e_i) = D_{eff}(e_j)$ for all $1 \leq i, j \leq k$. Consider the region $Q_{y_1} \times \ldots \times Q_{y_s}$, where $Q_{y_j} \subset \Delta_{y_j}$ is the simplex spanned by the projection of these vertices to Δ_{y_j}. For all j, take $D_{y_j} \subset Q_{y_j}$ to be the maximal sub-simplex with the property that $D_{eff}(x) = D^*$ for all $x \in D_{y_1} \times \ldots \times D_{y_s}$. Then the solution of (5) is the product of the barycenters of D_{y_i}.*

Proof. Section 5.

In the absence of symmetries case 3 is non-generic. Lemma 18 shows that to determine whether the function D_{eff} is constant on a region of the form $D_{y_1} \times \ldots \times D_{y_s}$, one only needs to check the value of D_{eff} at a single interior point.

Since the set of vertices is large, we implement a local search, linear in the order of the space Y, which leads, under modest assumptions, to a local maximum of (5) (Theorem 19).

Vertex search algorithm (9)

1. We start the search from uniform solution $q(y_N|y) = 1/N$ for all y and all classes y_N.

2. Select randomly y_1 and evaluate the function D_{eff} at all the vertices of Δ_{y_1}, so that $q(L|y_1) = 1$ for some class $y_N = L$ and zero for all other classes M. Select the assignment of y_1 to a class which gives the maximal value of D_{eff}.

3. repeat step 2 with $y_2, y_3, \ldots$ until all y_k are assigned classes. This yields a vertex e of Δ.

4. Starting from the vertex e found in step 3, we repeat K_1 times the steps 1-3 until a local maximum in the set E is found.

5. The steps 1-4 are repeated K_2 times to avoid local maxima.

Remark 6 Clearly the assignment of y_1 to a class is arbitrary, so the algorithm should start with y_2 after y_1 is assigned to a class at random.

4 Applications

4.1 Synthetic Data

We analyze the performance of the three optimization schemes on synthetic data drawn from the probability distribution shown in figure 3a. In this model we assume that X represents a range of possible stimulus properties and Y represents a range of possible spike train patterns. We have constructed four clusters of pairs in this stimulus/response space. Each cluster

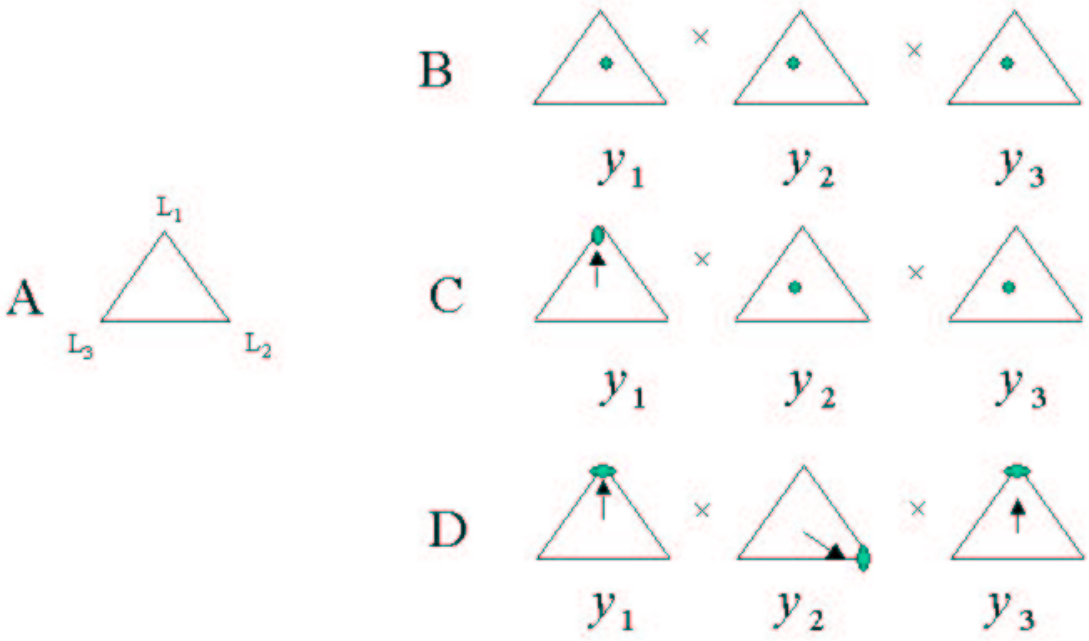

Figure 2: The vertex search algorithm, shown here for $N = 3$ and $|Y| = s = 3$. (A) A simplex Δ_y. Each vertex L_i corresponds to the value $q(L_i|y) = 1$. (B) The algorithm begins at some initial $q(y_N|y)$, in this case with $q(y_N|y) = 1/3$ for all y and y_N. (C) Randomly assign y_1 to a class L_1. (D) Assign y_2 consecutively to each class L_i, $i = 1, 2, 3$ and for each such assignment evaluate D_{eff}. Assign y_2 to the class L_i which maximizes D_{eff}. Repeat the process for y_3. Shown here is a possible classification of y_1, y_2 and y_3: y_1 and y_3 are put into class L_1, and y_2 is put into class L_2. Class L_3 remains empty.

corresponds to a range of responses elicited by a range of stimuli. The mutual information between the two sequences is about 1.8 bits, which is comparable to the mutual information conveyed by single neurons about stimulus parameters in several unrelated biological sensory systems [8, 20, 26, 31]. For this analysis we assume the relation between X and Y is known (the joint probability $p(x, y)$ is used explicitly).

The optimal quantizer $q^*(y_N|y)$ for $N = 2$, 3, 4 and 5 is show in panels b–f of figure 3. When an $N = 2$ class reproduction is forced as in panel (b), the algorithm recovers an incomplete representation of the coding scheme, in the sense that pairs of distinct classes in (a) are combined in (b). The representation is improved for the $N = 3$ class refinement (c). The next refinement (d) with $N = 4$ separates all the classes correctly and recovers most of the mutual information. Further refinements (e) fail to split the classes and are effectively identical to (d). Note that classes $y_N = 1$ and 2 in (e) are almost evenly populated and the class membership there is close to a uniform 1/2. That is, $q(y_N = 1|y) \approx q(y_N = 2|y) \approx 1/2$ for $y : 12 \leq y \leq 23$. The quantized mutual information in (f) increases with the number of classes approximately as $\log_2 N$ until it recovers about 90% of the original mutual information (at $N = 4$), at which point it levels off.

