Submitted to: *Network: Comput. Neural Syst.*

Analysis of neural coding through quantization with an information-based distortion measure

Alexander G. Dimitrov†, John P. Miller†, Tomáš Gedeon‡, Zane Aldworth† and Albert E. Parker†‡§

†Center for Computational Biology and
‡Department of Mathematical Sciences,
Montana State University, Bozeman MT 59717

AMS classification scheme numbers: 92C20,94A05,94A12,94A34,62H30

Abstract. We discuss an analytical approach through which the neural symbols and corresponding stimulus space of a neuron or neural ensemble can be discovered simultaneously and quantitatively, making few assumptions about the nature of the code or relevant features. The basis for this approach is to conceptualize a neural coding scheme as a collection of stimulus-response classes akin to a dictionary or 'codebook', with each class corresponding to a spike pattern 'codeword' and its corresponding stimulus feature in the codebook. The neural codebook is derived by quantizing the neural responses into a small reproduction set, and optimizing the quantization to minimize an information-based distortion function. We apply this approach to the analysis of coding in sensory interneurons of a simple invertebrate sensory system. For a simple sensory characteristic (tuning curve), we demonstrate a case for which the classical definition of tuning does not describe adequately the performance of the studied cell. Considering a more involved sensory operation (sensory discrimination), we also show that, for some cells in this system, a significant amount of information is encoded in patterns of spikes that would not be discovered through analyses based on linear stimulus-response measures.

1. Introduction

What stimulus features are encoded in neural activity patterns? What aspects of the neural activity patterns encode that information? Considerable progress has been made by approaching these questions independently. However, independent treatment of these interconnected questions often introduces multiple assumptions that prevent their complete solution. How can we be sure we have discovered the specific features to which an ensemble of cells is sensitive unless we know, with complete certainty, the symbols they use to represent those features? And, vice versa, how can we be sure of the symbols unless we know, with certainty, what stimulus features are being represented by those symbols?

§ Research supported in part by NIH grants MH12159 (AGD) and MH57179 (JPM, ZA, AGD), and NSF grants DGE9972824(ZA,AEP) and MRI9871191.

We recently presented an analytical approach [8] that enables the simultaneous solution to these two interconnected questions. The basis for this approach is to conceptualize a neural coding scheme as a collection of stimulus-response classes, where each class consists of a set of stimuli and a synonymous set of neural responses. The stimulus-response classes form a structure akin to a dictionary or 'codebook', with each class corresponding to a neural response 'codeword' and its corresponding stimulus feature in the codebook. This analytical approach enables the derivation of this neural codebook, which in turn allows any sequence of spike patterns in a neural response to be 'deciphered' into the corresponding sequence of stimulus features that elicited those responses.

This new approach uses tools from information theory and quantization theory to perform the tasks above. Specifically, we quantize the neural responses to a small reproduction set and optimize the quantization to minimize an information-based distortion function. Fixing the size of the reproduction set produces an approximation of the coding scheme. The number of distinguishable codeword classes is related to the mutual information between stimulus and response. This analytical approach has several advantages over other current approaches:

(i) it yields the most informative approximation of the encoding scheme given the available data (i.e., it gives the lowest distortion, by preserving the most mutual information between stimulus and response classes),

(ii) the cost function, which is intrinsic to the problem, does not introduce implicit assumptions about the nature or linearity of the encoding scheme,

(iii) the maximum entropy quantizer does not introduce additional implicit constraints to the problem,

(iv) it incorporates an objective, quantitative scheme for refining the codebook as more stimulus-response data becomes available,

(v) it does not need repetitions of the stimulus under mild continuity assumptions, so the stimulus space may be investigated more thoroughly.

In the following sections, we first summarize the essential theoretical background from our recent work. Second, we present results related to the practical computational implementation of the core algorithms for the analysis of neurophysiological recordings. Third, we present further analysis and extensions to this theoretical approach that enable the analysis of more complex encoding schemes. Finally, we demonstrate the application of this approach through an analysis of coding in sensory interneurons of a simple invertebrate sensory system: For a simple sensory characteristic (tuning curve), we demonstrate a case for which the classical definition of tuning is actually inadequate given the analyzed performance of cell. Specifically, the response region with maximal firing rate, usually considered to be the "preferred direction" of this cell actually offers much worse estimate of the stimulus direction compared to neighboring activity ranges with lower overall activity. Considering a more involved sensory operation

(sensory discrimination), we also show that, for some cells in this system, a significant amount of information is encoded in patterns of spikes that would not be discovered through analyses based on linear stimulus-response measures. Specifically, short-interval spike doublets were found to code for stimulus features that differ significantly from the waveforms predicted by the linear combination single-spike-based Wiener/Volterra kernels offset by that doublet interval, through the 'stimulus reconstruction' technique.

2. Theoretical background and previous results

2.1. A model of neural processing

Any neural code must satisfy several conflicting demands. On one hand the organism must recognize certain natural object in repeated exposures. Failures on this level may endanger an animal's well-being: e.g., if a predator is misidentified as a con-specific mate. On this level, the response of the organism needs to be *deterministic*. On the other hand, distinct stimuli need not produce distinguishable neural responses, if such a regime is beneficial to the animal (for example, a wolf and a fox need not produce distinct responses in a rabbit, just the combined concept of "predator" may suffice.) Thus the representation, albeit possibly deterministic, need not be bijective. Lastly, the neural code must deal with uncertainty introduced by both external and internal noise sources. Therefore the neural responses are by necessity *stochastic* on fine scale. In these aspects 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. With this in mind we represent the input/output relationship present in a biological sensory system as a *communication system* [37].

We will therefore consider a neural encoding process within an appropriate probabilistic framework [1, 22]. The *input signal* X to a neuron (or neural ensemble) may be a sensory stimulus or may be the activity of another set of (pre-synaptic) neurons. We will consider the input signal to be produced by a source with a probability $p(x)$. The *output signal* Y generated by that neuron (or neural ensemble) in response to Y will be a spike train (or ensemble of spike trains.) We will consider the encoding of X into Y to be a map from one stochastic signal to the other. This stochastic map will be the *encoder* $q(y|x)$, which will model the operations of this neuronal layer. The *output signal* Y is induced by $q(y|x)$ by $p(y) = \sum_x q(y|x)p(x)$.

The view of the neural code, which is probabilistic on a fine scale but deterministic on a large scale, emerges naturally in the context of Information Theory [6]. The Noisy Channel Coding Theorem suggests that, in this context, relations between individual elements of the stimulus and response spaces are not the basic building elements of the system. Rather, the defining objects are relations between *classes* of stimulus-response pairs. There are about $2^{I(X;Y)}$ such equivalence classes (i.e., codeword classes). When restricted to codeword classes, the stimulus-response relation is almost bijective. That

is, with probability close to 1, elements of Y are assigned to elements of X in the same codeword class. This framework naturally deals with lack of bijectivity, by treating it as effective noise. We decode an output y as any of the inputs that belong to the same codeword class. Similarly, we consider the neural representation of an input x to be any of the outputs in the same codeword class. Stimuli from the same equivalence class are considered indistinguishable from each other, as are responses from within the same class.

2.2. Finding the codebook

Given this model of neural function, we would like to recover the codebook. In this context, this equates to identifying the joint stimulus-response classes that define the coding relation. The approach we use ([8]) is to quantize (i.e., cluster) the response space Y to a small reproduction space of finitely many abstract classes, Y_N. This method allows us to study coarse (i.e., small N) but highly informative models of a coding scheme, and then to automatically refine them when more data becomes available. This refinement is done by simply increasing the size of the reproduction, N.

The quality of a quantization is characterized by a distortion function [6]. In engineering applications, the distortion function is often chosen in a fairly arbitrary fashion [6, 13]. By concentrating on a pair of interacting systems (stimulus and responses), we can avoid part of this arbitrariness: 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 $I(X;Y_N)$, which is the mutual information between X and the reproduction Y_N. With that in mind, we postulate the following distortion function [8]:

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

Following examples from rate distortion theory [6, 34], this problem of optimal quantization can be formulated as a maximum entropy problem [8, 18]. 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. Within this framework, the minimum distortion problem is posed as a maximum quantization entropy problem with a distortion constraint:

$$\begin{aligned} &\max_{q(y_N|y)} H(Y_N|Y) && \text{constrained by} \\ &D_I(q(y_N|y)) \leq D_0 && \text{and} \\ &\sum_{y_N} q(y_N|y) = 1 \quad \forall y \in Y \end{aligned} \tag{2}$$

More details are presented in Section 3 and [8]. Recently the same problem was reformulated as one of optimal decoding [35], interpreting the quantizer $q(y_N|y)$ as a channel decoder.

The optimal quantizer $q(y_N|y)$ induces a coding scheme from $X \to Y_N$ by $p(y_N|x) = \sum_y q(y_N|y)p(y|x)$ 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)).

The elements of Y_N can be interpreted as the labels of the equivalence classes which we want to find. The quantizer $q(y_N|y)$ gives the probability of a response y belonging to an equivalence class y_N. We have shown in [12] that the optimal quantizer is generically deterministic, that is, the optimal probability $q(y_N|y)$ is 1 or 0 (see also Appendix B). In this case the responses associated with class y_N are $\mathcal{Y}_N = \{y|q(y_N|y) = 1\}$. The induced coding scheme from $X \to Y_N$ also induces the quantization $X \to X_N$ by associating the class $x_N \in X_N$ with the stimulus set

$$\mathcal{X}_N = \{x|p(y_N|x) \geq p(y_M|x) \text{ for all other classes } y_M\}.$$

Clearly, each $x \in X$ belongs to at least one class $\mathcal{X}_N$ and thus $X = \cup_N \mathcal{X}_N$. If the inequality above is strict for each x, then the classes are non-intersecting. Hence the resulting relation $p(y_N|x_N)$ is bijective. In general we expect the set $\{x|p(y_N|x) = p(y_M|x) \text{ for some } M, N\}$ to be of measure zero, and therefore the relation $p(y_N|x_N)$ is almost bijective. Hence, we recover an almost complete reproduction of the coding scheme as a relation between equivalence classes, which we outlined earlier.

Examples of the application of this method to synthetic data were presented in [8]. We reproduce a similar figure here (Figure 1) to demonstrate essential aspects of this approach.

A similar approach, termed "The Information Bottleneck", was developed previously by Tishby et. al. in [30, 48]. This approach has been used successfully in applications of text clustering [2, 30, 40, 42, 43] and astronomical observations [41]. The preliminary results from its application to neural systems [14, 36] are discussed in the context of Section 4.1. The two approaches are related, due to the common term in the cost functions, $I(X;Z)$ ($Z \equiv Y_N$ in our notation), and the use of probabilistic ("soft") clustering techniques. In fact, as noted in [28], both can be seen as extensions of Grenander's method of sieves [16] for estimating expectations of probabilistic quantities. However, a more detailed inspection of the methods reveals several distinctions.

A cursory inspection would note the minor difference in cost functions: in this work and in [8], influenced by [18, 34], we use the maximum entropy formulation (2). The formulation in [48] follows more closely rate distortion theory ([6]) and uses the cost function $F = I(Y;Z) - \beta I(X;Z)$.

