8.581 Entropy, information, and the brain

Plan for the first two weeks (5 through 16 January 1998).
Lectures by W. Bialek (bialek@research.nj.nec.com).
Notes as of December 16, 1997.

In these two weeks of lectures I hope to convey three main points:

1. Entropy provides the unique measure of information that is consistent with certain simple and intuitive requirements, and many of the things we learn in statistical mechanics have direct analogs and implications for the flow and processing of information. All this will be squeezed into Lecture 1 (!).

2. The brain processes information, and information theory gives us a tool to quantify the responses of neurons, providing a precise formulation of several intuitive questions about how the brain might work. Lectures 2 – 4.

3. Many of the computational problems solved by the brain are equivalent to extracting and efficiently representing information, separating interesting features from uninteresting clutter; information theory suggests a unified view of these diverse tasks. Lectures 5 – 8.

Point 1 is material that I think every physicist should know, simply because it represents a remarkable confluence of ideas. Point 2 is about the very concrete problem of characterizing what the brain does, and we will learn a few facts about biology (with a physicist's twist) as well as trying to bridge the gap from real data to a more abstract description of what is happening. Point 3 reaches toward a much more ambitious goal, using information theory as a framework for a true theory of the brain; you can decide how far we have come, but I hope we have fun along the way.
We will go from Shannon's work fifty years ago through work in progress by several groups today. There are many classic papers that you should read, though you won't get to all of them in these weeks; I plan to assemble some of these together with current references into a reader for the course. There are also several books that you may find useful and thought provoking.[1] A collection of all these references, organized under the lectures to which they are most relevant, is given below; papers to be included in the reader are marked by *. Finally, as we discuss work in progress and connections among different approaches, I do not guarantee that I have the answers to all the questions you will ask, so sometimes we shall struggle together.

[1] One reference which should be useful throughout the first week (and even some of the second) is *Spikes: Exploring the Neural Code*, which my colleagues and I finished last year, although again we will go beyond what is covered there. This is *not* a required text, but it does give details for many things to be discussed in the lectures.

8.581 Entropy, information, and the brain

Abstracts of the lectures (5 through 16 January 1998).

Lecture 1 (5 Jan): **Introduction & foundations.** Almost all statistical mechanics textbooks make some vague remarks about the entropy of a gas measuring our lack of information regarding the microscopic state of the molecules. Shannon proved a theorem that makes this precise: entropy is the unique measure of available information consistent with certain simple and intuitively plausible requirements. I will relate some of the facts about entropy that we have learned in the statistical mechanics context to corresponding ideas in information theory, e.g. the implications of extensivity and the equivalence of canonical and microcanonical ensembles in the limit of infinite volume. In statistical mechanics we also use correlation functions (and their Fourier transform, or momentum space, analogs) to study various types of ordering and the response to external perturbations; we will see how the same quantities arise in the description of time dependent signals and images, and discuss the connection of correlation functions to entropy and information. It is important that these foundations be solid, so this lecture will be accompanied by a relatively thorough set of exercises, including the analysis of experimental data.[2]

Lecture 2 (6 Jan): **Electrical & chemical signals in the brain.** The brain contains no metallic or semiconductor components. Signals can still be carried by electrical currents and voltages, but now currents consist of ions, such as potassium or sodium, flowing through water or through specialized conducting pores. These pores, or channels, are large molecules (proteins) embedded in the cell membrane, and can thus respond to the electric field or voltage across the membrane. The coupled dynamics of channels and voltages makes each neuron into a nonlinear circuit, and this seems to be the molecular basis for neural computation. Many cells have the property that these nonlinear dynamics select stereotyped pulses; these action potentials or spikes are the dominant form of long distance cell to cell communication in the brain, and our understanding of how these pulses occur is one the triumphs of the (now) classical 'biophysics.' Signals can also be carried by small molecules, which trigger various chemical reactions when they arrive at their targets. In particular, signal transmission across the synapse, or connection between two neurons, involves such small molecule messengers or transmitters. Calcium ions can play both roles, moving in response to voltage gradients and regulating a number of important biochemical reactions in living

[2] In general, each lecture will be supplemented by one long problem. We will discuss these problems on Fridays, and I will also arrange office hours. For students enrolled in the course, the full set of eight problems will be due on Tuesday, 20 January.

cells, thereby coupling electrical and chemical events, and the chemical events can reach into the cell nucleus to regulate which protein molecules—which ion channels and transmitter receptors—the cell produces. We will try to get a feeling for this enormous range of phenomena in one lecture, starting on the back of an envelope and building our way up to the facts.