A random permutation of the rows and columns of the joint probability in figure 3a leaves the cost function unchanged. The optimal quantization is identical to the case presented in

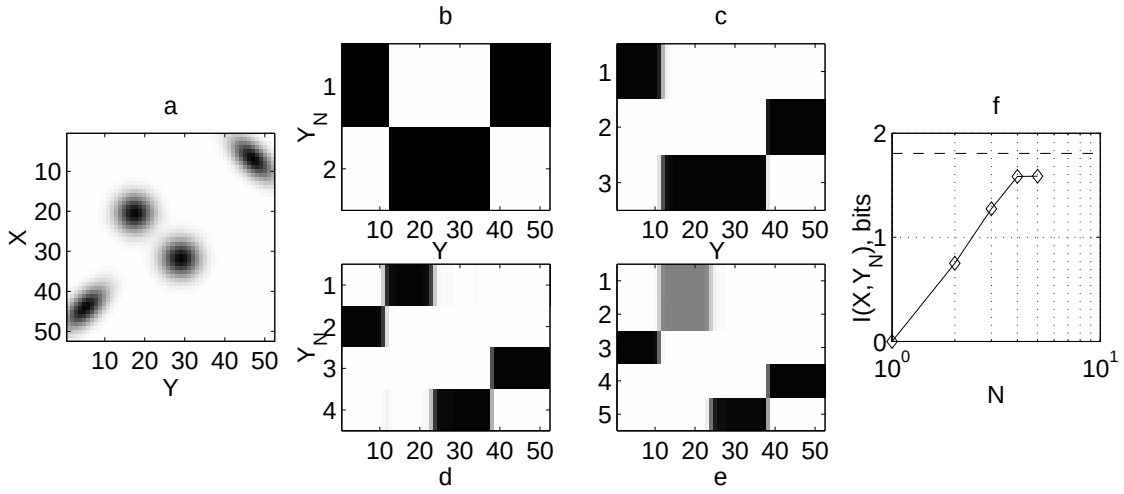

Figure 3: (a) A joint probability for the relation between two random variables X and Y, each with 52 elements. (b–e) The optimal quantizers $q(y_N|y)$ for $N = 2$, 3, 4 and 5 classes respectively. These panels represent the conditional probability $q(y_N|y)$ of a class y_N being associated with a response y. White represents $q(y_N|y) = 0$, black represents $q(y_N|y) = 1$, and intermediate values are represented by levels of gray. The behavior of the effective distortion $D_{eff} = I(X;Y_N)$ with increasing N can be seen in the log-linear plot (f). The dashed line is $I(X;Y)$, which is the least upper bound of $I(X;Y_N)$.

figure 3 after applying the inverse permutation and fully recovers the permuted classes (i.e., the quantization commutes with the action of the permutation group).

Further details of the course of the annealing optimization procedure (section 3.1) that lead to the optimal quantizer in panel (d) are presented in figure 4. The behavior of D_{eff} as a function of the annealing parameter β can be seen in the top panel. Snapshots of the optimal quantizers for different values of β are presented on the bottom row (panels 1 – 6). We can observe the bifurcations of the optimal solution (1 through 5) and the corresponding transitions of the effective distortion. The abrupt transitions ($1 \to 2$, $2 \to 3$) are similar to the ones described in [28] for a linear 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 seems to undergo a bifurcation.

Table 1 gives a comparison of our optimization algorithms for this data set. For $N = 2$, 3 and 4, left side of the table shows computational cost of each and the right side indicates the maximal value of D_{eff} procured by each algorithm. The vertex search was the fastest and the Augmented Lagrangian the slowest of the three with an order of magnitude difference between each two algorithms. The values of the cost function are almost identical. Each algorithm has its advantages, though, as the Augmented Lagrangian always gives a point that satisfies the *KKT* conditions and the vertex search does so under certain conditions (see Theorem 19). Although we do not have a complete theoretical understanding of the convergence of the implicit solution algorithm, it works very well in practice.

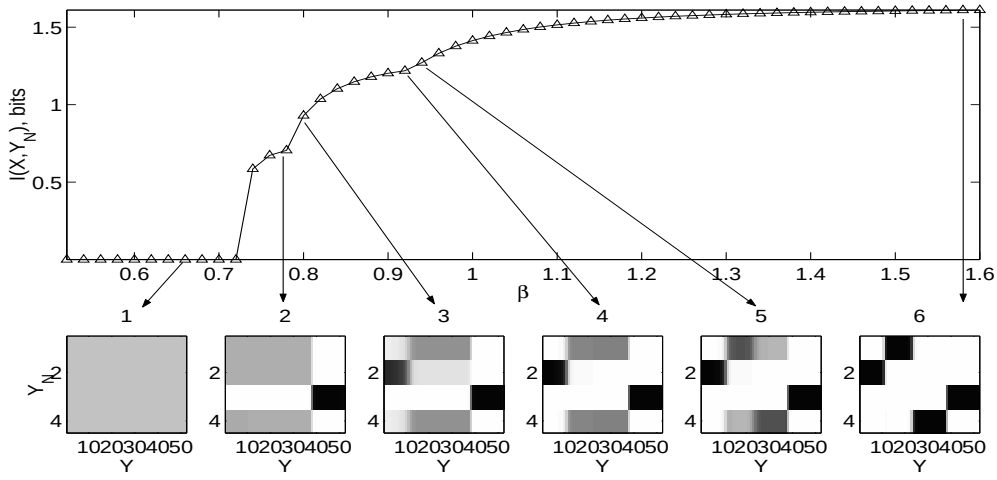

Figure 4: *For the data set in Figure 1A, the behavior of $D_{eff} = I(X;Y_N)$ (top) and the optimal quantizer $q^*(y_N|y)$ (bottom) as a function of the annealing parameter β.*

Algorithm	Cost in MFLOPs			$I(X;Y_N)$ in bits		
N	2	3	4	2	3	4
Lagrangian	431	822	1,220	0.8272	1.2925	1.6269
Implicit Solution	38	106	124	0.8280	1.2942	1.6291
Vertex Search	6	18	21	0.8280	1.2942	1.6291

Table 1: Comparison of the optimization schemes on synthetic data. The first three columns compare the computational cost in FLOPs. The last three columns compare the value of $D_{eff} = I(X;Y_N)$, evaluated at the optimal quantizer obtained by each optimization algorithm.

4.2 Physiological Data

4.2.1 Dealing with complex stimuli

To successfully apply our method to physiological data, we need to estimate the information distortion D_{eff}, which in turn depends on the joint stimulus/response probability $p(x,y)$. If the stimuli are sufficiently simple, $p(x,y)$ can be estimated directly as a joint histogram, and the method applied as described above. In general, we want to analyze conditions close to the natural for the particular sensory system, which usually entails observing stimulus sets of high dimensionality. Characterizing such a relationship non-parametrically is extremely difficult, since usually one cannot provide the large amounts of data this procedure needs. To cope with this regime, we model the stimulus/response relationship [10, 12]. The formulation as an optimization problem suggests certain classes of models which are better suited for this approach. We shall look for models that give us strict lower bounds $\tilde{D}_{eff}$ of the information distortion function D_{eff}. In this case, when we maximize the lower bound $\tilde{D}_{eff}$, the actual value of D_{eff} is also increased, since $I(X;Y) \geq D_{eff} \geq \tilde{D}_{eff} \geq 0$. This also gives us a quantitative measure of the quality of a model: a model with a larger $\tilde{D}_{eff}$ is better.