In fact, this difference in cost functions leads to drastically different solution strategies. In [48], both parts of the optimization function are functional and must be optimized at the same time. The parameter β is considered a tradeoff parameter, which controls the tradeoff between quality of representation and amount of compression of the data. Thus the choice of β left to the user. The number of classes (reproduction size, N) are not fixed, and in the implementation in [48] more classes are introduced as

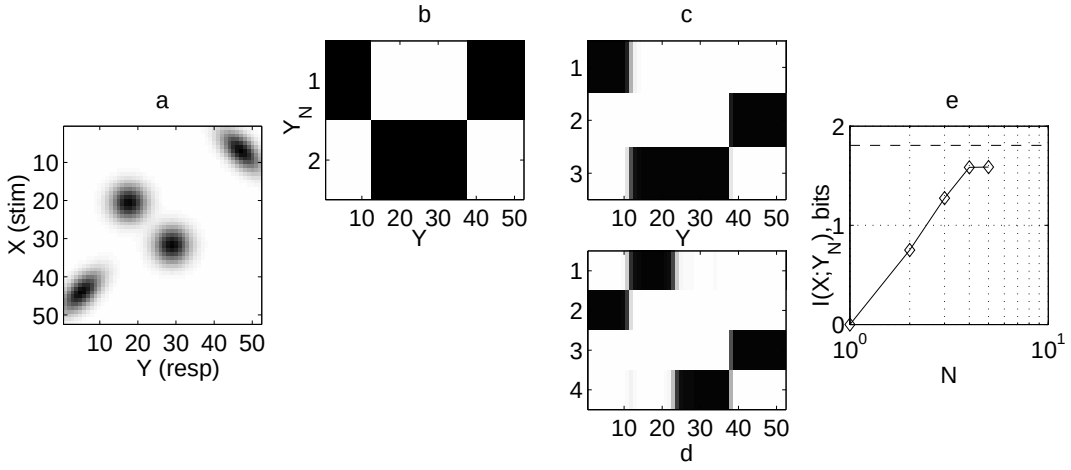

Figure 1. (a) A joint probability for a discrete relation between two random variables X (stimulus) and Y (response), with 52 elements each. (b–d) The optimal quantizers $q(y_N|y)$ for different numbers of classes. These panels represent the conditional probability $q(y_N|y)$ of a pattern y from a) (horizontal axis) belonging to class y_N (vertical axis). White represents zero, and black represents one. Intermediate values are represented by levels of gray. The behavior of the mutual information with increasing N can be seen in the log-linear plot (e). The dashed line is $I(X;Y)$, which is the least upper bound of $I(X;Y_N)$.

they are needed. In the implementation used for the agglomerative bottleneck approach [42], a greedy algorithm is used, in which points are grouped to larger classes as β descends from infinity to prescribed value of β_0.

In contrast, here and in [8] we are mostly concerned with minimizing the distortion function (1) for fixed number of classes, N. To that effect, in our approach the entropy H is simply a regularizer that helps with continuous optimization, since it has a unique maximum where the optimization starts. For the purposes of the numerical implementation, any strictly concave function has the same effect. For the same reason we fix the reproduction size N and always take β to infinity. As demonstrated here (Figure 1, Section 4), and in [8], the reproduction is refined (N increased) only if both a) there is sufficient data and b) the approximation is markedly improved. This regime allows us to prove (in [12], and Section 3.3) the important result that the solution to our problem is generically deterministic. We have used this result to design a combinatorial search algorithm (Section 3.1.2), which searches for optimal solution only among the deterministic solutions. We have found this algorithm very useful to our data analysis, however, it is inapplicable to the problem discussed in [48].

Because of the above differences, it is not obvious *a priori* whether the solutions of the two approaches are the same in the parameter regime where they can be compared (the same β and N).

3. Theoretical Results

3.1. Numerical algorithm for optimal quantization

In order to implement this analytical approach for the analysis of experimental data, we have devised several algorithms using various reformulations of the following problem:

$$\begin{aligned} &\max_{q(y_N|y)} H(Y_N|Y) \qquad \text{constrained by} \\ &D_I(q(y_N|y)) \le D_0, \qquad \sum_{y_N} q(y_N|y) = 1 \quad \forall y \in Y, \quad \text{and} \quad q(y_N|y) \ge 0. \end{aligned} \tag{3}$$

We have thus turned our problem to an optimization problem, similar to problems which appear in Rate Distortion Theory [6, 34]. Below we present two distinct approaches, which use different properties of the cost function and the feasible space to design efficient algorithm for solving (3). The first (3.1.1) involves a continuous optimization scheme which uses the probabilistic formulation of the problem. We present two implementations of this scheme. The second (3.1.2) makes use of special properties of the cost function in this space to replace the optimization with a combinatorial search in the set of vertices of the feasible region. Both formulation have their strengths and weaknesses. We use them interchangeably in further analytical work.

3.1.1. Annealing Using the method of Lagrange multipliers we can reformulate the optimization problem as finding the maximum of the cost function as

$$\begin{aligned} &\max_{q(y_N|y)} F(q(y_N|y)) \equiv \max_{q(y_N|y)} \Big(H(Y_N|Y) - \beta D_I(q(y_N|y)) \Big) \\ &\text{constrained by} \qquad q(y_N|y) \in \Delta, \end{aligned} \tag{4}$$

where $\Delta := \{q(y_N|y) \mid \sum_{y_N} q(y_N|y) = 1 \text{ and } q(y_N|y) \ge 0\}$. This construction removes the nonlinear parametric constraint from the problem and replaces it with a parametric search in $\beta = \beta(D_0)$. For small β the obvious optimal solution is the uniform solution $q(y_N|y) = 1/N$ [34]. It can be shown that as $\beta \to \infty$, the solution of the problem (4) converges to a solution of the problem (3) ([12]). Therefore we need to track the optimal solution from $\beta = 0$ to $\beta = \infty$. We can do this by incrementing 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 (4) at a fixed value of β. We have implemented two different algorithms to solve this problem.

The first algorithm is a Projected Newton Conjugate Gradient Line Search with an Augmented Lagrangian cost function [12]. This is a relatively standard numerical method for which the convergence property to a local maximum is assured.

The second 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 [8]. Solving this system produces the implicit equation (∇D_I depends on $q(y_N|y)$)

$$q(y_N|y) = \frac{e^{-\beta \frac{\nabla_q D_I}{p(y)}}}{\sum_{y_N} e^{-\beta \frac{\nabla_q D_I}{p(y)}}}. \tag{5}$$

Here ∇_q denotes the gradient operator with respect to the quantizer. The expression (5) can be iterated for a fixed value of β to obtain a solution for the optimization problem, starting from a particular initial state. In practice this scheme has demonstrated very fast convergence to a fixed point, and linear to quadratic dependence on the size of the problem. We are currently investigating the reasons for this beneficial behavior.

Tracking the solution from small values to large values of β can be also formulated as a *continuation* problem [3, 10], which finds efficiently the solution of (4) for the next step in β given the current solution. Instead of using the optimal solution at the last β value as the initial condition for the next step (as explained above), the initial condition (as well as the magnitude of the next β step) can be computed by taking a fixed step along the vector which is tangent to the curve defined by $\nabla_q F(q, \beta) \equiv 0$. A more extensive discussion of this technique can be found in [29].

3.1.2. Combinatorial search The special structure of our cost function and feasible region allows us to approach the optimization from a different perspective and design an optimization scheme which involves a combinatorial search in a discrete space of events. Applying standard results from information theory [6] we have shown in previous studies that the function D_I is concave in $q(y_N|y)$ ([8]). The domain $\Delta := \{q(y_N|y) \mid \sum_{y_N} q(y_N|y) = 1 \ \forall y \in Y \text{ and } \ q(y_N|y) \geq 0\}$ is a product of simplices and therefore convex. We have shown in [12] that these two facts imply that the optimal solution of (3) lies generically in a vertex of Δ (see Appendix B). Since the set of vertices may become large, we implemented a local search, bilinear in the sizes of the spaces Y and Y_N, which leads, under modest assumptions [12], to a local maximum of (3). Empirically, this search is very fast for small problem sizes (coarse quantizations with a small reproduction size N). However the increased computational cost makes it prohibitively slow for large reproductions. This drawback is offset by its massively parallel nature, which makes it a prime candidate for implementing on parallel computing environments.

3.2. Analysis of complex sensory stimuli

In general, we want to analyze the operation of any sensory system under conditions which are close to its natural set of conditions. This usually entails observing rich stimulus sets of high dimensionality. Characterizing such a relationship non-parametrically is very difficult, and usually requires prohibitively large datasets [19]. To cope with this regime, we choose to model the stimulus/response relationship. 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 upper bounds $\tilde{D}_I$ to the information distortion function D_I. In this case, when we minimize the upper bound, the actual value of D_I is also decreased, since $0 \leq D_I \leq \tilde{D}_I$. This also gives us a quantitative measure of the quality of a model: a model with smaller $\tilde{D}_I$ is better.

We start the modeling process by noting that D_I can be expressed as

$$D_I(Y, Y_N; X) = H(X) - H(X|Y) - (H(X) - H(X|Y_N)) \tag{6}$$

by using standard equalities from information theory [6]. The only term in (6) that depends on the quantizer $q(y_N|y)$ is $H(X|Y_N)$, so minimizing D_I is equivalent to minimizing

$$D_{eff} := H(X|Y_N).$$

Thus the models we need to consider should produce upper bounds of $H(X|Y_N)$. One way to achieve this is by constructing a maximum entropy model [18] of the corresponding probability.