Lecture 3 (7 Jan): **Reproducibility and variability in neural responses.** The signals that describe the sensory world are continuous: the reflectance of surfaces, the trajectories of motion, the time dependence of sound pressure or odorant concentrations, ... Similarly, the motor output of an organism consists of forces and motions that are not obviously discrete or quantized. Nonetheless, most of the neurons in our brain do not produce continuous or graded electrical signals, but rather generate sequences of discrete, identical pulses called action potentials or spikes. Evidently these spike trains constitute an encoding of the sensory and motor signals. Seventy years ago, Adrian observed that, in response to constant stimuli, sensory neurons produce spikes at a rate that increases with the strength or amplitude of the stimulus, and this picture of 'rate coding' has guided the exploration of previously uncharted areas of the brain. Adrian's observations left open several questions, usually grouped under the rubric of 'spike timing': With constant stimuli, can the temporal pattern of action potentials convey information beyond that carried by the spike rate? In a collection of cells, can the simultaneous occurrence of spikes in multiple cells 'mean' something special? With naturalistic, dynamic stimuli, can the timing of individual spikes mark particular events in the sensory input, and if so with what precision can this timing operate? Each of these questions can be given a precise formulation using ideas from information theory, and at least in some cases this approach has led to clear and decisive answers. What has emerged also is a clear definition of what it would mean to understand the neural code.

Lecture 4 (8 Jan): **Decoding the information in spike trains.** In the preceding lecture we discussed how much information neurons provide about the world, as well as evaluating which features of the neural response we need to record in order to preserve this information. Now we want to turn the problem around, and ask which features of the sensory world are represented in particular neurons; put another way, now that we know *how much* the spike train represents, we would like to know *what* it represents. The difficulty is that sensory inputs to the brain come from a very high dimensional space, and it is impossible to explore all possible mappings from this huge space to the (also rather large) space of spike trains. What I will do instead is to introduce several plausible simplifying assumptions—a notion of linearity or superposition of the contributions from different spikes in a train, the possibility that neurons are sensitive only to low dimensional features of the inputs, and so on. We will see how one can analyze experiments to provide, in some cases, direct evidence for these simplifications, and more generally how these simplifications are translated into algorithms for decoding the signals carried by individual neurons or by groups of neurons. In several instances these rather simple decoding procedures have been strikingly successful, squeezing out all the information that could be available, and allowing us to draw conclusions about the structure and precision of the computations that lead to feature extraction. In other cases the situation is less clear, and in some cases we know that information is present but haven't succeeded (yet!) in decoding it; this is very much the edge of our understanding, and I hope to give a feeling for some of the currently exciting issues.

Lectures 5/6 (12/13 Jan): **The search for efficient representations.** This a vast area, and even an introductory taste will spill over two lectures. Imagine describing an image by giving the light intensity in each pixel. Alternatively, we could give a description in terms of objects and their locations. The latter description is almost certainly more efficient in a sense that can be formalized using information theory. The idea of Barlow and Attneave (in the 1950s) was to turn this around—perhaps by searching for maximally efficient representations of natural scenes we would be forced to discover and recognize the objects out of which our perceptions are constructed. Efficient representation would have the added advantage of allowing the communication of information from one brain region to another (or from eye to brain along the optic nerve) with a minimal number of nerve fibers or even a minimal number of action potentials. Evidently the structure of a maximally efficient representation depends on the statistical structure of the signals we are trying to represent, leading immediately to several questions: What are the statistics of natural signals? Is there any sense in which the brain is more efficient at processing signals that come from this 'natural' statistical class? More generally, we can ask how to build efficient representations, and we can phrase this either as a mathematical problem (this is hard enough) or as a problem for the dynamics of development and learning in neural circuits. One well developed set of ideas concerns the search for optimal linear filters that enhance efficiency by removing second order correlations. Another idea is to use local nonlinear transformations to fit the distribution of input signals into the dynamic range of our neurons, and a related concept is to adapt the spike encoding properties of neurons to (hopefully) slowly varying low order statistics of the inputs. Evidence for all of these strategies can be found in the retinae of animals as diverse as flies and rabbits, and in some cases I think it fair to say that theory led to the discovery of interesting new dynamics in these neural circuits. A very different approach to efficient representation is the organization of input signals into discrete clusters which (perhaps) embody our intuition that a whole continuum of input signals will be labelled with the same identity in our perceptions. There is a statistical mechanics approach in which the emergence of clear clusters and subclusters can be seen as a phase transition. One of the most intriguing applications of these ideas is to the clustering of words, where (for example) nouns can be clustered based on the distribution of verbs with which they occur in sentences; as clusters emerge they seem to reflect groupings of nouns by meaning, with more subtle meanings distinguished in subclusters. These observations allow us to take a first quick look at the deep questions surrounding the idea of 'meaning' or 'value' for information. Finally, efficient representation can be phrased as a general information theoretic optimization problem, which frees us from specific references to filters, clusters, or even a metric on the space of inputs.

Lecture 7 (14 Jan): **Gathering information and learning rules.** The world around us is, thankfully, a rather structured place. Whether we have constructed a careful experiment or are gathering sense data on a walk through the woods, the signals that arrive at our brains are far from random noise; instead there appear to be some underlying regularities or rules. Surely one task our brain must face is the learning or extraction of these rules and regularities. Perhaps the simplest example of learning a rule is fitting a function to data—we believe in advance that the 'rule' belongs to a class of possible rules that can be parameterized, and as we collect data we learn the values of the parameters. This example, however, contains the kernel of many deep issues: if there is noise in the data then we really are trying to learn a probability distribution, not just a functional relation; we would like to compare models (classes of possible rules) that have different numbers of parameters, and incorporate the intuition that 'simpler' models are better; we might like to step outside the restrictions of finite parameterization and consider the possibility that the data are described by functions that are merely 'smooth' to some degree; at the opposite extreme we might want to learn Boolean functions that express logical relations among the observed data; ... ; finally we might like to quantify how much we are learning (or how much *can* be learned) about the underlying rules. In the last decade or so, a rich literature has emerged, tackling these problems with a sophistication far beyond the curve fitting exercises that we all performed in our student physics laboratories. I will try to take a path through these developments, emphasizing throughout the connections of these learning problems to problems in statistical mechanics and the implications of this statistical approach for an information theoretic characterization of how much we learn.

Lecture 8 (15 Jan): **Predictability, complexity, and meaning (maybe).** Our sense organs gather information at enormous rates. The task of the brain is, in part, to generate an "understanding" of these observations. One component of this understanding is data compression: the knowledge, for example, that there are objects called human heads, and that these objects have certain geometrical regularities, allows us to give a much more concise description of the images falling on our retina than if we had to enumerate the outputs of every photoreceptor cell, as discussed in Lectures 5/6. But we need to do more than provide an efficient representation: the precise location of every hair on a human head is seldom of interest, and true understanding seems to require that we discard these uninteresting components of the data. Is there a general principle that determines the relevance or irrelevance of different features in the continuous stream of sense data? In this last lecture I consider the hypothesis that data are interesting to the extent that they provide hints about what will happen in the future, and show how we can use information theory to quantify this predictive information. The magnitude of the predictive information in a data stream correlates well with our intuition about the complexity of the processes generating the data, and the extraction of predictive information provides a unified description of several different neural computations, from signal detection to learning, connecting to the optimization principle presented at the end of Lectures 5/6. Finally, this analysis identifies a class of truly complex data streams in which we cannot learn to make efficient predictions without some additional knowledge of what patterns are significant; the available statistical data suggest that language falls in this class.

8.581 Entropy, information, and the brain

References.

Lecture 1 (5 Jan): **Introduction & foundations.**

- C. E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.* **27,** 379–423, 623–656 (1948). Reprinted in Shannon and Weaver (1949).
- *C. E. Shannon, Communication in the presence of noise, *Proc. I. R. E.* **37,** 10–21 (1949).
- C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).
- L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1962).
- *R. R. de Ruyter van Steveninck and S. B. Laughlin, The rate of information transfer at graded–potential synapses, *Nature* **379,** 642–645 (1996).
- F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, 1997).

Lecture 2 (6 Jan): **Electrical & chemical signals in the brain.**

- *A. L. Hodgkin and A. F. Huxley, Currents carried by sodium and potassium ions through the membrane of the giant axon of *Loligo, J. Physiol.* **116,** 449–472 (1952a).
- *A. L. Hodgkin and A. F. Huxley, The components of membrane conductance in the giant axon of *Loligo, J. Physiol.* **116,** 473–496 (1952b).
- *A. L. Hodgkin and A. F. Huxley, The dual effect of membrane potential on sodium conductance in the giant axon of *Loligo, J. Physiol.* **116,** 497–506 (1952c).
- *A. L. Hodgkin and A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.* **117,** 500–544 (1952d).
- B. Katz, *Nerve, Muscle, and Synapse.* (McGraw–Hill, New York, 1966). Note: this is a brief and delightful book by one of the great masters.
- *E. Neher, Ion channels for communication between and within cells, *Neuron* **8,** 605–612 (1992).
- *B. Sakmann, Elementary steps in synaptic transmission revealed by currents through single ion channels, *Neuron* **8,** 613–629 (1992).

– B. Hille, *Ionic Channels of Excitable Membranes,* 2d ed. (Sinauer Associates, Sunderland MA, 1992).

– *L. Y. Jan and Y. N. Jan (1994). Potassium channels and their evolving gates, *Nature* **371,** 119–122.

Lecture 3 (7 Jan): **Reproducibility and variability in neural responses.**

– D. MacKay and W. S. McCulloch, The limiting information capacity of a neuronal link, *Bull. Math. Biophys.* **14,** 127–135 (1952).

– *H. B. Barlow, Summation and inhibition in the frog's retina, *J. Physiol. (Lond.)* **119,** 69–88 (1953).

– *S. W. Kuffler, Discharge patterns and functional organization of mammalian retina, *J. Neurophys.* **16,** 37–68 (1953).

– *D. H. Hubel and T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol. (Lond.)***160,** 106–154 (1962).

– *Z. F. Mainen and T. J. Sejnowski, Reliability of spike timing in neocortical neurons, *Science* **268,** 1503–1506 (1995).

– *R. de Ruyter van Steveninck, G. Lewen, S. Strong, R. Koberle, and W. Bialek, Reproducibility and variability in neural spike trains, *Science* **275,** 1805–1808 (1997).

– *M. J. Berry II, D. K. Warland, and M. Meister, The structure and precision of retinal spike trains, *Proc. Nat. Acad. Sci. (USA)* **94,** 5411–5416 (1997).[3]

– *S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, Entropy and information in neural spike trains, *Phys. Rev. Lett.* in press.

– *N. Brenner, S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, The economy of impulses and the rigidity of spike trains, preprint.

– *G. Buracas, A. Zador, M. DeWeese, and T. Albright, Efficient discrimination of temporal patterns by motion sensitive neurons in primate visual cortex, preprint.

Lecture 4 (8 Jan): **Decoding the information in spike trains.**

– *R. de Ruyter van Steveninck and W. Bialek, Real–time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences, *Proc. R. Soc. Lond. Ser. B* **234,** 379–414 (1988).

– *W. Bialek and A. Zee, Coding and computation with neural spike trains, *J. Stat. Phys.* **59,** 103–115 (1990).

[3]In the lectures, I will describe more recent results from this group, based on posters presented by Berry et al. at two conferences in 1997. I thank M. J. Berry II, J. R. Keat, T. Jordan, D. K. Warland and M. Meister for permission to show figures from their work.

– *W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, Reading a neural code, *Science* **252,** 1854–1857 (1991).

– *H. S. Seung and H. Sompolinsky, Simple models for the reading neuronal population codes, *Proc. Nat. Acad. Sci. (USA)* **90,** 10749–10753 (1993).

– *F. Rieke, D. Warland, and W. Bialek, Coding efficiency and information rates in sensory neurons, *Europhys. Lett.* **22,** 151–156 (1993).

– *E. Salinas and L. Abbott, Vector reconstruction from firing rates, *J. Comp. Neurosci.* **1,** 89–107 (1994).

– F. Rieke, D. Bodnar, and W. Bialek, Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory neurons, *Proc. R. Soc. Lond. Ser. B* **262,** 259–265 (1995).

– *D. K. Warland, P. Reinagel, and M. Meister, Decoding visual information from a population of retinal ganglion cells, *J. Neurophysiol.* in press.

– Note: much of what is discussed in the preceding references is also described in *Spikes: Exploring the Neural Code.*

– *W. Bialek and R. R. de Ruyter van Steveninck, What do motion sensitive visual neurons compute?, preprint.

– *E. N. Brown, L. M. Frank, D. Tang, M. C. Quirk, and M. A. Wilson, A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells, preprint.

Lectures 5/6 (12/13 Jan): **The search for efficient representations.**

– *H. B. Barlow, Possible principles underlying the transformation of sensory messages. In *Sensory Communication,* W. Rosenblith, ed., pp. 217–234 (MIT Press, Cambridge, 1961).

– H. B. Barlow, Single units and sensation: A neuron doctrine for perception, *Perception* **1,** 371–394 (1972).

– H. B. Barlow, The absolute efficiency of perceptual decisions, *Philos. Trans. R. Soc. Lond. Ser. B* **290,** 71–82 (1980).

– *H. B. Barlow, Critical limiting factors in the design of the eye and visual cortex, *Proc. R. Soc. Lond. Ser. B* **212,** 1–34 (1981).

– S. B. Laughlin, A simple coding procedure enhances a neuron's information capacity, *Z. Naturforsch.* **36c,** 910–912 (1981).

– *D. Field, Relations between the statistics of natural images and the response properties of cortical cells, *J. Opt. Soc. Am. A* **4,** 2379–2394 (1987).

– *K. Rose, E. Gurewitz, and G. Fox, Statistical mechanics and phase transitions in clustering, *Phys. Rev. Lett.* **65,** 945–948 (1990).

– J. J. Atick and A. N. Redlich, Towards a theory of early visual processing, *Neural Comp.* **2,** 308–320 (1990).

- J. J. Atick, Z. Li, and A. N. Redlich, Understanding retinal color coding from first principles, *Neural Comp.* **4,** 559–572 (1992).
- *J. J. Atick, Could information theory provide an ecological theory of sensory processing?. In *Princeton Lectures on Biophysics,* W. Bialek, ed., pp. 223–289 (World Scientific, Singapore, 1992).
- *F. Pereira, N. Tishby, and L. Lee, Distributional clustering of english words, *30th Annual Meeting of the Association for Computational Linguistics,* pp. 183–190 (1993).
- *D. L. Ruderman and W. Bialek, Statistics of natural images: Scaling in the woods, *Phys. Rev. Lett.* **73,** 814–817 (1994).
- *N. Barkai and H. Sompolinsky, Statistical mechanics of the maximum likelihood density estimation, *Phys. Rev. E* **50,** 1766–1769 (1994).
- M. DeWeese, *Optimization Principles for the Neural Code* (Dissertation, Princeton University, 1995).
- S. Becker, Mutual information maximization: Models of cortical self-organization, *Network* **7,** 7–31 (1996).
- *S. Smirnakis, M. Berry, D. Warland, W. Bialek and M. Meister, Adaptation of retinal processing to image contrast and spatial scale, *Nature* **386,** 69–73 (1997).

Lecture 7 (14 Jan): **Gathering information and learning rules.**

- E. Levin, N. Tishby, and S. A. Solla, A statistical approach to learning and generalization in layered neural networks, *Proc. I. E. E. E.* **78,** 1568–1574 (1990).
- *H. S. Seung, H. Sompolinsky, H., and N. Tishby, