
Analyzing neural codes using the information bottleneck method

Elad Schneidman,^{1,2} Noam Slonim,¹ Naftali Tishby,¹
Rob R. de Ruyter van Steveninck,³ William Bialek³

¹School of Computer Science and Engineering and the Center for Neural Computation,
Hebrew University, Jerusalem 91904, Israel

²Molecular Biology, Princeton University, Princeton, NJ 08544, USA

³NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA
{elads,noamm,tishby}@cs.huji.ac.il, {bialek,ruyter}@research.nj.nec.com

Abstract

A basic aspect of understanding the neural code of a neuron (or a neural system), is the ability to form a dictionary from the stimuli presented to the neuron (or the system) and the patterns of spikes that the neuron responds with. As neurons may respond unreliably to their stimuli, such a dictionary will be stochastic by nature. If the neuron responds to many different stimuli in a similar way (i.e. the number of stimulus features that the neuron 'cares about' is small), then the dictionary can be compressed, without a significant loss of its properties. Here we apply the *agglomerative information bottleneck* algorithm to study the properties of the dictionary (and neural code) of the identified H1 neuron in the fly visual system. We find that the neural code dictionaries of different flies are highly compressible, suggesting that a small number of features are the key components of the H1 neural code. We also compare the encoded features of the different flies and find similar general structure, but differences in the details.

1 Introduction

The problem of neural coding is to understand how neurons use patterns of action potentials (spike trains) to represent and transmit information [1]. Ideally, one would like to form some kind of a dictionary from the stimuli presented to a neuron (or the animal) to the resulting spike trains (and vice versa). Such a dictionary must be probabilistic [2], since when presented with the same stimulus repeatedly, neurons (may) respond with unreliable spike trains, both *in vitro* [3] and *in vivo* [4, 5]. Thus, the neuronal encoding of a stimulus s by the spike train $\{t_i\}$, can be described by the conditional distribution $p(\{t_i\}|s)$. Therefore, the dictionary between stimuli and response would contain the set of all possible stimuli and their corresponding conditional response distributions. This set of distributions, and the relations between them, capture the fundamental characteristics of the neural code. For example, it describes the set of codewords which the neuron uses, the effects of noise in the coding process and how much information about a stimulus would be

encoded by the spike trains [1, 6]. Similarly, decoding is described by the response conditional ensemble [2] $p(s|\{t_i\})$.

In general, the dictionary from stimulus segments to conditional responses and back may be very large (due to the size of possible stimuli space). However, if many of the conditional response distributions are similar to one another (which means that from the neuron’s point of view, these stimuli are similar), then it should be possible to construct a reduced version of the dictionary which would be a good approximation to the full dictionary. I.e., if there is a clear clustering of the response distributions, then replacing the actual response of the neuron with that of the average response of the cluster it belongs to, would result in a small loss of encoded information about the stimulus. This notion of clustering responses is complementary to the idea of dimensionality reduction in the space of stimuli [2]. An obvious question regarding the nature of the neural code is then: how large does the dictionary need to be, to capture (most of) the relation between the stimuli to a neuron and its responses?

The recently introduced *information bottleneck method* [7], provides a general information theoretic framework for building *compact* dictionaries of this sort. In this approach, given the joint distribution of two random variables $p(x, y)$, one looks for a compact representation of x , which preserves as much information as possible about y . In the context of the stimuli and the neural response, we would seek a clustering of stimulus features (or times during the stimulus) that maximizes the information about the resulting spike trains.

Here, we apply the *agglomerative information bottleneck* algorithm [8], (a “hard” clustering approximation to the general method) to study the nature of the neural code of the H1 motion sensitive neuron in the fly’s visual system [9]. We find that a small number of response clusters captures most of the information about the stimulus (compared with the information carried by the full set of responses). The average stimulus associated with each cluster is different, and these cluster-triggered averages may be interpreted as (first order approximations of) the different features of the stimulus which H1 encodes. We also present the average of the response distributions that clustered together, which reflect the nature of “noise model” of the coding mechanism. Finally, we compare the clustering results for different flies, and find them to differ in their dictionary content and in the details of their features.

2 Experimental setup and the construction of the neural code dictionary

Flies were presented repeatedly with the same 40 second long movie, during which, the responses of their H1 neuron were recorded. The stimulus presented to each of the flies was a rigidly moving image of random black and white bars pattern; pattern position was defined by a pseudorandom sequence, simulating diffusive random walk.¹ The spike trains are discretized into time bins of size $\Delta t = 2$ msec. At this resolution there are almost never two spikes in a single bin, so we can think of the neural responses as binary strings. We examine the responses in windows of length T , and so the individual responses become binary words W with $T/\Delta t$ ‘letters’. Any choice of specific T and δt values is arbitrary – in this paper we use $T = 14$ msec, as previous work has shown it to be long enough words to reflect the temporal structure of the spike train (see also discussion) but short enough that we can sample the relevant probability distributions.

¹Recordings were made from the H1 neuron of immobilized flies, using standard methods. The flies used were freshly caught female *Calliphora*; average intensity of the stimulus was $\bar{I} \approx 100$ mW/(m² · sr).

As the movie is presented repeatedly, the responses of H1 to the stimulus, can be described using the conditional probability distribution of binary words at specific times along the stimulus, $p(W|t)$. If the movie (which is the result of some underlying statistical source) is sufficiently long then the set of time conditional distributions $p(W|t)$, is a fair replacement of the stimulus segment conditioned distribution $p(W|s_t)$ (where s_t is the stimulus portion preceding time t) [6]. Figure 1 shows a portion of the spike trains which H1 of one of the flies responds with, and the construction of the conditional word distribution.

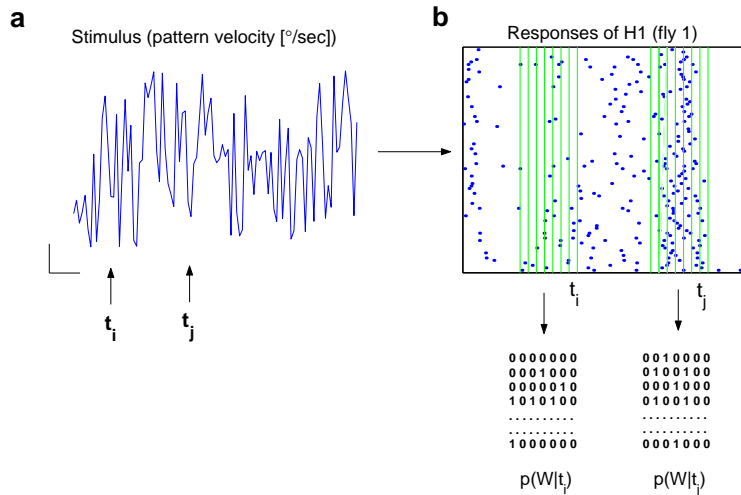


Figure 1: **A sample of H1 spike trains and estimation of $p(W|t)$.** (a) All flies view the same random vertical bar pattern moving across their visual field with a time dependent velocity, part of which is shown on the left panel (scale bars mark 10 msec horizontally and $50^\circ/\text{sec}$ vertically). In the experiment, a 40 sec waveform is presented repeatedly, 90 times. (b) The responses of one of the flies (fly 1) to part of the stimulus shown in a (top). Example of construction of the time dependent word distributions $p(W|t)$, (with 7-letter words) is shown for two times, t_i and t_j , with arrows marking the spots both in the stimulus and the responses (bottom panel). The fly's spike trains are divided into contiguous 2 msec bins (vertical lines), and the value of each bin is given by the number of spikes in the bin (for each of the stimulus repeats separately).

The full version of the stimulus-response dictionary would then be the whole set of stimulus segments (or times) and their corresponding word distributions. In order to investigate how 'compressible' this dictionary is, we would like to cluster the conditional word distributions, and quantify the effect of compression on the quality of the dictionary (the compressed dictionary would be from stimuli to average cluster responses). While one could apply a wide range of clustering algorithms and dictionary quality measures to this problem, the question at hand falls within the scope of a more general question about the relation between two random variables or the relation between the inputs and outputs of a system.

3 The (agglomerative) information bottleneck and the clustering of neural responses

Finding the features of the input X to a system, which are relevant for the prediction of its output Y , has been shown to have a natural information theoretic formulation, termed the *information bottleneck method* [7]: find a compact representation

of X , denoted \tilde{X} , such that a fixed amount of meaningful information about Y , $I(\tilde{X}; Y)$, while minimizing the mutual information between X and \tilde{X} , $I(X; \tilde{X})$ (i.e. maximizing the compression). As shown in [7] this formulation amounts to the minimization of the *Lagrangian*:

$$L[p(\tilde{x}|x)] = I(X; \tilde{X}) - \beta I(\tilde{X}; Y) \quad (1)$$

with respect to $p(\tilde{x}|x)$, where β is a positive *Lagrange multiplier*. This minimization yields a set of self consistent equations for $p(\tilde{x}|x)$, $p(y|\tilde{x})$ and $p(\tilde{x})$, which can be solved iteratively.

The *agglomerative information bottleneck* algorithm [8], introduces a simple bottom-up “hard” clustering approach to the information bottleneck problem. Using a greedy agglomerative procedure it finds a set of clusters \tilde{X} that directly maximizes (locally) the information $I(\tilde{X}; Y)$ for any prescribed number of clusters².

In the current context, this formulation translates to seeking a set of stimuli features \tilde{s} (clusters of times along the stimulus), such that the information between these features and the neural responses is maximized. In the framework of the agglomerative algorithm this means that we iteratively group together the times (with respect to the stimulus) on which the distribution of responses were similar. We start with the full set of conditional distributions and assign each of them to a unique cluster. Thus $p(W|t_i)$ is assigned to cluster c_i , and the cluster’s probability distribution is given by $p_{c_i}(W) = p(W|t_i)$. The size of each of the clusters is $\lambda_i = 1$. We measure the pairwise distance between all clusters using the *Jensen-Shannon* divergence [11], given by³,

$$d_{ij} = D_{JS}[p_{c_i}(W)||p_{c_j}(W)] \quad (2)$$

and find the nearest clusters. We merge these two clusters into a new cluster c_k , with $p_{c_k}(W) = \frac{1}{\lambda_i + \lambda_j}(\lambda_i p_{c_i}(W) + \lambda_j p_{c_j}(W))$, and the new cluster size is $\lambda_k = \lambda_i + \lambda_j$. We recalculate the set of pairwise distances between the clusters (only the distance between the new cluster and all the remaining clusters needs to be recalculated) and iteratively merge the nearest ones as before. This process proceeds until we are left with a single cluster.

As shown in [8], for any given number of clusters, this algorithm approximates the full bottleneck problem, searching for clusterings of the times along the stimulus that (locally) maximize the amount of mutual information between the clusters and the neural responses. Alternatively, one can interpret this as the information about the stimulus, that the neuron would convey if for all the times that clustered together, the neuron responded to the stimulus according to the centroid distribution of the cluster. For every number of clusters, we compare this information to the information that the words of length T conveys about the stimulus (and vice versa), $I_T(\text{stimulus}; \text{spike times})$, derived from the full set of conditional distributions [6].

Figure 2 shows the fraction of information contained about the stimulus in such a compact representation of the responses, calculated for the responses of 3 different flies (separately), using $T = 14 \text{ msec}$ and $\Delta t = 2 \text{ msec}$. We note that these three flies differ considerably in the amount of information they encode about the stimulus (as well as in their average firing rate). Evidently, a rather small number of clusters

²This is the limit of the information bottleneck for $\beta \rightarrow \infty$.

³The Jensen-Shannon divergence between two probability distributions $p(x)$ and $q(x)$ is given by $D_{JS}[p(x)||q(x)] = \lambda D_{KL}(p(x)||r(x)) + (1 - \lambda) D_{KL}(q(x)||r(x))$ where λ_p is the relative weight of the p distribution (Similarly for q) and D_{KL} is the *Kullback-Leibler* divergence [10], defined as $D_{KL}[p(x)||r(x)] = \sum_x p(x) \log \frac{p(x)}{r(x)}$

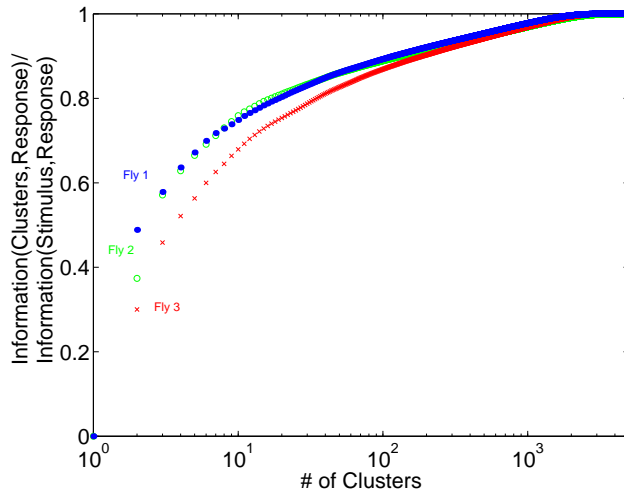


Figure 2: **Information bottleneck curve for 3 different flies.** The information that the clusters convey about the neural response (normalized by the information that the full set of $p(W|t)$ conveys about the stimulus), is shown as a function of the number of clusters that the algorithm uses (see text for details). We have used 5000 time points (bins) from the experiment as the initial set of $\{p(W|t)\}$. The time points were chosen randomly from the experiment; different choices of the time points give similar results (not shown). The information curve is shown for 3 different flies (same time points used for all the flies). Fly 1 and 2 have a similar firing rate (and information rates), whereas fly 3 has roughly twice as high firing rate (and information rates).

captures most of the information that the full spike trains holds about the stimulus – suggesting that the number of stimulus features that H1 encodes is relatively small. For example, 5 clusters (which is a reduction factor of 3 orders of magnitude!), preserve 50% to 70% of the original information is preserved (exact value depends on the specific fly).

4 Extracting features of the neural code and comparing neural codes

By our clustering construction, all of the times along the stimulus that belong to the same cluster have similar response distributions. If all these distributions really originate from one underlying distribution, then the average distribution describes the noise model of the neuron. Figure 3 shows the centroids of the 5 clusters (for one of the flies), i.e. the average $p(W|t)$ of each of the clusters.

We also ask what are the common features of the stimulus segments preceding the set of word distributions (i.e. times) which are assigned to the same cluster. Generalizing the spike (or word) triggered average stimulus [1], we calculate the average over the stimulus segments preceding the times that clustered together. These average stimuli waveform (“cluster triggered average”) for each of the clusters is shown for fly 1 and the case of 5 clusters in the left panel of Figure 4.

Repeating the same analysis for the other 2 flies whose information curve is shown in Figure 2, we compared the clustering results of the 3 flies. For the case of 5 clusters for each of the flies, we identify corresponding clusters as ones which have the largest number of shared times (since the flies watched the same movie). The centroid word

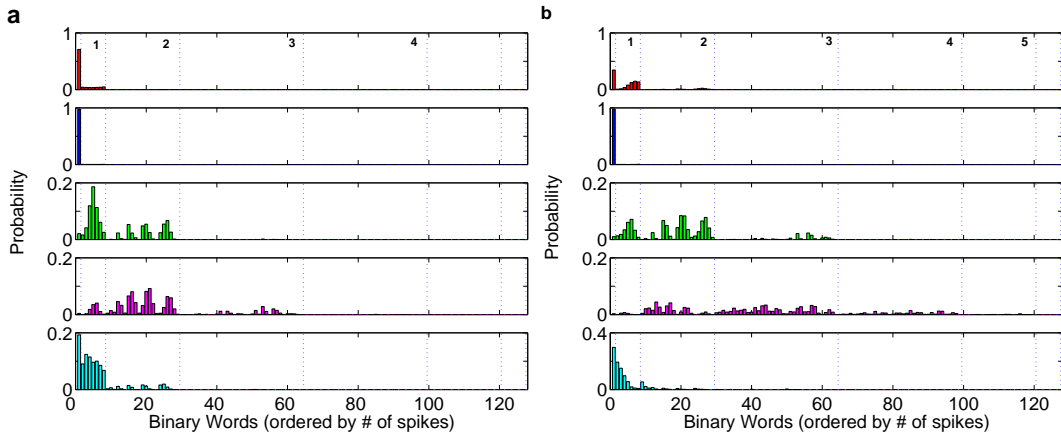


Figure 3: **Average $p(w|t)$ for the clusters of 2 of the flies.** (a) Average $p(W|t)$ for each of the clusters of fly 1, is shown for the case of 5 clusters. Words are ordered by the number of spikes ; vertical broken lines mark the boundary between words with different number of spikes (shown in the top panel). (b) Average $p(W|t)$ for fly 3. (We note that the resulting distributions are not very sensitive to different randomized choice of the 5000 time points.)

distributions have a similar general structure (a large cluster of no spiking, another with sporadic single spike etc.), but differ in the details of these distributions (not shown). Right panel in Figure 4 shows the cluster triggered averages of the 3 flies for 2 corresponding clusters, reflecting similar general structure, but large difference in the details.

5 Discussion

We have found that, for the fly H1 neuron, the dictionary that described the neural coding of complex dynamic stimuli is highly compressible. Correspondingly, H1 must be selective for a small number of stimulus features in the high dimensional stimulus space. Moreover, while the neural codes of different flies may differ considerably in their firing rates and information rates, the compressibility is almost

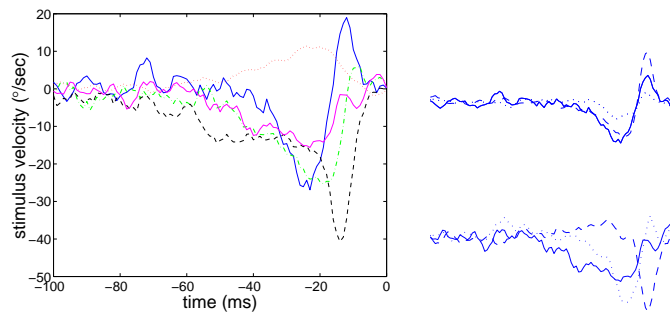


Figure 4: **Cluster-triggered stimulus averages.** For each of the 5 clusters in Fig. 3a, we compute the average stimulus waveform which preceded all the times that clustered together (left panel). Comparison of the waveforms of 2 corresponding clusters of the 3 flies (right panel). Waveforms were lowpass filtered for presentation clarity.

a universal function (Fig. 2). Finally, we find that there is a considerable overlap between the features of the different flies, but that the flies do differ in the specific features details.

Future work should further explore the range of stimulus parameters and responses (i.e. T and Δt), and compare the results for different stimuli and response segment. Also, our analysis of the stimulus features should include go beyond the mean of the stimulus waveform. It would also be of interest to apply to this problem the “full” bottleneck method rather than the agglomerative version.

One might object that H1 is an identified neuron in an invertebrate, and hence these results might not be typical of vertebrate or especially cortical codes. This is an empirical question. We believe that our approach to the compressibility of the coding dictionary sharpens considerably our questions about the structure of the neural code in any system, and we hope that the computational methods we have introduced will contribute to answering these questions.

References

- [1] Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: Exploring the Neural Code*, MIT Press, Cambridge (1997).
- [2] de Ruyter van Steveninck, R. & Bialek, W. Real-time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences, *Proc. R. Soc. London Ser. B* **234**, 379–414 (1988).
- [3] Mainen, Z.F. & Sejnowski, T.J. Reliability of spike timing in neocortical neurons. *Science* **268**, 1503–1506, (1995).
- [4] Bair, W. & Koch, C. Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation* **8**, 1185–1202, (1996).
- [5] Reich D.S., Victor J.D., Knight B.W., Ozaki T. & Kaplan E. Response variability and timing precision of neuronal spike trains in vivo., *J Neurophysiol.* **77**, 2836–41, (1997).
- [6] Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. & Bialek, W. Entropy and information in neural spike trains, *Phys. Rev. Lett.* **80**, 197–200 (1998).
- [7] Tishby, N., Pereira, F., & Bialek, W. The Information Bottleneck Method, in Proceedings of *The 37th annual Allerton conference on communication, control, and computing*, University of Illinois (1999).
- [8] Slonim, N. & Tishby, N. Agglomerative information bottleneck, *Advances in Neural Information Processing systems (NIPS)* **12**, 617–623 (2000).
- [9] Hausen, K. The lobular complex of the fly, in *Photoreception and vision in invertebrates* (ed Ali, M.), 523–559, Plenum (1984).
- [10] Cover, T.M. & Thomas J.A. *Elements of Information Theory*, Wiley, (1991).
- [11] Lin, J., Divergence measures based on the Shannon entropy, *IEEE Trans. Inf. Theory*, **37**, 145–151, (1991).
- [12] Schneidman, E., Brenner, N., Tishby N., de Ruyter van Steveninck, R. & Bialek, W. Universality and individuality in a neural code, *Advances in Neural Information Processing systems (NIPS)* **13**, in press (2001).