We can further express $H(X|Y_N)$ as $H(X|Y_N) = E_{y_N} H(X|y_N)$ [7, 9], where each term $H(X|y_N)$ is the entropy of X conditioned on y_N being the observed response class. Here E_{y_N} denotes the expectation in Y_N. As a first attempt, we constrained the class conditional mean and covariance of the stimulus to the ones observed from data:

$$\begin{aligned} x_N &= \sum_x p(x|y_N)\; x \\ C_{X|y_N} &= \sum_x p(x|y_N)(x - x_N)^2. \end{aligned} \tag{7}$$

Here and later we use x^2 as a shorthand for $x\; x^T$ (direct product, non-commutative). The maximum entropy model under such constraints is a Gaussian $N(x_N, C_{X|y_N})$. with the estimated mean and covariance. Each entropy term is then bounded by

$$H(X|y_N) \le H_G(X|y_N) \equiv \frac{1}{2} \log(2\pi e)^{|X|} \det C_{X|y_N}$$

where $|X|$ is the dimensionality of the stimulus space X. This produces an upper bound $\tilde{D}_{eff}$ of D_{eff} by

$$D_{eff} \le \tilde{D}_{eff} \equiv E_{y_N} H_G(X|y_N) = E_{y_N} \frac{1}{2} \log(2\pi e)^{|X|} \det C_{X|y_N}. \tag{8}$$

The class conditioned covariance $C_{X|y_N}$ can be expressed explicitly as a function of the quantizer. Since $p(x|y_N) = \sum_y p(x|y)p(y|y_N)$, equation (7) implies

$$x_N = \sum_{xy} p(x|y)p(y|y_N)x = \sum_y p(y|y_N) \sum_x p(x|y)x = \sum_y p(y|y_N)x_y \tag{9}$$

and

$$\begin{aligned} C_{X|y_N} &= \sum_{xy} p(x|y)p(y|y_N)(x - x_N)^2 \\ &= \sum_y p(y|y_N) \sum_x p(x|y) \Big((x - x_y) + (x_y - x_N) \Big)^2 \\ &= \sum_y p(y|y_N) \Big(C_{X|y} + (x_y - x_N)^2 \Big) \\ &= \sum_y p(y|y_N) \Big(C_{X|y} + x_y^2 \Big) - \Big(\sum_y p(y|y_N) x_y \Big)^2. \end{aligned} \tag{10}$$

Since $p(y_N) = \sum_y q(y_N|y)p(y)$ and $p(y|y_N) = q(y_N|y)\frac{p(y)}{p(y_N)}$ by Bayes' theorem, the last expression (10) is a function of the quantizer (through $p(y|y_N)$). The parameters $(C_{X|y}, x_y)$ are independent of the quantizer and can be estimated from data. When substituted back in (8), this yields an explicit formula for the upper bound of the effective distortion

$$\tilde{D}_{eff} = \sum_{y_N} p(y_N)\frac{1}{2}\log(2\pi e)^{|X|}\det\Big[\sum_y p(y|y_N)\Big(C_{X|y}+x_y^2\Big)-\Big(\sum_y p(y|y_N)x_y\Big)^2\Big] \quad (11)$$

which can be used in place of D_I in the optimization scheme (3). The stimulus model obtained in this manner is effectively a Gaussian mixture model (GMM), with weights $p(y_N)$ and Gaussian parameters $(x_N, C_{X|y_N})$. Each element of the mixture is determined by L parameters for the class conditioned mean x_N and $L(L+1)/2$ parameters for the symmetric class conditioned covariance matrix $C_{X|y_N}$, for a total of $L(L+3)/2$ parameters per class. Here $L = |X|$ is the size of the input space. Hence the number of parameters for this model is $NL(L+3)/2$. The number of parameters grows linearly in the reproduction size N, but quadratically in L.

Reduced models

The full covariance model may quickly become impractical because of the large number of parameters. In practice we detect this by observing the error bars of the cost function estimates and stop if these increase too much. These complex models often gives us very good estimates of the cost function, but they limit the level of refinement we can achieve in practice. For that reason we also developed several reduced GMMs with fewer parameters: probabilistic Principal Components Analysis (PCA) model, spherical covariance model, common covariance model and common PCA model.

Probabilistic PCA model A model closest to the full Gaussian above is the probabilistic PCA (PPCA) model [4]. It is essentially the same GMM, but the covariance matrix is restricted in the following way: The largest K eigenvalues and corresponding principal component (PC) eigenvectors of $C_{X|y_N}$ are preserved, the remaining $L-K$ eigenvalues are forced to have the same value. This takes the original covariance structure and preserves the first K principal directions, while modeling the remaining directions with a spherical noise model. In effect, the class conditioned covariance $C_{X|y_N}$ is restricted to the class of block diagonal matrices of the form $C_{X|y_N} = [C_K \ \sigma_N^2 I_{L-K}]$, that is, it is block diagonal, with covariance C_K along the first K principal components, and covariance $\sigma_N^2 I$ along the rest. Each class is determined by L parameters for the mean x_N, $K(K+1)/2$ for the preserved covariance matrix C_K, and one additional parameter for σ_N along the orthogonal noise dimensions. The total number of parameters for this model is $N(L+K(K+1)/2+1)$, which is linear in N, linear in L, and quadratic in K. The number of preserved dimensions, K, is a free parameter for this model. The full Gaussian model can be seen as the limiting case $K = L$.

We now show that the PPCA model gives an upper bound of the function $\tilde{D}_I$ modeled by a full Gaussian. Recall that $\tilde{D}_I \propto \sum_N \log\det C_{X|y_N} = \sum_N \log\prod_s \sigma_{N,s}$ and denote by $\tilde{D}_I^{PCA}$ the distortion function obtained from the PCA model. We fix the last $L-K$ eigenvalues of the matrix $C_{X|y_N}$ to be the eigenvalue σ_{L-K}. In this case, $\sigma_s \leq \sigma_s^{PCA} = \sigma_{L-K}$ for all $s \geq L-K$ and $\sigma_s = \sigma_s^{PCA}$ for all $s \leq L-K$. Therefore $\prod_s \sigma_s \leq \prod_s \sigma_s^{PCA}$. Since the logarithm is a monotonically increasing function and this computation holds for all classes N the result

$$\tilde{D}_I \leq \tilde{D}_I^{PCA}$$

follows. Moreover, if $K_1 < K_2$ then

$$\tilde{D}_I^{PCA_{K_1}} \geq \tilde{D}_I^{PCA_{K_2}}.$$

Spherical model Another limiting case for the above model is the spherical Gaussian model, for which $K = 0$, that is, all principle directions are forced to have the same variance. In this case, $C_{X|y_N} = \sigma_N^2 I_N$ is proportional to the identity matrix I_N. Each class is determined by L parameters for x_N and one parameter for σ_N. The total number of parameters is $N(L+1)$, linear in N and L.

Common covariance model In the full covariance model and variants, every class-conditioned stimulus may have a different covariance matrix $C_{X|y_N}$. Here we impose a different type of restrictive structure on the input, by requiring that all class-conditioned stimuli have the same covariance structure

$$C_{X|N} = E_{y_N} C_{X|y_N} \tag{12}$$

This produces the following estimate of the cost function

$$\tilde{D}_{eff}^C = \frac{1}{2}\log(2\pi e)^{|X|}\det C_{X|N} \tag{13}$$

By a result of Ky Fan [11], the function $\log\det$ is concave, implying that $\tilde{D}_{eff} \leq \tilde{D}_{eff}^C$ and hence the common Gaussian model produces an upper bound to the full Gaussian bound $\tilde{D}_{eff}$, and to the cost function D_I as well. The number of parameters for each class is L for x_N, and there is a common covariance structure with $L(L+1)/2$ parameters, independent of the number of classes. The total number of parameters is $L(L+1)/2 + NL$, linear in N, quadratic in L, but the quadratic part is independent of N.

Common PPCA model Similarly to the previous PPCA model, we can restrict the common covariance $C_{X|N}$ even further, by imposing the PPCA structure on it: we preserve the K highest eigenvalues and corresponding principal directions, and force the remaining $L-K$ eigenvalues to have the same value, thus modeling the orthogonal subspace with a spherical Gaussian. The total number of parameters is $K(K+1)/2 + 1 + NL$.

Common spherical model For completeness we also present the $K = 0$ limiting case of the above model, which represents the whole variance structure with a single parameters, σ. In this case $C_{X|N} = \sigma^2 I$ and the total number of parameters is $1 + NL$, linear both in N and L. We have found this model to be too restrictive for the case of the cricket cercal sensory system and have not used it, except in few test cases. It may prove to be useful for other sensory systems.

3.3. Properties of the optimal solution

Solution is deterministic when models are used. By results of [12], the cost function

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

is concave in $q(y_N|y)$ for the most general model of the data, i.e. the full Gaussian model. More restrictive models, described in Section 3.2, are special cases of the general model. This means that some of the parameters estimated from the data are forced to have common values, but the overall structure of the model (i.e. Gaussian) remains the same. Therefore the results of [12] extend to reduced models and we have

Theorem 1 *The optimal solution of the problem (3) with D_I replaced by $\tilde{D}_I = I(X,Y) - \tilde{D}_{eff}$ with any of the four models of the input classes, lies generically on the vertex of Δ. In other words the optimal quantizer is generically deterministic when using any of the suggested models.*

Stability of the optimal solution. An important question which needs to be addressed is how stable our quantizer is with respect to small changes in data. These small changes can come from a variety of sources, among them recording errors, adaptation, round off errors when handling the data,etc. The function F which we optimize is a continuous function of the joint probability $p(x, y)$, and, in the case of function $\tilde{D}_{eff}$, continuous function of the estimated quantities $C_{X|y_N}$ and x_N. These estimates depend in turn continuously on the collected dataset. This means that small difference in collected data will yield function F only slightly different from the original function. The assignment of the optimal solution of the optimization problem (3) can be thought of as a continuous function from the space of possible values of estimated quantities $C_{X|y_N}$ and x_N to a discrete set of vertices of Δ. Every continuous function whose range is discrete must be locally constant. In other words, if values of the quantities $C_{X|y_N}$ and x_N change slightly, then the new optimal quantizer will be not only close to the old one, but actually the same. Clearly this is the strongest possible stability statement one can make. This stability property is another attractive feature of our approach.

4. Analysis of stimulus/response relations in the cricket cercal sensory system

The preparation we study is the cercal sensory system of the cricket. In the following sections, we briefly introduce this system, describe the experimental methods used to

collect the data, and then discuss the application of this new approach to analysis of coding by single sensory interneurons in this system.

Functional organization of the cercal system. This system mediates the detection and analysis of low velocity air currents in the cricket's immediate environment. This sensory system is capable of detecting the direction and dynamic properties of air currents with great accuracy and precision [15, 17, 20, 21, 25, 38, 39, 44, 46, 47], and can be thought of as a near-field, low-frequency extension of the animal's auditory system.

Receptor organs. The receptor organs for this modality are two antenna-like appendages called cerci at the rear of the abdomen. Each cercus is covered with approximately 1000 filiform mechanosensory hairs, like bristles on a bottle brush. Each hair is constrained to move along a single axis in the horizontal plane. As a result of this mechanical constraint, an air current of sufficient strength will deflect each hair from its rest position by an amount that is proportional to the cosine of the angle between the air current direction and the hairs movement axis. The 1000 hairs on each cercus are arrayed with their movement axes in diverse orientations within the horizontal plane, insuring that the relative movements of the ensemble of hairs will depend on the direction of the air current. The filiform hairs also display differential sensitivity to aspects of the dynamics of air displacements, including the frequency, velocity, and acceleration of air currents [27, 33].

Sensory receptor neurons. Each hair is innervated by a single spike-generating mechanosensory receptor neuron. These receptors display directional and dynamical sensitivities that are derived directly from the mechanical properties of the hairs [20, 23, 24, 33, 39, 38]. In particular, the amplitude of the response of each sensory receptor cell to any air current stimulus depends upon the direction of that stimulus, and these directional tuning curves of the receptor afferents are well-described by cosine functions [23]. The set of approximately 2000 receptors innervating these filiform hairs have frequency sensitivities spanning the range from about 5 Hz up to about 1000 Hz.