In [12, 11] we modeled the class conditioned stimulus $p(x|y_N)$ with the Gaussian: $p(x|y_N) =$

$N(x;\ x_{y_N}, C_{X|y_N})$. The class conditioned stimulus mean x_{y_N} and covariance matrix $C_{X|y_N}$ can be estimated from data. The stimulus estimate obtained in this manner is effectively a Gaussian mixture model [2]

$$\hat{p}(x) = \sum_{y_N} p(y_N) N(x;\ x_{y_N}, C_{X|y_N})$$

with weights $p(y_N)$ and Gaussian parameters $(x_{y_N}, C_{X|y_N})$. This model produces an upper bound [27] $\tilde{H}(X|Y_N)$ of $H(X|Y_N)$:

$$\tilde{H}(X|Y_N) = \sum_{y_N} p(y_N) \frac{1}{2} \log(2\pi e)^{|X|} det\Big[\sum_{y} p(y|y_N)(C_{X|y} + x_y^2) - (\sum_{y} p(y|y_N) x_y)^2 \Big]. \quad (10)$$

Here x_y^2 is the matrix $x_y x_y^T$.

Since $\tilde{H}(X|Y_N)$ is an upper bound on $H(X|Y_N)$ and

$$D_{eff} = I(X;Y_N) = H(X) - H(X|Y_N),$$

the quantity

$$\tilde{D}_{eff}(q(y_N|y)) := H(X) - \tilde{H}(X|Y_N) \quad (11)$$

is the lower bound to D_{eff}. This transforms the optimization problem (5) for physiological data to

$$\begin{aligned} &\max_{q(y_N|y)} H(Y_N|Y) && \text{constrained by} \quad (12) \\ &\tilde{D}_{eff}(q(y_N|y)) \geq I_0 && \text{and} \\ &\sum_{y_N} q(y_N|y) = 1 \quad \text{and} \quad q(y_N|y) \geq 0 \quad \forall y \in Y. \end{aligned}$$

It is not immediately obvious that solutions to (12) have properties similar to the solutions of (5). We show later in Theorem 11 that $\tilde{D}_{eff}$ is convex in $q(y_N|y)$. It follows that Theorem 4 and Theorem 5 hold for problem (12) as well and that the optimal quantizer $q^*(y_N|y)$ will be generically deterministic. This means that $\tilde{D}_{eff}$ can be used in place of D_{eff} in all optimization schemes discussed so far .

4.2.2 Results

A biological system that has been used very successfully to address aspects of neural coding [3, 5, 23, 24, 34] is the cricket's cercal sensory system. It provides the benefits of being simple enough so that all output signals can be recorded, yet sufficiently elaborate to address questions about temporal and collective coding schemes. The cricket's cercal system is sensitive to low frequency, near-field air displacement stimuli [18]. During the course of the physiological recording, the system was stimulated with air current stimuli, drawn from a band-limited (5-500Hz) Gaussian white noise (GWN) source [33]. We apply the method to intra-cellular recordings from identified inter-neurons in this system.

When applying the method to this data, the joint stimulus/response probability $p(x, y)$ needs to be estimated. We use $\tilde{D}_{eff}$ (11) in place of D_{eff}, and the optimization scheme (12). Figure 5 illustrates the dataset and optimal quantizers for this system. Sequences 2 through 105 in A were obtained by choosing 10 ms sequences from the recording which started with a spike (at time 0 here). Sequences in which the initial spike was preceded by another spike closer than 10 ms were excluded. Sequence 2 contains a single spike. Sequences 3-59 are doublets. Sequences 60-105 are triplets. Sequence 1 is a well isolated empty codeword (occurrences were chosen to be relatively far from the other patterns). Each pattern was observed multiple times (histogram not shown).

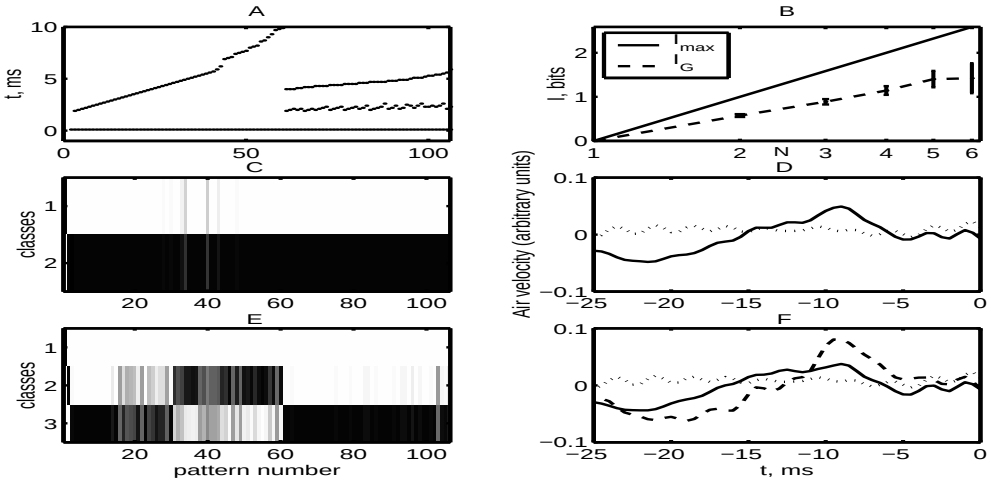

Figure 5: *Results from the information distortion method. A) All the response spike patterns that were analyzed. Each dot represents the occurrence of a single spike. Each column of dots represents a distinct sequence of spikes. The y axis is the time in ms after the occurrence of the first spike in the pattern. The x axis here and below is an arbitrary number, assigned to each pattern. B) The lower bound of I (dashed line) obtained through the Gaussian model can be compared to the absolute upper bound $I = \log_2 N$ for an N class reproduction (solid line). C) The optimal quantizer for $N = 2$ classes. This is the conditional probability $q(y_N|y)$ of a pattern number y from A) (horizontal axis) belonging to class y_N (vertical axis). White represents zero, black represents one, and intermediate values are represented by levels of gray. D) The means, conditioned on the occurrence of class 1 (dotted line) or 2 (solid line). E) The optimal quantizer for $N = 3$ classes. F) The means, conditioned on the occurrence of class 1 (dotted lne), 2 (solid line) or 3 (dashed line).*

Panels C–F show the results of applying the information distortion approach to this dataset. The optimal quantizer for the $N = 2$ reproduction is shown in panel C. It isolates the empty codeword in one class (class $y_N = 1$) and all other patterns in another class (class $y_N = 2$). The mean of the stimuli conditioned with the zero codeword (panel D, dotted line), does not significantly deviate from a zero signal. Panels E and F show the results of extending the analysis to a reproduction of $N = 3$ classes. The zero codeword remains in class 1. The former class 2 is split into two separate classes: class 2, which contains the single spike codeword and codewords with an inter-spike interval $ISI > 5ms$, and class 3, which contains all doublets with $ISI < 2ms$ and all triplets. The mean in (D, solid line) is

split into two separate class conditioned means (F, solid and dashed line).

Algorithm	Cost in GFLOPs				$I(X, Y_N)$ in bits		
N	3	4	5		3	4	5
Lagrangian	13	29	59		0.18	0.18	0.16
Implicit Solution	7	11	9		0.43	0.80	1.14
Vertex Search	31	84	141		0.44	0.85	1.81

Table 2: Comparison of the optimization schemes on physiological data. The first four columns compare the computational cost in gigaFLOPs. The last four columns compare the value of $D_{eff} = I(X; Y_N)$, evaluated at the optimal quantizer obtained by each optimization algorithm.

In table 2 we compare the three algorithms on the physiological data set. We see that the cost is lowest for the Implicit solution algorithm, but the vertex search finds the best solution measured in value of D_{eff}.

5 Analytical results

Recall that the admissible region Δ for the linear constraints in (5) is a direct product of simplices $\Delta := \Pi_y \Delta_y$ where each $\Delta_y := \{q(y_N|y) \mid \sum_N q(y_N|y) = 1\}$. Observe that $\Delta \subset \mathbf{R}^{ns}$ where n is the number of quantization classes and s is the number of elements in Y ($s = |Y|$). We show that the optimal solution generically occurs at a vertex of Δ. This allows us to reformulate (5) as a maximization of D_{eff} on the set of vertices. Then we prove that under certain conditions, the vertex search algorithm (9) always finds a local maximum.

5.1 Maximum on the boundary

Lemma 7 *The function $I(X; Y_N)$ is a convex function of $q(y_N|y)$.*

Proof. Lemma B.1 of [9]. □

Lemma 8 *Given a convex function $f(x)$, $x \in \mathbf{R}^k$, the set $S(k) := \{x \mid f(x) \leq k\}$ is convex.*

Proof. Let $x_\theta := \theta x_0 + (1 - \theta)x_1$. Assume $x_0, x_1 \in S(k)$. Then

$$f(x_\theta) \leq \theta f(x_0) + (1 - \theta) f(x_1) \leq k$$

where the first inequality follows from convexity of the function f and the second from the fact that $x_0, x_1 \in S(k)$. □

Recall that E denotes the set of all vertices of the set Δ. An element e in this set can be written as

$$e = \Pi_y e_y$$

where e_y is a vertex of the simplex Δ_y.

Lemma 9 *The set Δ is the convex hull of E.*

Proof. Let $C :=$ convex hull (E). We show first that $\Delta \subset C$. Select a point $w \in \Delta$. Such a point is determined by a collection of barycentric coordinates $s_y^1, \ldots, s_y^n$ in Δ_y for each y. To show that $w \in C$ we need to find numbers λ_j such that

$$w = \sum_{e(j) \in E} \lambda_j e(j), \quad \sum \lambda_j = 1. \tag{13}$$

We denote the vertices of the simplex Δ_y by $v_y^1, v_y^2, \ldots, v_y^n$. Observe that (13) will be satisfied if

$$\sum_{e_y(j) = v_y^k} \lambda_j = s_y^k, \quad \text{for all } y = 1, \ldots, s, \quad k = 1, \ldots, n. \tag{14}$$

We construct the numbers λ_j satisfying (14) explicitly. We start our construction with a collection of sn barycentric coordinates s_y^k, for $k = 1, \ldots, n$ and for each $y \in Y$, which specify the point w. Let

$$S_y := max_k \; s_y^k$$

for each y and let $m(y) = \text{argmax}_k \; s_y^k$. Hence $S_y = s_y^{m(y)}$. Let $e(1)$ be a vertex of Δ such that

$$e_y(1) := v_y^{m(y)} \quad \text{for each } y. \tag{15}$$

Finally, we select

$$\lambda_1 := min_y \; S_y.$$

Notice that $\lambda_1 \neq 0$ since for each y at least one $s_y^k \neq 0$. We let $s_y^k(0) := s_y^k$ for all y and all k. We construct a new set of numbers $s_y^k(1)$ (which are no longer barycentric coordinates) in the following way: we replace each number $s_y^{m(y)}$ by number $s_y^{m(y)} - \lambda_1$

$$s_y^{m(y)}(1) := s_y^{m(y)}(0) - \lambda_1. \tag{16}$$

We note two facts about this construction

1. After replacement (16), the sum

$$\sum_{k=1}^{n} s_y^k(1) + \lambda_1 = 1 \quad \text{for each } y.$$

2. At least one number $s_y^{m(y)}(1)$ is zero.

We repeat the construction with the new set of numbers $s_y^k(1)$ instead of the set $s_y^k(0)$. In general, in the l-th step we construct vertex $e(l)$, coefficient λ_l, and a new set of numbers $s_y^k(l)$. After the l-th step of the construction we observe that

1. For each fixed y, the sum

$$\sum_{k=1}^{n} s_y^k(l) + \sum_{i=1}^{l} \lambda_i = 1. \tag{17}$$

2.

$$\text{At least } l \text{ numbers } s_y^k(l) \text{ are zero.} \tag{18}$$

Claim 10 *If for $y = t$ there is only one nonzero element s_t^k and for some $y = q$ there are $u > 1$ nonzero elements $s_q^{p_1}, \ldots, s_q^{p_u}$, then in next step of the algorithm s_t^k will not be selected as λ_j.*

Proof. Observe that by (17) above

$$s_t^k = 1 - \sum_{i=1}^{j-1} \lambda_i = \sum_{i=1}^{u} s_q^{p_i}$$

and so the maximum of the set $s_q^{p_i}, i = 1, \ldots, u$ is smaller than s_t^k. □

It follows from the Claim and (18) above that after at most $s(n-1)$ steps the algorithm comes to the situation where for each y there is precisely one $s_y^{k(y)} \neq 0$ and all other s_y^l are zero. Again, by (17), the numbers $s_y^{k(y)}$ are equal

$$s_y^{k(y)} = 1 - \sum_{i=1}^{j-1} \lambda_i \text{ for all } y.$$

Hence, in the next step, $\lambda_j := s_t^{k(t)}$, $s_y^k(j) = 0$ for all y, k and the algorithm ends. For the vertices $e(i)$ which did not come up in the construction step (15), we set $\lambda_i = 0$. Note that it follows immediately from (17) that

$$\sum_{i=1}^{j} \lambda_i = 1.$$

By construction, $s_y^k(l) \neq s_y^k(l+1)$ for some l if and only if $\lambda_l = s_y^k(l) - s_y^k(l+1)$ and the corresponding vertex $e(l)$ has y-th component, $e_y(l)$, equal to v_y^k. It follows that (14) is satisfied. Hence (13) holds and this proves $\Delta \subset C$.

To show that $C \subset \Delta$ it is enough to realize that Δ, being a product of convex sets Δ_y, is convex. Since C is the smallest convex set containing E and $E \subset \Delta$, we have $C \subset \Delta$. □

Proof of Theorem 4.

Denote $M := max_E D_{eff}$ and let

$$A := \{q(y_n|y) \mid D_{eff} \leq M\}.$$

By Lemma 7 and Lemma 8, A is a convex set. Since $E \subset A$, then for C, which is the convex hull of E (and hence the smallest convex set containing E), we have $C \subset A$. By Lemma 9, $C = \Delta$ and thus $\Delta \subset A$. □

5.2 Cost function for physiological data

In applications to physiological data, one can either estimate the joint probability $p(x, y)$ directly and then use the cost function $D_{eff} = I(X; Y_N)$, or, as the authors did in [12, 11], one can use a different cost function, $\tilde{D}_{eff}$, which is a lower bound of D_{eff}. Recall that

$$\tilde{D}_{eff} = H(X) - \tilde{H}(X|Y_N),$$

and $\tilde{H}(X|Y_N)$ is defined in (10).

Theorem 11 *The function $\tilde{H}(X|Y_N)$ is concave in $q(y_N|y)$ and hence the function $\tilde{D}_{eff}$ is convex in $q(y_N|y)$.*

Our argument is based on four Lemmas.

Lemma 12 (Ky-Fan [22]) *The function* log *det A is concave in A.*

Lemma 13 *For all i and j, the (i, j)-th component of the matrix*

$$\mathcal{F} := \sum_y p(y|y_N)(C_{X|y} + x_y^2) - (\sum_y p(y|y_N)x_y)^2$$

is concave in $p(y|y_N)$.

Proof. The first part of $\mathcal{F}$ is linear in $p(y|y_N)$. We look at the second part. Fix i and j and look at the (i, j)-th component of the matrix. After taking out the constants we get that the second part is a function of the form

$$g_{ij} := -(\sum_y a_y p(y|y_N))(\sum_y b_y p(y|y_N)) \tag{19}$$

where a is a vector of i-th components $([x_{y_1}]^i, [x_{y_2}]^i, \ldots, [x_{y_n}]^i)$ and b is a similar vector of j components of x_y. Denote the vector $u := (p(y_1|y_N), p(y_2|y_N), \ldots, p(y_n|y_N))$. Differentiating g_{ij} we arrive at

$$\nabla^2 g_{ij} = -(ba^T + ab^T).$$

The function g_{ij} is concave if the quadratic form

$$u^T(ba^T + ab^T)u$$

is positive semidefinite. Observe that both matrix ba^T and matrix ab^T have rank 1 and so the rank of matrix $M := (ba^T + ab^T)$ is at most two. To show positive semi-definiteness we need to show that the nonzero eigenvalues are nonnegative.

Note that equation $Mv = \lambda v$ leads to

$$(ba^T)v + (ab^T)v = b(a^T v) + a(b^T v) = \lambda v.$$

Since both $(a^T v)$ and $(b^T v)$ are scalars this shows that eigenvectors with nonzero eigenvalues must be in $span\{a, b\}$.

We compute the eigenvalues and eigenvectors by setting $v = c_1 a + c_2 b$ where the constants c_1, c_2 are to be determined.

$$\begin{aligned}(ba^T)v + (ab^T)v &= (ba^T)(c_1 a + c_2 b) + (ab^T)(c_1 a + c_2 b) \\ &= b(c_1(a^T a) + c_2(a^T b)) + a((c_1(b^T a) + c_2(b^T b)) \\ &= \lambda(c_1 a + c_2 b).\end{aligned}$$

Collecting on both sides we get a system

$$c_1(a^T a) + c_2(a^T b) = \lambda c_1, \quad c_1(b^T a) + c_2(b^T b) = \lambda c_2.$$

In matrix form, this is $A(c_1, c_2)^T = \lambda(c_1, c_2)^T$, where

$$A = \begin{bmatrix} a^T a & a^T b \\ b^T a & b^T b \end{bmatrix}.$$

So λ is also an eigenvalue of the matrix A. Observe that $a^T a > 0$ and the determinant of A is

$$detA = (a^T a)(b^T b) - (a^T b)^2 \geq 0$$

by Cauchy-Schwartz inequality. So A has nonnegative eigenvalues which are also eigenvalues of M. □.

Lemma 14 *Let $F : \mathbf{R}^n \to \mathbf{R}$ and $f : \mathbf{R}^n \to \mathbf{R}^n$ such that*

1. *$\nabla^2 F$ is negative semidefinite*
2. *If we denote $f = (f_1, f_2, \ldots, f_n)$, then for each i, $x^T \nabla^2 f_i x = 0$.*

Then for $G = F \circ f : \mathbf{R}^n \to \mathbf{R}$ the matrix $\nabla^2 G$ is negative semidefinite.

Proof. Straightforward computation shows that

$$\nabla G = Df \, \nabla F,$$

where Df is $n \times n$ matrix and both ∇G and ∇F are n-vectors. We write out the l-th component of ∇G

$$\frac{\partial G}{\partial x_l} = \sum_{j=1}^{n} \frac{\partial F}{\partial f_j} \frac{\partial f_j}{\partial x_l}.$$

Now compute the (l,k)-th element of the matrix $\nabla^2 G$

$$\begin{aligned}(\nabla^2 G)_{lk} = \frac{\partial}{\partial x_k}(\frac{\partial G}{\partial x_l}) &= \sum_{j=1}^{n}\sum_{s=1}^{n} \frac{\partial^2 F}{\partial f_j \partial f_s} \frac{\partial f_s}{\partial x_k} \frac{\partial f_j}{\partial x_l} + \sum_{j=1}^{n} \frac{\partial F}{\partial f_j} \frac{\partial^2 f_j}{\partial x_k \partial x_l} \\ &= \frac{df}{dx_k} \nabla F \frac{df}{dx_l} + \sum_j \frac{\partial F}{\partial f_j} (\nabla^2 f_j)_{k,l}\end{aligned}$$

Finally, we compute $x^T \nabla^2 G x$:

$$\begin{aligned} x^T \nabla^2 G x &= \sum_l \sum_k (\nabla^2 G)_{l,k} x_l x_k \\ &= \sum_k \sum_l x_k \frac{df}{dx_k} \nabla^2 F \frac{df}{dx_l} x_l + \sum_k \sum_l \sum_j \frac{\partial F}{\partial f_j} (\nabla^2 f_j)_{k,l} x_k x_l \\ &= (Dfx)^T \nabla^2 F (Dfx) + \sum_j \frac{\partial F}{\partial f_j} (x^T \nabla^2 f_j x). \end{aligned}$$

By the second assumption the last term is zero and so

$$x^T \nabla^2 G x = (Dfx)^T \nabla^2 F (Dfx).$$

The first assumption now guarantees that $\nabla^2 G$ is negative semidefinite. □

Lemma 15 *Fix the value of the random variable $y_N = M$. Let*

$$f_i(q(M|y)) := p(y_i|M) = \frac{q(M|y_i)p(y_i)}{p(M)} = \frac{q(M|y_i)p(y_i)}{\sum_j q(M|y_j)p(y_j)}.$$

Then, if we denote $q = (q(M|y_1), q(M|y_2), \ldots, q(M|y_n))$, we have

$$q^T \nabla^2 f_i q = 0$$

for all i.

Proof. To simplify notation we let $a_i := p(y_i)$, $u_i := q(M|y_i)$ and $u = (u_i, \ldots, u_n)$. Then

$$f_i(u) = \frac{a_i u_i}{\sum_j a_j u_j}.$$

We compute

$$\begin{aligned} \frac{\partial f_i}{\partial u_l} &= \delta_{li} \frac{a_l (\sum_j a_j u_j)}{(\sum_j a_j u_j)^2} - \frac{a_l a_i u_i}{(\sum_j a_j u_j)^2} \\ &= \delta_{li} \frac{a_l}{\sum_j a_j u_j} - \frac{a_l a_i u_i}{(\sum_j a_j u_j)^2}, \end{aligned}$$

where $\delta_{li} = 1$ if $l = i$ and zero otherwise. The second derivative is

$$\frac{\partial f_i}{\partial u_l \partial u_k} = -\delta_{li} \frac{a_l a_k}{(\sum_j a_j u_j)^2} - \delta_{ki} \frac{a_l a_k}{(\sum_j a_j u_j)^2} + 2 \frac{a_k a_l a_i u_i u_k u_l}{(\sum_j a_j u_j)^3}.$$

Then $u^T \nabla^2 f_i u$ is

$$\begin{aligned} u^T \nabla^2 f_i u &= \sum_{k,l} \frac{\partial f_i}{\partial u_l \partial u_k} u_k u_l \\ &= \frac{1}{(\sum_j a_j u_j)^2} \left[\sum_k -a_i a_k u_i u_k - \sum_l a_i a_l u_i u_l + \frac{2 a_i u_i}{\sum_j a_j u_j} \sum_{k,l} a_k u_k a_l u_l \right] \\ &= \frac{1}{(\sum_j a_j u_j)^2} (a_i u_i) \left[-\sum_k a_k u_k - \sum_l a_l u_l + \frac{2 \sum_l a_l u_l \sum_k a_k u_k}{\sum_j a_j u_j} \right] \\ &= 0. \end{aligned}$$

Proof of Theorem 11

Lemma 15 and Lemma 13 verify the assumptions of Lemma 14 where we set $f := f_i$ (f_i from Lemma 15) and $F = g_{lk}$ (g_{lk} from Lemma 13), for any k, l, i. Hence, by Lemma 14, each (k,l)-th component g_{lk} of $\mathcal{F}$ is a concave function of $q(M|y_i)$ for all i. By Lemma 12 the function $\log det \mathcal{F}$ is concave in $\mathcal{F}$ and thus in $q(M|y_i)$ for any i. At this point we should write $\mathcal{F}_M$ instead of $\mathcal{F}$ since we have the value $y_N = M$ fixed in computation of $\mathcal{F}$. Clearly our argument is true for any such M. Finally, since $p(y_N) = \sum_y q(y_N|y)p(y)$ is a linear combination of $q(y_N|y)$, then the function $\tilde{H}(X|Y_N) = \tilde{H}(X|Y_N)(q(y_N|y))$ (compare (10)) is a linear combination of concave functions

$$\log det \mathcal{F}_M,$$

where $\mathcal{F}_M$ has fixed value $y_N = M$. This finishes the proof. □

Theorem 16 *If E is the set of vertices of the domain Δ, then*

$$\max_E \tilde{D}_{eff} \geq \max_\Delta \tilde{D}_{eff}.$$

Proof. Analogous to the proof of Theorem 4, where we use the Theorem 11 instead of Lemma 7. □

5.3 Equivalent problem

Proof of Theorem 5

We select a vertex e such that $D_{eff}(e) = max_E D_{eff}$. Then either case [2.] or case [3.] happens.

Assume that [2.] happens. Then, by Theorem 4 the function D_{eff} achieves a global maximum at e over all Δ. Therefore, in problem (5) the feasible domain of the maximization problem with value $I_0 = D_{eff}(e)$ consists of at most a finite number of isolated vertices of Δ. Each such vertex corresponds to a deterministic quantizer $q(y_N|y)$. Entropy $H(Y_N|Y)$ of any such quantizer is zero. So the solution of (5) with maximal possible value of D_{eff} is e.

Now assume that the case [3.] above happens. Then the feasible domain of the maximization problem with value $I_0 = D^*$ is $D_{y_1} \times \ldots \times D_{y_s}$. Then the solution with maximum entropy is the product of barycenters of D_{y_i}. □

Corollary 17 *The problem for physiological data (12) is equivalent to the problem described in Theorem 5, where the function D_{eff} is replaced by function $\tilde{D}_{eff}$.*

Proof. The only difference in the proof for the function $\tilde{D}_{eff}$ is that we use Theorem 16 instead of Theorem 4. □

We also prove the following Lemma which shows that in order to determine whether the cost function D_{eff} is constant on a set $D_{y_1} \times \ldots \times D_{y_s}$, one needs to check the value at a single interior point.

Lemma 18 *Let f be a convex function, $f : D \to R$, where $D := D_1 \times \ldots \times D_k$ and D_i is a simplex. Assume that $\max_D f(x) \leq \max_E f(x)$, where E is the vertex set of D, and that $f(e) = k$ for every vertex $e \in E$. Assume also that there exists an interior point p of D such that $f(p) = k$. Then $f(x) = k$ for all $x \in D$.*

Proof. Fix a set $U := \mathrm{Int}(D_1 \times \ldots \times D_k)$, $U \subset D$. Let $A := \{x \in U \mid f(x) = k\}$. This set is clearly closed since $A = f^{-1}(k)$ and f is continuous. We show that A is open in U. Let us first consider $x \in A$. Since U is open, there is an open neighborhood $N(x) \subset U$. Pick an arbitrary $y \in N(x)$. Since $N(x)$ is open there is a $z \in N(x)$ such that

$$(y + z)/2 = x.$$

By convexity $k = f(x) \leq f(y)/2 + f(z)/2$. By assumption $f(y) \leq f(x)$ and $f(z) \leq f(x)$ and so

$$f(x) \leq f(y)/2 + f(z)/2 \leq f(x)/2 + f(x)/2 = f(x).$$

It follows that $f(x) = f(y) = f(z) = k$. Hence if $x \in A$ then $N(x) \subset A$. Since every U is connected, either $A = U$ or $A = \emptyset$. By assumption $p \in A$ and so $A = U$. By continuity of the function f

$$f(x) = k \quad \text{for all} \quad x \in D.$$

□.

5.4 Vertex Search Algorithm and convergence to a local maximum

In this section we show that the vertex search algorithm converges to a local maximum under certain conditions.

Theorem 19 *The point e, obtained by a vertex search, is a local maximum of D_{eff} if for each k, when $q(y_N|y_k)$ is determined, we have*

$$p(x, y_k) << \sum_{y_i \in L, i \neq k} p(x, y_i), \quad p(y_k) << \sum_{y_i \in L, i \neq k} p(y_i)$$

for each class L.

Proof. Assume that the points y_i, $i = 1, \ldots, k-1$ were assigned to their prospective classes by steps 2 and 3. The algorithm decides where to assign y_k based on the value of $D_{eff} - I(X; Y_N)$ of different assignments at this point. We can write $I(y_k \to L)$ for the value of the mutual information when we assign $q(L|y_k) = 1$ and $q(M|y_k) = 0$ for $M \neq L$.

Let

$$S(L, x) := \frac{\sum_y q(L|y)p(x, y)}{p(x) \sum_y q(L|y)p(y)},$$

(compare to (6)). We denote $S_L(N,x)$ the function $S(N,x)$ where we assigned y_k to class L (i.e. $q(y_N = L|y_k) = 1$ and zero otherwise). We compute

$$\begin{aligned} I(y_k \to L) &= \sum_{x,y,y_N} q(y_N|y)p(x,y)\log S(y_N,x) \\ &= \sum_{y_N}\sum_{x}\log S(y_N,x)[\sum_{y_i \in y_N} p(x,y)] \\ &= \sum_{y_N \neq L}\sum_{x}\log S(y_N,x)[\sum_{y_i \in y_N, i\neq k} p(x,y)] \\ &+ \sum_{x}\log S_L(L,x)[\sum_{y_i \in L, i\neq k} p(x,y) + p(x,y_k)] \end{aligned} \tag{20}$$

We select $q(L|y_k) = 1$ if and only if

$$d_{LM} := I(y_k \to L) - I(y_k \to M) \geq 0$$

for all $M \neq L$. Observe that most summands in the first term in (20) are the same for $I(y_k \to L)$ and $I(y_k \to M)$. Then d_{LM} for fixed classes L and M is

$$\begin{aligned} d_{LM} &= \sum_{x}\log S_L(L,x)[\sum_{y_i \in L, i\neq k} p(x,y) + p(x,y_k)] \\ &+ \sum_{x}\log S_L(M,x)[\sum_{y_i \in M, i\neq k} p(x,y)] \\ &- \sum_{x}\log S_M(L,x)[\sum_{y_i \in L, i\neq k} p(x,y)] \\ &- \sum_{x}\log S_M(M,x)[\sum_{y_i \in M, i\neq k} p(x,y) + p(x,y_k)] \end{aligned} \tag{21}$$

Denote $\epsilon_1 := \frac{p(x,y_k)}{\sum_{y\in M} p(x,y)}$ and $\epsilon_2 := \frac{p(y_k)}{\sum_{y\in M} p(y)}$. We compute

$$\begin{aligned} \frac{S_L(M,x)}{S_M(M,x)} &= \frac{\sum_{y\in M, y\neq k} p(x,y)}{\sum_{y\in M, y\neq k} p(x,y) + p(x,y_k)} \frac{\sum_{y\in M, y\neq k} p(y) + p(y_k)}{\sum_{y\in M, y\neq k} p(y)} \\ &= (\frac{1}{1+\epsilon_1})(1+\epsilon_2) \\ &\approx 1 - \epsilon_1\epsilon_2. \end{aligned}$$

Similarly,

$$\frac{S_L(L,x)}{S_M(L,x)} \approx 1.$$

Then from (21) we get

$$\begin{aligned} d_{LM} &= \sum_x p(x,y_k)[\log S_L(L,x) - \log S_M(M,x)] \\ &+ \sum_x \log \frac{S_L(L,x)}{S_M(L,x)}[\sum_{y\in L, y\neq k} p(x,y)] \\ &+ \sum_x \log \frac{S_L(M,x)}{S_M(M,x)}[\sum_{y\in M, y\neq k} p(x,y))] \\ &= \sum_x p(x,y_k)[\log S_L(L,x) - \log S_M(M,x)] + O(\epsilon_1\epsilon_2). \end{aligned} \tag{22}$$

Now we look at conditions under which a vertex e is a local maximum of the function D_{eff}. These conditions are equivalent to the Karush-Kuhn-Tucker conditions for a local maximum. At a point where D_{eff} achieves a local maximum the projection of the gradient ∇D_{eff} onto each affine space forming the boundary of Δ must fall outside Δ. A boundary near a vertex is a collection of affine faces, each spanned by the vectors $e - e_i$, where e_i is a vertex of Δ which differs from e in i-th component only. If the projection

$$(\nabla D_{eff})_e \cdot (e - e_i) \geq 0 \tag{23}$$

for all i then e is a local maximum. For each i there are n vectors $e_i = \{e_i^L\}_{y_N = L}$. From [9],

$$(\nabla D_{eff})_{q(L|\bar{y})} = \sum_x p(x,\bar{y}) \log S(L,x).$$

Select $\bar{y} = y_k$. Assume that at the vertex e we have $q(L|y_k) = 1$. Taking the dot product of ∇I with the vector $e - e_k^M$, where the gradient is evaluated at the point e gives

$$(\nabla D_{eff})_e \cdot (e - e_k^M) = \sum_x p(x,y_k)[\log S_L(L,x) - S_M(M,x)]. \tag{24}$$

Observe that in the course of the algorithm $q(L|y_k)$ is selected to be 1, if and only if $d_{LM} \geq 0$ for all $M \neq L$. This condition, by (22), is equivalent for small $\epsilon_1\epsilon_2$ to condition for local maximum ((23) with (24)) at the point e. □

6 Discussion and Conclusions

Our goal in this paper is to develop the mathematical theory behind an approach aimed at discovering the neural code in biological neural systems. The main parts of this approach are [9]:

- We model the neural coding problem in biological systems as a communicational chanel.
- We develop a strategy to obtain an approximation of a coding scheme and its natural size, N_c. This strategy is based on quantizing the output random variable to a smaller reproduction space.

- We formulate an optimal quantization problem, using a distortion function based on the expected Kullback-Leibler distortion.

In this paper we developed some of the mathematical ideas underlying this approach. We have shown that the optimal quantizer $q^*(y_N|y)$ is deterministic and lies on the vertex of the feasible region for the original problem (5) and the problem (12), which arises when we deal with physiological data. Using this fact we designed a vertex search algorithm and have shown that it converges under mild conditions to a local maximum. We have developed two annealing algorithms to solve the optimization problem as well. Each algorithm has its advantages and disadvantages and we tested them on synthetic and physiological data sets.

Most current approaches to studying neural coding rely on formulating a hypothesis about the coding scheme and then using observations to estimate parameters of the hypothesis. The complexity of the hypothesis determines the amount of data needed for reliable estimates of the necessary parameters. The method presented in this report offers a means for data-driven hypothesis formulation. When we stop the refinement of the reproduction due to lack of data, we effectively formulate a hypothesis about the most informative coding scheme that can be supported with the available amount of data. When more observations become available, the hypothesis can be refined automatically to include them for a better approximation.

We have successfully applied our approach to physiological data sets from the cercal system of the cricket [12, 11]. Our input in the physiological recordings was drawn from a band-limited Gaussian white noise source to avoid making unnecessary assumptions about the nature of the input signal that is relevant to the cricket. Our method did not make any assumptions about the character of the coding process, except the assumptions implicit in the fact that we selected to model this process as an information chanel. In particular, we did not assume that the process is linear, as the methods based on kernel reconstructions do [27], nor have we assumed that the coding is through the mean spike rate [29]. Our approach is more general than existing approaches in the literature and subsumes most of them as special cases (see [9] for a detailed discussion).

It is interesting to note that, although we had neural coding in mind while developing the information distortion method, the ensuing analysis is in no way limited to nervous systems. Indeed, the constraints on the pair of 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.

References

[1] H. B. Barlow. Possible princilples underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communications*. MIT Press, Cambridge, MA, 1961.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1998.

[3] D. A. Bodnar, J. Miller, and G. A. Jacobs. Anatomy and physiology of identified wind-sensitive local interneurons in the cricket cercal sensory system. *J. Comp. Physiol. A*, 168:553–564, 1991.

[4] L. Breiman. *Probability*. Addison-Wesley Publishing Companyy, Menlo Park, CA, 1968.

[5] H. Clague, F. Theunissen, and J. P. Miller. The effects of adaptation on neural coding by primary sensor interneurons in the cricket cercal system. *J. Neurophysiol.*, 77:207–220, 1997.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley Series in Communication, New York, 1991.

[7] A. G. Dimitrov and J. P. Miller. Analyzing sensory systems with the information distortion function. In R. B. Altman, editor, *Pacific Symposium on Biocomputing 2001*. World Scientific Publushing Co., 2000.

[8] A. G. Dimitrov and J. P. Miller. Natural time scales for neural encoding. *Neurocomputing*, 32-33:1027–1034, 2000.

[9] A. G. Dimitrov and J. P. Miller. Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

[10] A. G. Dimitrov, J. P. Miller, and Z. Aldworth. Neural coding and decoding. New Orleans, November 2000. Society for Neuroscience Annual Meeting.

[11] A. G. Dimitrov, J. P. Miller, Z. Aldworth, T. Gedeon, and A. E. Parker. Coding schemes based on spike patterns in a simple sensory ssystem. *J. Neurosci.*, 2002. *(under review)*.

[12] A. G. Dimitrov, J. P. Miller, Z. Aldworth, and A. Parker. Spike pattern-based coding schemes in the cricket cercal sensory system. *Neurocomputing*, 2002. *(to appear)*.

[13] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[14] R. M. Gray. *Entropy and Information Theory*. Springer-Verlag, 1990.

[15] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (London)*, 195:215–243, 1961.

[16] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proc. IEEE*, 70:939–952, 1982.

[17] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of the neural code. *J. Comp. Neurosci*, 10(1):47–70, 2001.

[18] G. Kamper and H.-U. Kleindienst. Oscillation of cricket sensory hairs in a low frequency sound field. *J. Comp. Physiol. A.*, 167:193–200, 1990.

[19] C. T. Kelley. *Iterative Methods for Optimization.* SIAM, Philadelphia, 1999.

[20] T. W. Kjaer, J. A. Hertz, and B. J. Richmond. Decoding cortical neuronal signals: Network models, information estimation and spatial tuning. *J. Comp. Neurosci*, 1(1-2):109–139, 1994.

[21] S. Kullback. *Information Theory and Statistics.* J Wiley and Sons, New York, 1959.

[22] Ky-Fan. On a theorem of weyl concerning the eigenvalues of linear transformations ii. Proc. NAtional. Acad. Sci. U.S., 36:31–35, 1950.

[23] M. A. Landolfa and J. P. Miller. Stimulus-response properties of cricket cercal filiform hair receptors. *J. Com. Physiol. A.*, 177:749–757, 1995.

[24] J. P. Miller, G. A. Jacobs, and F. E. Theunissen. Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons. *J. Neurophys*, 66:1680–1689, 1991.

[25] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, 2000.

[26] P. Reinagel and R. Reid. Temporal coding of visual information in the thalamus. *J. Neurosci.*, 20(14):5392–5400, 2000.

[27] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the neural code.* The MIT Press, 1997.

[28] K. Rose. Deteministic annealing for clustering, compression, classification, regerssion, and related optimization problems. *Proc. IEEE*, 86(11):2210–2239, 1998.

[29] M. N. Shadlen and W. T. Newsome. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.*, 4:569–579, 1994.

[30] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:623–656, 1948.

[31] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Let.*, 80(1):197–200, 1998.

[32] F. Theunissen and J. P. Miller. Temporal encoding in nervous systems: A rigorous definition. *J. Comp. Neurosci*, 2:149–162, 1995.

[33] F. Theunissen, J. C. Roddey, S. Stufflebeam, H. Clague, and J. P. Miller. Information theoretic analysis of dynamical encoding by four primary sensory interneurons in the cricket cercal system. *J. Neurophy.*, 75:1345–1359, 1996.

[34] F. E. Theunissen and J. P. Miller. Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning curve width of four primary interneurons. *J. Neurophysiol.*, 66:1690–1703, 1991.