Primary sensory interneurons. The sensory afferents synapse with a group of approximately thirty local interneurons and approximately twenty identified projecting interneurons that send their axons to motor centers in the thorax and integrative centers in the brain. It is a subset of these projecting interneurons that we study here. Like the afferents, these interneurons are also sensitive to the direction and dynamics of air current stimuli [21, 25, 46, 47]. Stimulus-evoked neural responses have been measured in several projecting and local interneurons, using several different classes of air current stimuli [5, 25, 46, 47]. The stimuli that have been used range from simple unidirectional air currents to complex multi-directional, multi-frequency waveforms. Each of the interneurons studied so far has a unique set of directional and dynamic response characteristics. Previous studies have shown that these projecting interneurons

encode a significant quantity of information about the direction and velocity of low frequency air current stimuli with a linear rate code [5, 46, 47]. More recent studies demonstrate that there is also substantial amount of information in the spike trains that cannot be accounted for by a simple linear encoding scheme [8, 32]. Evidence suggests the implementation of an ensemble temporal encoding scheme in this system.

Experimental approach. Stimulus-response properties of sensory interneurons were measured using intracellular electrodes. Stimuli consisted of controlled air currents directed across the animals' bodies, and the responses consisted of the corresponding spike trains elicited by those air currents. The preparations were mounted within a miniature wind tunnel, which generated laminar air currents having precisely controlled direction and velocity parameters. Details of the dissection, stimulus generation, and electrophysiological recording procedures are presented in Appendix A.

4.1. Analysis of simple stimulus/response relations

In some cases the relationships between sensory stimuli and neural responses are simple enough to be captured relatively easily in a non-parametric manner. Here we demonstrate one such case. Consider the relation between stimulus direction and neural response in one of the cercal sensory interneurons. For this experiment, the set of stimulus waveforms consisted of a set of simple uni-directional air current "puffs" of identical shape and duration, presented sequentially from a variety of directions around the animal's body. Stimulus angle can be represented as a one-dimensional variable. The neuron's response in this simple case was represented by the number of elicited spikes in a 50ms window. The particular cell studied here shows pronounced directional selectivity, i.e., the number of spikes it fires depends on the direction from which the air puff originates. This directional tuning has been measured and analyzed in earlier studies, and shown to be well approximated by a truncated sine wave function [25, 47].

Using the approach presented here, this relation has been captured in the series of quantizations shown in Figure 2. Panel a) shows the histogram of the raw data, where the horizontal axis is the stimulus direction (here called 'Y, deg') and the vertical axis plots the distribution of the spike rates elicited at each stimulus direction. We use this as the estimate of the joint probability $p(x, y)$, which is used explicitly in estimating the information distortion D_I. Panels b) through e) show successive steps in the refinement of the quantizer as the number of classes increases from 2 to 5, respectively. The quantizer in panel e) corresponds to the point $N = 5$ on the plot in panel f), which shows the mutual information yielded by this scheme. This case, with N=5 different distinguishable classes, yields over 1.5 bits of information. This is close to the theoretical maximum, and corresponds closely to the value calculated in earlier studies based on an alternate approach [47]. It is interesting to note that what is usually referred to as the "preferred" direction of the cell (the stimulus direction eliciting maximum activity) is actually less-well discriminable than the neighboring directions. In particular, the

finest reproduction in Figure 2e discriminates the "preferred" direction of 45^o with a uncertainty of more than 60^o (class 1), while the direction near 120^o can be discriminated much better, with uncertainty of less than 20^o (class 4).

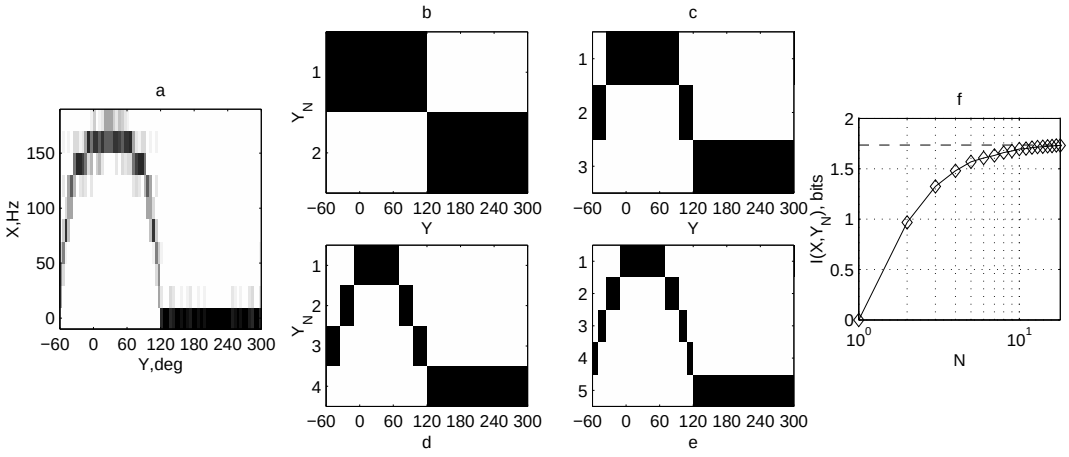

Figure 2. (a) The joint probability for the relation between stimulus angle (Y, degrees) and neural response (X, spikes / 50 msec). (b–e) The optimal quantizers $q(y_N|y)$ for different numbers of classes, from 2 classes in b) to 5 classes in e). These panels represent the conditional probability $q(y_N|y)$ that a stimulus y from a) (horizontal axis) belongs to class y_N (vertical axis). White represents a probability of zero, black represents a probability of one, and intermediate probabilities are represented by levels of gray. The behavior of the mutual information with increasing N can be seen in the log-linear plot (f). The dashed line is $I(X;Y)$.

This way of applying the method is similar in its use of data to the "direct method" of estimating mutual information [45]. The optimal quantization makes our approach less demanding than the direct method regarding the amount of data needed to achieve equivalent significance, since it quantizes the large response space to a relatively small reproduction space. In addition, the quantization also produces a simple functional relation between stimulus and response classes, while the direct method produces only an estimate of the mutual information. We do, however, obtain a lower bound of the mutual information, albeit with higher precision.

There are several drawbacks to attempting direct estimates of the joint probability, as used here in our analysis and also in applications of the direct method [14, 26, 31, 36, 45]. In principle, estimating the joint probability with a histogram is feasible only for relatively small stimulus and response spaces. For this reason a single-dimensional stimulus space (stimulus angle here, arm direction in [14]), and a small response space (number of spikes in a small temporal window, a number between zero in nine for Figure 2) are used here and in [14]. This allows the direct estimate of the joint probability, but limits us to relatively uninteresting cases. The other attempts to use estimates of the joint or conditional probabilities [26, 31, 36, 45] try to do so in relatively complex spaces (high dimensional white noise [36, 31, 45], or naturalistic stimuli [26, 31]). However, their reliance on repeated stimulus presentations means that just a tiny portion of the whole

input space can be sampled. This may bias the responses and corresponding estimates of the mutual information quite dramatically, even when enough data is avaliable for the estimates to be performed precisely.

4.2. Analysis of complex stimulus/response relations

We applied this analytical approach to characterize the encoding characteristics of single identified sensory interneurons to stimuli that were more complex and biologically relevant than simple unidirectional air puffs. The goal of the experiments and analyses were to discover (jointly) a) the dynamic stimulus waveform features encoded by the cells, and b) the spike train "codeword" patterns that encoded those features. Within the jargon of our approach, the goal was to discover the "codeword classes" for these cells. For this analysis, a variety of complex air current stimulus waveforms were used, ranging from bandlimited (5-400Hz) Gaussian white noise (GWN) to waveforms that combined stochastic and deterministic components that are suspected to be of more behavioral relevance [46]. Typically stimuli are not repeated on a single preparation, so the stimulus space can be sampled in more detail. For a more detailed description of the experimental procedures, see Appendix A.1.

After the responses are recorded, they are pre-processed to a form suitable for the algorithm. We use a binary representation of spikes, where at certain time a spike can either be present or absent. To be considered a pattern and further processed, a sequence of spikes must start with a spike and be preceded by a quite period of at least $D\ ms$. For a single neuron this is a binary string of length $T\ ms$. For an ensemble of neurons, this is a string of symbols of length T. Each symbol represents the activity of the ensemble as a binary string of labeled lines, as described in [19]. The parameters of the initial processing, D and T, may be varied to verify their effects on the final results. In the example shown on Figure 3, $D = 5ms$ and $T = 10ms$, which is a typical set of parameters.

Using the algorithms presented above, we proceeded to derive quantizers that identified synonymous classes of feature/spike-pattern pairs. In the illustrations below, the stimulus features are represented as the mean voltage waveforms, and voltage ranges of the stimulus that drove the air currents immediately preceding the elicited spike pattern codewords, and the response codewords are represented as the actual spike patterns that corresponded to those stimulus features. For visualization purposes we use a representation of spike patterns that is similar to a peristimulus time histogram (PSTH). We call this representation a Class Conditioned Time Histogram (CCTH.) The procedure is illustrated for a relatively simple case in Figure 4.2, for which the stimulus-response space has been quantized into three classes (i.e., N=3). The response space Y in panel a) consists of spike patterns y_i. Here each y_i is a spike doublet with a certain interspike interval (ISI). Each dot in the panel represents the time of occurrence of a single spike in a doublet. All doublets start with a spike at time 0, hence the vertical line along the left border at t=0. For this figure, the doublets have

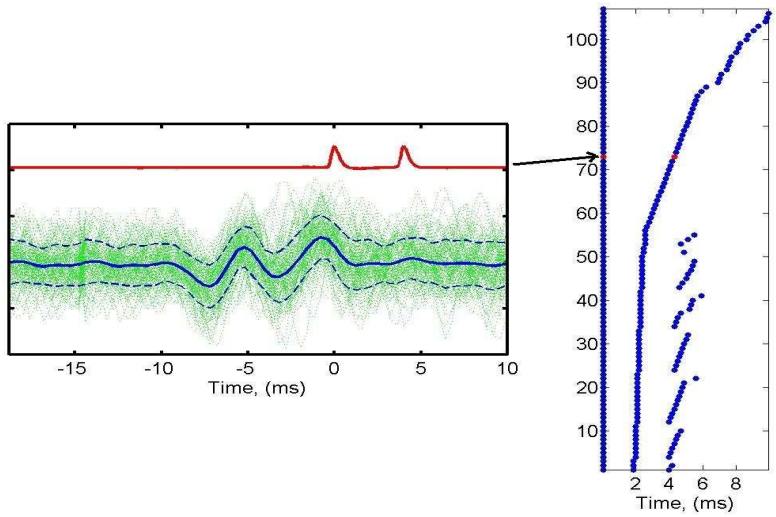

Figure 3. *Data extraction for subsequent quantization. The left panel show an intracellular trace (top, red). Here we observe the occurrence of a doublet with ISI* $\approx 5ms$*. Such a pattern occurs many times during a regular physiological recording. The lower panel show a set of stimuli that occurred in conjunction with this pattern (evoked this response). The actual stimuli are represented with green. The mean, and variance of the stimulus conditioned on this particular doublet occurring is shown in blue.*

The right panel shows **all** *spike sequences of length less than 10ms, which were observed in the course of the same experiment. The doublet seen on the left panel is highlighted with red. The horizontal axis is time, in ms, relative to the first spike in a pattern. The vertical axis shows pattern number. Here patterns are ordered in ISI, from small (bottom) to large (top). This ordering is irrelevant to the subsequent analysis.*

been arranged in order of descending inter-spike interval. (In the more general case, the spike codewords can be any arbitrary sequence of spikes in time, and might be distributed across several or many cells in an ensemble.) In b) and c) we see the two probabilities that completely define the quantization: $p(y)$ in b) and $q(y_N|y)$ in c). Using Bayes' Theorem, we obtain $p(y|y_N)$ from $p(y)$ and $q(y_N|y)$ (not shown). The final result in d) is the expectation $\sum_y y_i p(y_i|y_N)$. The pattern y_i can be considered as the conditional probability $p(t_j|y_i) = p($ spike occurs at time $t_j|$ the observed pattern is $y_i)$. This probability is 1 at times when a spike occurs and zero otherwise. In this case, panel d) can be interpreted as showing $p(t_j|y_N) = \sum_y p(t_j|y_i)p(y_i|y_N)$ - the conditional probability of a spike at time t_i given class y_N. The similarity to a PSTH is that we present the distribution of spikes in time, conditioned on the occurrence of an event. For the PSTH, the event is a particular stimulus. For this representation, the event is a certain response class, hence the name CCTH.

This representation has problems similar to the PSTH, since it assumes that

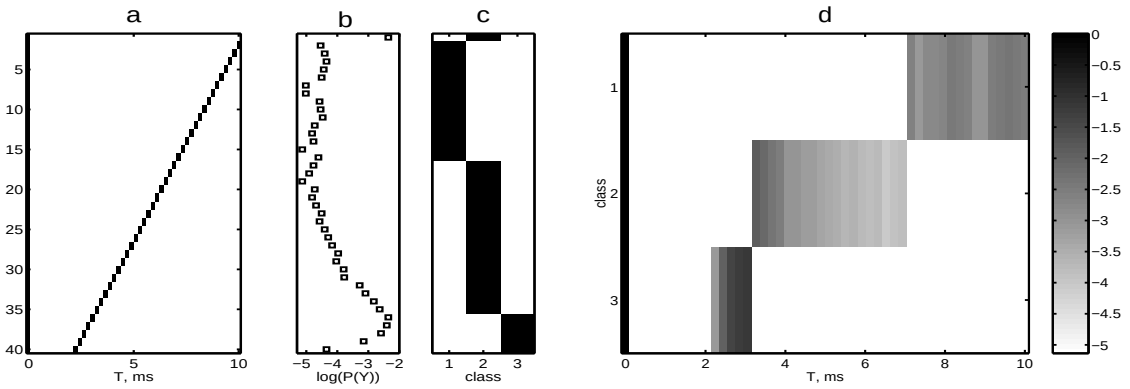

Figure 4. *The response space and its probabilistic representation. a) The response space for this example consists of patterns of doublets, which always start with a spike at time 0. The second spike follows with a delay between 2.5 ms (lower left) and more than 10 ms (upper right, only single spike visible). The patterns are ordered according to decreasing interspike interval. The vertical scale is the consecutive number of a spike pattern. b) The log probability of occurrence of a particular pattern, estimated by counting frequency of occurrence (histogram). c) A particular quantizer, as in Figure 2, groups several of the patterns in a) in a single class. In this case, all doublets with ISI* $\in$ *[6.9 10]*ms *are grouped in class 1, doublets with ISI* $\in$ *[3.1 6.9]*ms *and single spikes are grouped in class 2, and doublets with ISI*$<$ *3*ms *are in class 3. d) The CCTH of a spike at time T given the pattern is in a certain class. See details in text explaining the CCTH. We plot the conditional probability of spike occurrence vs. time for each pattern on a logarithmic scale, with black indicating a probability of one for the occurrence of a spike, and a lighter shade of gray representing a lower probability.*

spike at different times are independent†. Hence it cannot discriminate whether the spikes are from two different patterns (an *'or'* event, denoting combined patterns) from the possibility that there are two spikes from the same pattern (an *'and'* event, denoting a different pattern). However, since there is a refractory period, and since different patterns occur with different frequencies, it is relatively easy to discriminate the signature of a triplet from that of a doublet. For example, in Figures 5, 6C, the darker regions are due mostly to the second spike in a doublet, while the lighter regions preceding or following them are due to more rare triplets, for which one of the spikes is in the corresponding dark region.

How do we know when to stop the process of model refinement? The model of a coding scheme we use suggests that $I(X;Y_N) \propto \log N$ for $N \leq N_c \approx 2^{I(X;Y)}$ and $I(X;Y_N) \approx const$ for $N \geq N_c$. Since we in general don't know $I(X;Y)$, in practice we stop the refinement at an N_c for which the rate of change of D_I with N appears "empirically" to decrease dramatically. The estimate of $I(X;Y_{N_c})$ is the best lower

† Note that this was used only for visualization purposes and nowhere in our analysis do we assume that spike occurrences are independent events!

bound estimate of $I(X;Y)$ at this level of detail.

If there is not enough data to support such refinement, the algorithm is stopped earlier. The criterion we use in such a case is that the estimate of D_I 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. Thus we can recover a somewhat impoverished picture of the actual input/output relationship which can be refined automatically as more data becomes available, by increasing N and repeating the optimization procedure.

Below we present equivalent analyses of several other identifiable interneurons from the cricket's cercal sensory system to illustrate specific aspects of the procedures. In Figure 6 we present a full quantization sequence for one cell. For later examples, we present only the finest reproduction supported by data for the particular cell. We also suppress showing confidence intervals to the class conditioned means for reasons of visualization clarity. Details of the procedures and results are in the figure captions.

Figures 5 through 11 illustrate this analytical approach, and show results for several different cell classes. Figures 8, 9 and 11 also illustrate the different results obtained with this approach vs. the 'stimulus reconstruction' approach. Estimation of a stimulus waveforms with these class-conditioned means would be significantly more accurate than estimates based on a linear stimulus reconstruction kernel.

Figure 10 demonstrates the applicability of the method to analyzing multi-cell ensembles. The data for this case was actually obtained from intracellular recording of a single cell, to which we presented a GWN stimulus followed by the identical but sign-inverted stimulus. The responses of the cell to the second stimulus were taken to represent the activity of its complementary cell, which is sensitive to stimuli from the opposite direction. In this way we have a synthetic two-cell ensemble, in which the cells are forced to be conditionally independent (i.e., their activity is not related except through the common stimulus). Figure 10A demonstrates the appearance of this independence in the analysis: the first two classes contain isolated single spikes from one cell irrespective of the activity of the other cell. The class conditioned means (which are also the linear reconstruction kernels) also show that the cells are rectifying the stimulus.

An interesting case which needed a more detailed model is shown in Figure 11. In this case, the single class/single Gaussian model that we outlined in Section 3.2 was too restrictive, and we had to use a 2 component GMM to explain the stimulus conditioned on a single class. This is a minor extension of the quantization method. The stimulus reconstruction method cannot handle this case in principle, since the nonlinearity is not in the interaction between spikes, but in the generation of a single spike.

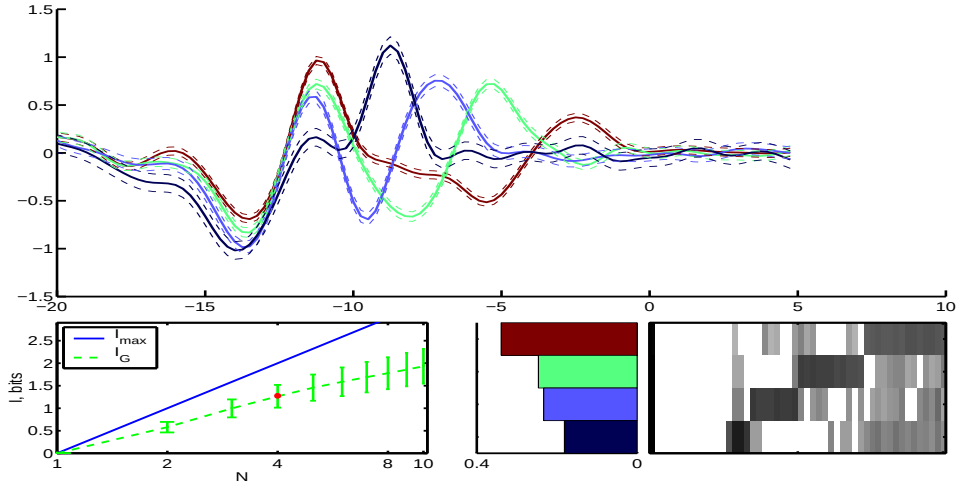

Figure 5. *A quantization with 4 classes. The top panel shows four class-conditioned mean stimulus waveforms, corresponding to the four spike pattern codewords derived through this quantization. The 4 mean waveforms are each plotted in the color of the corresponding class. The horizontal axis of this top plot denotes time, in ms, relative to the occurrence of the first spike in a class. That is, time 0 is the time at which the first spike on the codeword pattern occurred. The dashed lines denote 95% confidence intervals of the means, which depend on the reproduction size, N. The lower right panel plots the CCTH spike codewords for these four classes, as described in Figure 4.2. These patterns are aligned in time with the mean stimulus waveforms that elicited them, in the panel directly above. These are the classes of spike patterns that served as the basis for extracting the corresponding mean stimulus waveforms. Every class starts with a spike (line at 0ms). The amplitudes of the colored bars in the panel to the left of these CCTH plots show the relative proportion of spike patterns belonging to the different classes, as GMM priors (weights). These bars are color-coded to indicate the class-conditioned mean stimulus waveform to which the spike pattern to the right corresponds. This particular quantization groups the spike patterns roughly according to interspike intervals: The top class (brown) consists mostly of doublets with a second spike 7-10 ms after the initial spike (dark gray range to the right), and a few triplets (light gray bars in front), for which the third spike is in the same range. The 4th class (dark blue) consists mostly of short doublets, with a second spike 2.5-3.3 ms after the first spike, and a range of triplets with a third spike 6-10ms after the first spike. The lower left panel shows the estimate of the lower bound to the mutual information (green), and the absolute upper bound for the same level of quantization (blue, $\log_2 N$). The errorbars mark the uncertainty of the estimate, which depend on the reproduction size. The estimate for the current quantization level is denoted with a red marker.*

5. Discussion

The general goals of the research presented here were a) to develop algorithms through which the relevant stimulus space and the corresponding neural symbols of a neuron

or neural ensemble could be discovered simultaneously and quantitatively, making no assumptions about the nature of the code or relevant features, and b) to test the algorithms on an experimental preparation. The approach presented here makes a significant step in these directions. The essential basis for this approach is to conceptualize a neural coding scheme as a collection of stimulus-response classes akin to a dictionary or 'codebook', with each class corresponding to a neural response 'codeword' and its corresponding stimulus feature in the codebook. The analysis outlined here enables the derivation of this neural codebook, by quantizing the neural responses into a small reproduction set and optimizing the quantization to minimize an information-based distortion function.

The major advantage of this analytical approach over other current approaches is that it yields the most informative approximation of the encoding scheme given the available data. That is, it gives the representation with the lowest distortion, by preserving the most mutual information between stimulus and response classes. Moreover, the cost function (which is intrinsic to the problem) does not introduce implicit assumptions about the nature or linearity of the encoding scheme, nor does the maximum entropy quantizer introduce additional implicit constraints to the problem.

Many of the current analytical approaches for studying coding schemes can be seen as special cases of the method we present here. A rate code can be described as a deterministic quantization to the set of integers within an encoding window. The quantizer assigns all spike patterns with the same number of spikes to the same equivalence class. A spike latency code can be seen as a quantization to classes determined by the latency and jitter of the spike's timing. In this case, a stimulus feature is decoded as in the rate code case, based on which latency "class" a spike falls into. The metric space approach [49] uses an explicit cost (distortion) function to determine which different sequences are identical: they are equivalent if, according to the cost function, their difference is below a certain threshold. The cost function and identification threshold induce a deterministic quantization of the response space to a smaller reproduction space of equivalent classes.

We chose to formulate the problem explicitly in the language of information theory, so that we could use the powerful methods developed in this context for putting all these ideas in a unified framework. By doing so, we immediately realized one problem with this general approach: the distortion functions impose an assumed structure on the neural response (albeit a very natural one in the case of [49]) that may or may not be there in reality. Therein lies on important benefit of the method we present here: the information distortion cost function in (1) is intrinsic to the system, and does not introduce any additional assumptions about its function or structure. This benefit is somewhat decreased from the point at which we introduce models of the stimulus in Section 3.2, since now the models implicitly impose assumptions about the structure of the stimulus-response space. We partially resolve this issue by allowing for flexible models, that can partition the input space on small enough chunks, so that the distortions that the models introduce are small compared to the relevant structures in

the space (Figure 11).

Appendix A. Experimental protocols

Appendix A.1. Dissection and preparation of specimens

All experiments were performed on adult female crickets obtained from commercial suppliers (Bassett's Cricket Ranch, Visalia, CA, and Sunshine Mealworms, Silverton, OR). Specimens were selected that had undergone their final molt within the previous 24 h. The legs, wings and ovipositor were removed from each specimen, and a thin strip of cuticle was removed from the dorsal surface of the abdomen. After removal of the gut, the body cavity was rinsed and subsequently perfused with hypotonic saline. Hypotonicity facilitated microelectrode penetration of the ganglionic sheath.

The preparation was pinned to the center of a thin disc of silicone elastomer approximately 7 cm in diameter, located within the central arena of a air-current stimulation device, described below. Once the preparation was sealed and perfused with saline, the ganglion was placed on a small platform and gently raised from the ventral surface of the abdomen. This increased the accessibility of the ganglion to electrodes while at the same time improving the stability of electrode penetration by increasing surface tension on the ganglion.

Appendix A.2. Electrophysiological recording

Sharp intracellular electrodes were pulled from glass capillary tubes by a model P*80/PC electrode puller (Sutter Instrument Co.) The electrodes were filled with a mixture of 2% neurobiotin and 1 M KCl, and had resistances in the range of 30 to 50 megohms. During recordings the neurobiotin would diffuse into the nerve cell, allowing for subsequent staining and identification. Data were recorded using an NPI SEC-05L Intracellular amplifier and sampled at 10 kHz rate with a digital data acquisition system running on a Windows 2000 platform.

Appendix A.3. Stimulus generation

The cricket cercal sensory system is specialized to monitor air currents in the horizontal plane. All stimuli for these experiments were produced with a specially-designed and fabricated device that generated laminar air currents across the specimens' bodies. Air currents were generated by the controlled, coordinated movement of loudspeakers. The loudspeakers were mounted facing inward into an enclosed chamber that resembled a miniature multi-directional wind tunnel. The set of speakers were sent appropriate voltage signals to drive them in a "push-pull" manner to drive controlled, laminar air-current stimuli through an enclosed arena in the center of the chamber, where the cricket specimens were placed after dissection.

Stimulus waveforms were constructed prior to the experiment using Matlab. During experiments, the stimulus waveforms were sent out through a DAC to audio

amplifiers and then to the set of loudspeakers. Stimuli for determining directional selectivity consisted of half-cosine waves interspersed with silent periods, which created unidirectional air puffs. Additional stimuli consisted of 30 minute Gaussian white noise voltage waveforms, low passed below 500 Hz. Stimuli were either played along a single axis relative to the cricket, or were allowed to change angle at a maximum rate of 50 Hz.

Appendix B. The optimal solution is generically deterministic

Here we explain some technical background for the results in [12] which we cite in the text. Recall, that our goal is to solve the minimization problem

$$\mathbf{I} \qquad min_{q(y_N|y)\in\Delta} D_I(Y, Y_N)$$

where

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

and

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

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

$$I_N := I(X;Y_N)$$

in our optimization schemes.

In [12] we showed that D_I is a concave function of $q(y_N|y)$, that the domain Δ is convex and therefore the solution of problem (**I**) is either a vertex of Δ or, in a degenerate case, a product of simplices D_i which lie on the boundary of Δ. More precisely, for a fixed size of X and Y we let $\mathcal{P}$ to be the set of all joint probability distributions $p(X,Y)$. Since both X and Y are discrete spaces, the set $\mathcal{P}$ can be identified with the set of all $|X| \times |Y|$ matrices A with each row and column summing to one. This allows us to put on $\mathcal{P}$ a subspace topology from $R^{|X|\times|Y|}$. Then there is an open and dense set $\mathcal{D} \subset \mathcal{P}$ such that if $p(x,y) \in \mathcal{D}$, then the solution of the problem (**I**) is in the vertex of Δ. We say that this is a *generic case*. This means that, unless $p(x,y)$ has a special symmetry, the solution will be a vertex. The presence of noise in the system and the finite amount of data should break any symmetries, and so for all practical purposes one can assume that indeed the solution of (**I**) is a vertex of Δ.

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] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 2003. *to appear*.

[3] W. J. Beyn, A. Champneys, E. J.Doedel, Y. A. Kutznetsov, and B. Sandstede. *Handbook of Dynamical Systems III: Towards applications*. World Scientific, to appear.

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

[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. Natural time scales for neural encoding. *Neurocomputing*, 32-33:1027–1034, 2000.

[8] 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.

[9] A. G. Dimitrov, J. P. Miller, Z. Aldworth, and A. Parker. Spike pattern-based coding schemes in the cricket cercal sensory system. *Neurocomputing*, 44-46:373–379, 2002.

[10] E. J. Doedel, H. B. Keller, and J. P. Kernevez. Numerical analysis and control of bifurcation problems: (I) Bifurcation in finite dimensions. *International Journal of Bifurcation and Chaos*, 1:493–520, 1991.

[11] K. Fan. On a theorem of Weyl concerning the eignevalues of linear transformations II. *Proc. Natl. Acad. Sci. U.S.*, 36:31–35, 1950.

[12] T. Gedeon, A. E. Parker, and A. G. Dimitrov. Information distortion and neural coding. *Can. Math. Q.*, 2002. *(at press)*.

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

[14] A. Globerson, G. Chechick, N. Tishby, O. Steinberg, and E. Vaadia. Distributional clustering of movements based on neural responses. unpublished manuscript, 2001.

[15] Gnatzy and Heusslein. Digger wasp against crickets. I. Receptors involved in the antipredator strategies of the prey. *Naturwissenschaften*, 73:212–215, 1986.

[16] U. Grenander. *Abstract Inference*. Wiley, New York, 1981.

[17] H. G. Heinzel and M. Dambach. Traveling air vortex rings as potential communication signals in a cricket. *J. Comp. Physiolo. A.*, 160:79–88, 1987.

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

[19] 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.

[20] 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.

[21] M. Kanou and T. A. Shimozawa. Threshold analysis of cricket cercal interneurons by an alternating air-current stimulus. *J. Comp. Physiol. A*, 154, 1984.

[22] 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.

[23] M. Landolfa and G. A. Jacobs. Direction sensitivity of the filiform hair population of the cricket cercal system. *J. Comp. Physiol. A*, 1995.

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

[25] 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.

[26] S. Nirenberg, S. M. Carcieri, A. L. Jacobs, and P. E. Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411(7):698–701, 2001.

[27] L. C. Osborne. *Biomechanical Properties Underlying Sensory Processing in Mechanosensory Hairs in the Cricket Cercal Sensory System.* PhD thesis, University of California, Berkeley., 1997.

[28] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 2003. *at press.*

[29] A. E. Parker, T. Gedeon, A. G. Dimitrov, and B. Roosien. Annealing and the rate distortion problem. In *Advances in Neural Information Processing Systems*. 2002. *(to appear)*.

[30] F. Pereira, N. Z. Tishby, and L. Lee. Distributional clustering of english words. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio, 1993. Association for Computational Linguistics.

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

[32] J. C. Roddey, B. Girish, and J. P. Miller. Assessing the performance of neural encoding models in the presence of noise. *J. Comp. Neurosci.*, 8:95–112, 2000.

[33] J. C. Roddey and G. A. Jacobs. Information theoretic analysis of dynamical encoding by filiform mechanoreceptors in the cricket cercal system. *J. Neurophysiol.*, 75:1365–1376, 1996.

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

[35] I. Samengo. Information loss in an optimal maximum likelihood decoding. *Neural Computation*, 14:771–779, 2002.

[36] E. Schneidman, N. Slonim, N. Tishby, R. R. de Ruyter van Steveninck, and W. Bialek. Analyzing neural codes using the information bottleneck method. unpublished manusript, 2001.

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

[38] T. Shimozawa and M. Kanou. The aerodynamics and sensory physiology of a range fractionation in the cercal filiform sensilla of the cricket *gryllus bimaculatus*. *J. Comp. Physiol. A.*, 155:495–505, 1984.

[39] T. Shimozawa and M. Kanou. Varieties of filiform hairs: range fractionation by sensory afferents and cercal interneurons of a cricket. *J. Comp. Physiol. A.*, 155:485–493, 1984.

[40] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proceeding of SIGIR'02, 25th ACM international Conference on Research and Development of Information Retireval*, Tampere, Finland, 2002. ACM Press, New York, USA.

[41] N. Slonim, R. Somerville, N. Tishby, and O. Lahav. Objective classification of galaxy spectra using the information bottleneck method. *Mon. Not.R. Astron. Soc.*, 2001.

[42] N. Slonim and N. Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 617–623. MIT Press, 2000.

[43] N. Slonim and N. Tishby. The power of word clusters for text classification. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, Darmstadt, Germany, 2002.

[44] J. F. Stout, C. H. DeHaan, and R. W. McGhee. Attractiveness of the male acheta domesticus calling song to females. I. Dependence on each of the calling song features. *J. Comp. Physiol.*, 153:509–521, 1983.

[45] 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.

[46] 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.

[47] 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.

[48] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of The 37th annual Allerton conference on communication, control and computing*. University of Illinios, 1999.

[49] J. D. Victor and K. Purpura. Metric-space analysis of spike trains: theory, algorithms, and application. *Network: Computation in Neural Systems*, 8:127–164, 1997.

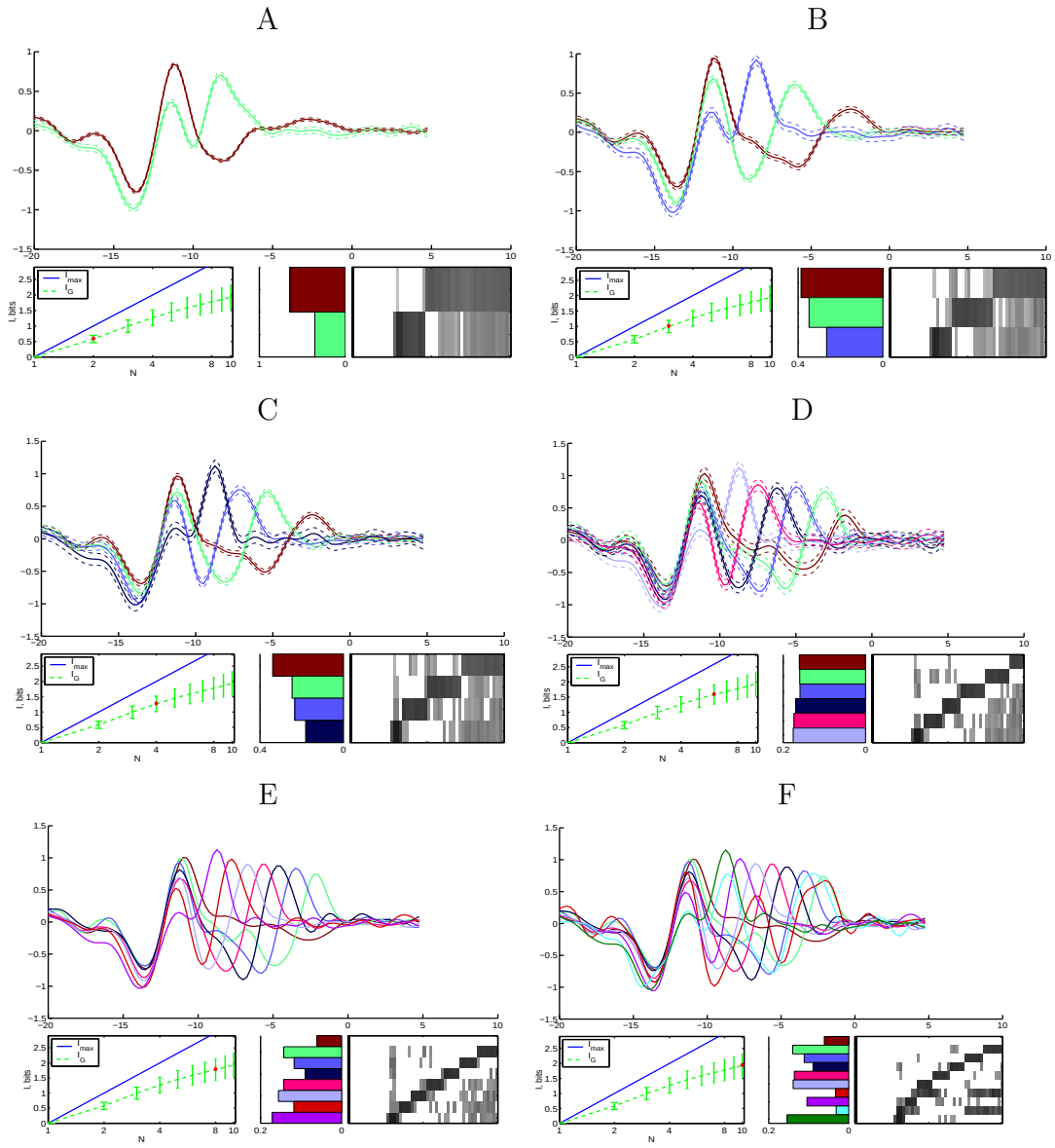

Figure 6. *Six steps in refining the quantizations. The format for each panel is equivalent to that described in Figure 5. A) The coarsest nontrivial quantization, containing only 2 classes. (B–E) Increased levels of refinement, from 3 (B) to 8 (E) classes. The structure evident in the initial coarse quantizations (Figure 5) remains unchanged: The patterns are grouped mostly according to the ISI of a doublet, with additional spikes appearing infrequently with a relatively uniform distribution (light gray region in the lower right corner, and light gray stripe at about 2.5 ms). F) A refinement in which the triplets were isolated in separate classes (class 2 and 4 from the bottom). All the uncertainty previously associated with the light gray range of the third spike is now almost completely collapsed in the triplet classes. The corresponding class conditioned stimulus reflect this class structure as well (light blue and red classes). The confidence ranges of (E,F) are not displayed in order to show the means more clearly. In general, the uncertainty increases with N.*

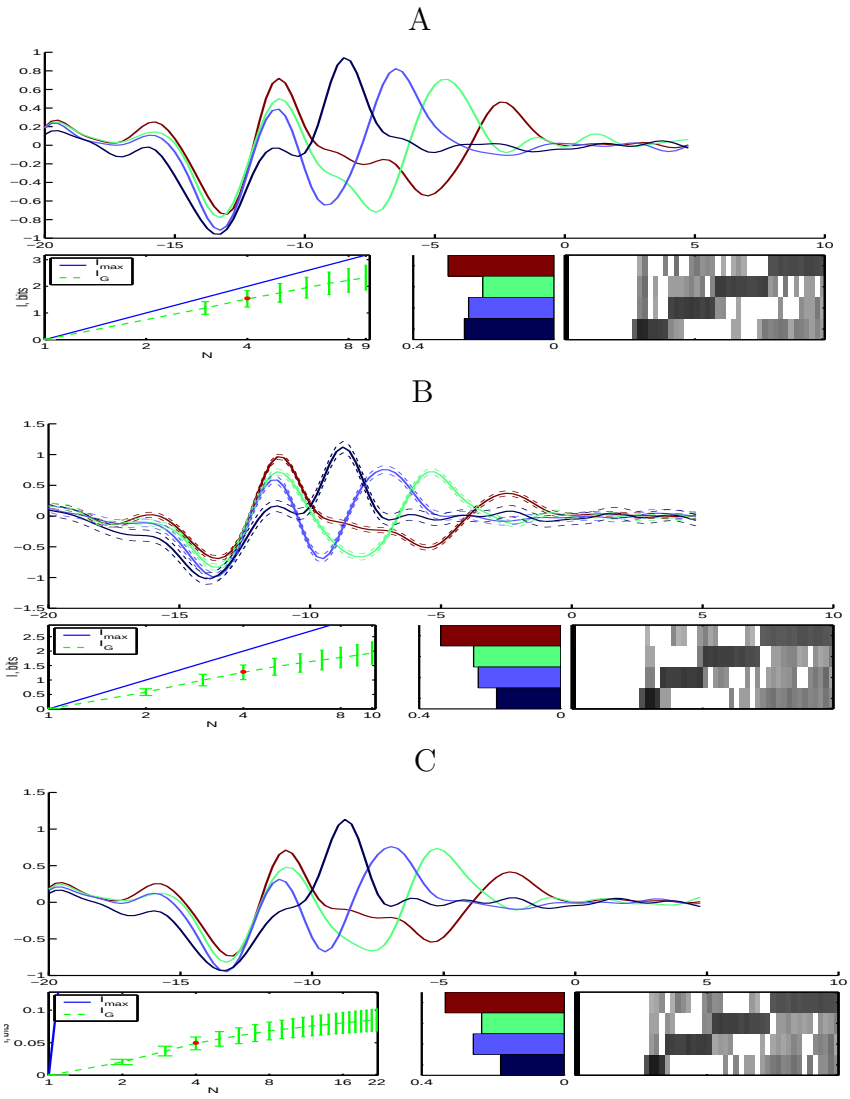

Figure 7. *A set of quantizations for a fixed reproduction size $N = 4$ derived with three different models. A)PPCA model with $K = 10$ dimensions preserved. B) common PPCA model with $K = 25$ dimensions preserved. C) Spherical Gaussian model. The resulting clustering is mostly consistent between the models (lower right panel in (A-C). The class conditioned means are also practically identical. The estimate of the mutual information (lower left panel in (A-C)) changes with the complexity of the model used. A) provides a relatively tight lower bound of I_N (green trace) closest to the absolute upper bound (blue trace), but the uncertainty grows rather rapidly (errorbars). B) produces a lower estimate of I_N (green trace), which is less uncertain (errorbars). The lower bound to I_N in C) is very poor (note different vertical scale). However, the estimate of this bound is very precise (again, different vertical scale makes the errorbars look big). The complexity of the models also affects the maximum reproduction size N that can be used. The more complex model in A) allows refinements with $N \approx 7 - 8$ classes. The intermediate model in B) allows refinements with $N \approx 12 - 14$ classes. The simplest model in C) allows additional refinements in excess of $N = 16$ classes. For this model, we can also observe the decreased rate of change of I_N with N around 8 classes.*

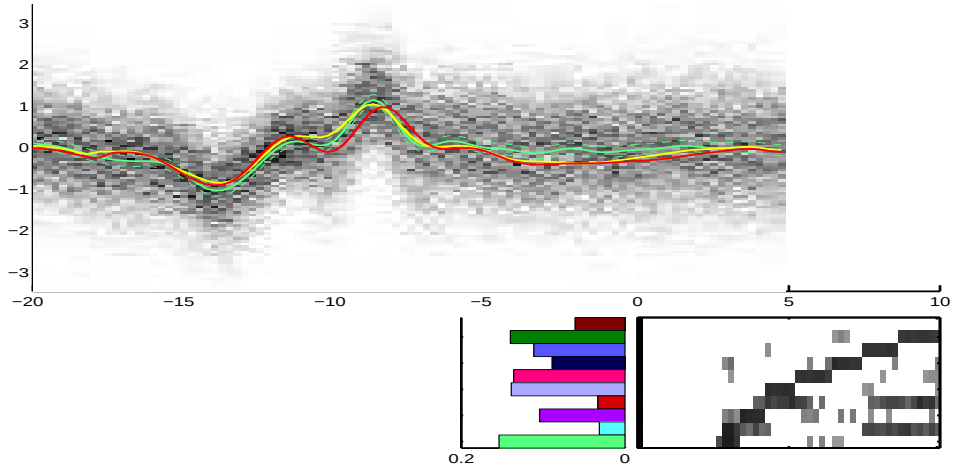

Figure 8. *Comparison to the stimulus reconstruction method. The bottom panel is replicated from Figure 6F to present a common reference. In the top panel, we show the 8th class-conditioned mean from Figure 6F (green trace), superimposed on a histogram of the stimulus (grayscale background). For each time t on the horizontal axis there is a histogram of the amplitudes V on the vertical axis at that time. This gives a visual representation of the variance around this class conditioned mean. The red trace shows the linear stimulus reconstruction for this class. It lies outside the confidence ranges of the class conditioned mean for $T \in [-10 \ -8]$ ms. Throughout the rest of the time it is inside the confidence limits of the green trace. The yellow trace was obtained by calculating the linear stimulus reconstruction when the set of second spikes were moved $0.4ms$ closer to the first spike. This indicates a mild sigmoidal nonlinearity – very high firing rates are reduced, presumably due to refractory effects. If we detect and correct for this nonlinearity, the behavior of the cell is well predicted by linear stimulus reconstruction. It is possible that second order reconstruction will be able to provide a better approximation for this particular nonlinearity*

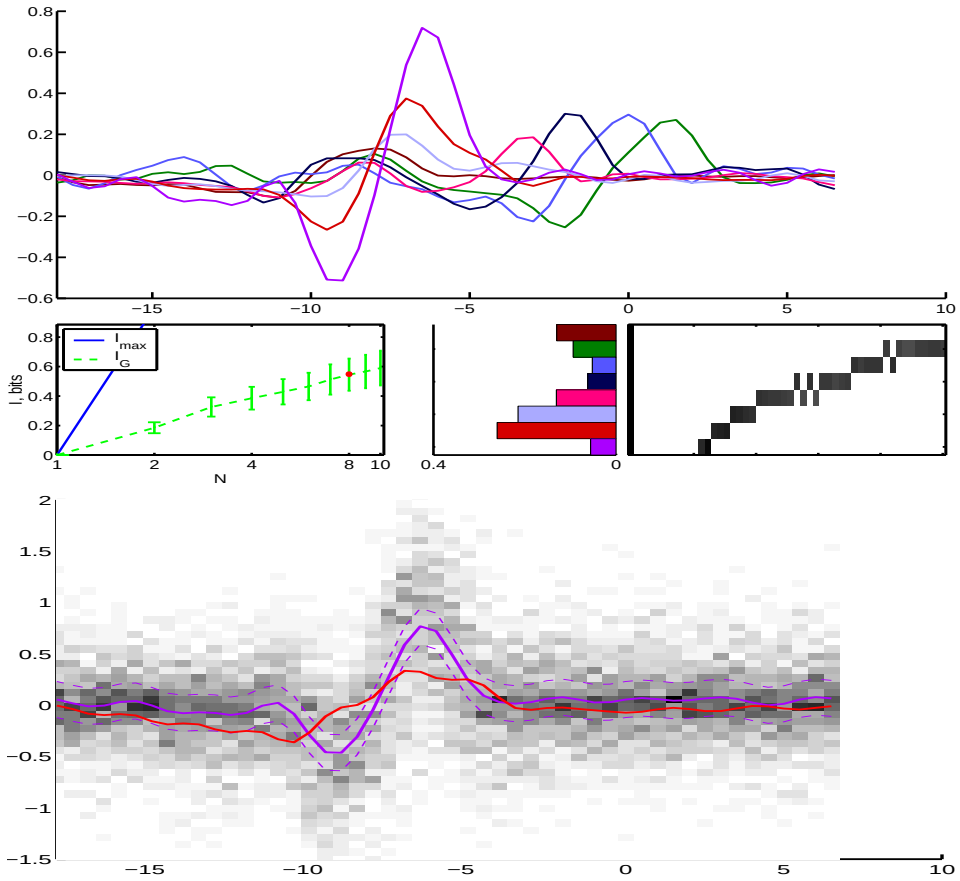

Figure 9. *Eight-class reproduction for another type of interneuron. This cell has a stimulus signature quite different from the previous cell: for short ISIs (last 3 classes) the class-conditioned means differ mainly in amplitude. For long ISIs, similarly to the previous cell, each spike seems to be associated with a biphasic sine-like input. The stimulus density around one of the class-conditioned means (purple) is shown in the lower panel. The linear stimulus reconstruction for this class is show in red. In this case, the stimulus reconstruction kernel is very different than any class-conditioned mean, and would yield a very poor stimulus estimate. Unlike the case in Figure 8, we were unable to find a simple nonlinearity that could account for this discrepancy. It is possible that adding second- or higher- order Volterra kernels in the stimulus reconstruction approach would account for this nonlinearity as well.*

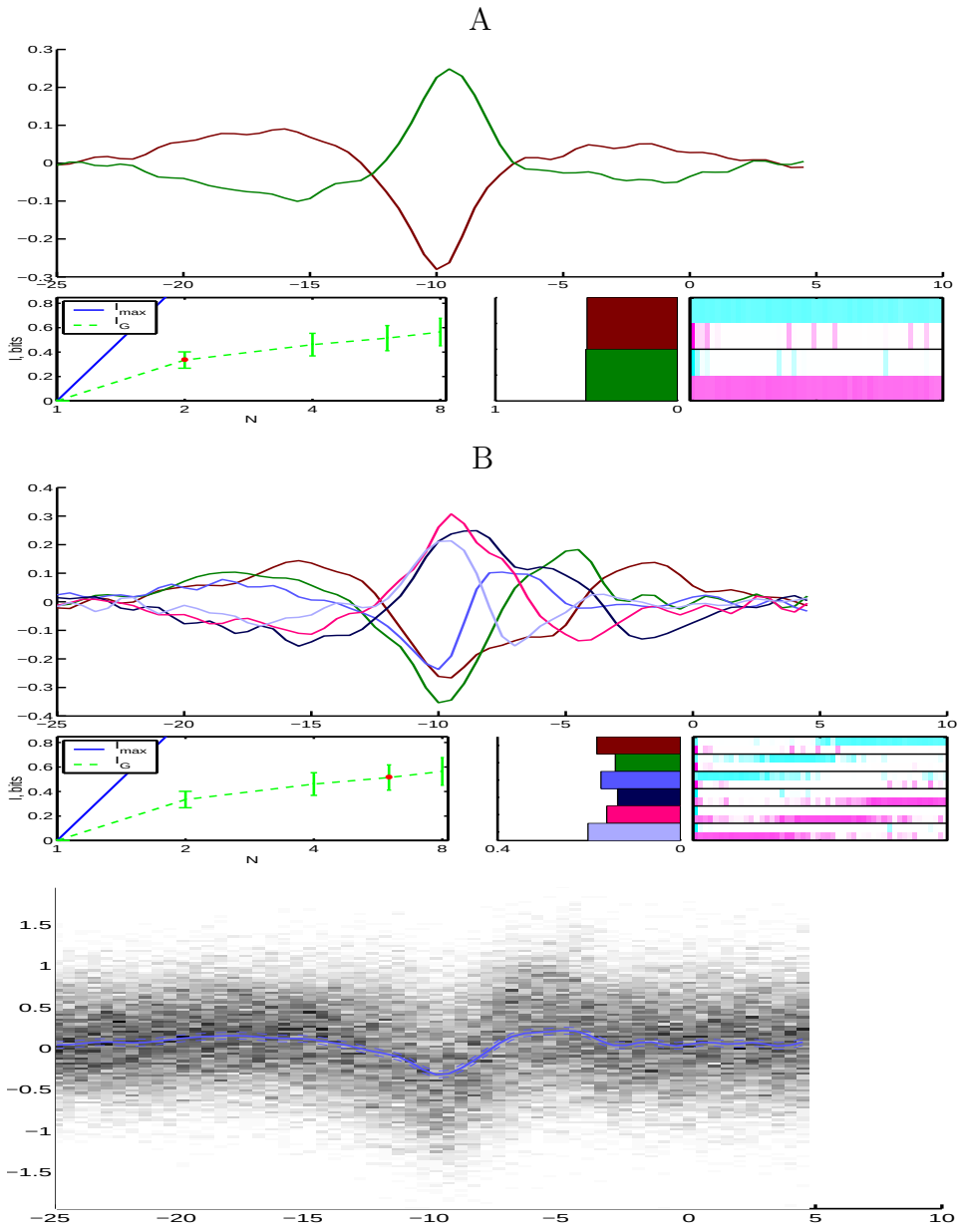

Figure 10. *Quantization of a synthetic two-cell ensemble (see text for explanation). A) The coarsest quantization to two classes recovers the single spike-conditioned average. The class-conditioned spike histogram in the lower right corner now has two traces for each class (labeled lines), showing the activity of the two cells in cyan and magenta, respectively. Class 1 consists of a single spike in cell 2 (magenta) and any activity from cell one (cyan). Class 2 consists of a single spike in cell 1 and any activity from cell 2. B) One of the finest quantizations supported by the available data. Classes continue isolate the activity of each individual cell. However, the activity of the other cell can now be discriminated better. For example, class one consists of cell 1 firing, followed by cell 2 firing 5-10 ms later, while class 2 has cell 1 firing and cell 2 firing 3-5 ms later. The class conditioned means follow roughly the same relation: there is a stimulus deviation associated with the first spike, and an anti-phase deviation associated with the second spike (for example, consider the green trace). The stimulus density around one of the classes is shown in the lower panel.*

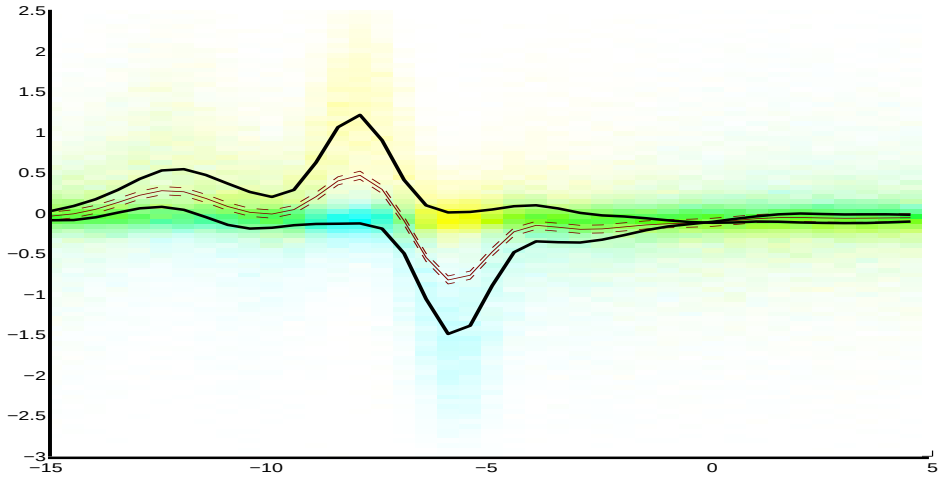

Figure 11. *A single class from another cell. This class consists of one spike at 0 ms with no activity before or after it for 10 ms. This dataset supported only the coarsest quantization to two classes, with everything but the isolated single spike combined in class two. There was not enough data for additional refinements. The single spike conditioned mean, which coincides with the linear reconstruction kernel, is shown in brown. A simple visual inspection of the data revealed that there were actually two distinct stimulus conditions which lead to a single spike. These are shown with yellow and blue densities (green denotes overlap of the densities). We used a two-component GMM for this class to explain the data (black traces overlayed over the corresponding densities). The stimulus reconstruction method cannot handle this case in principle, since the nonlinearity is not in the interaction between spikes, but in the generation of a single spike.*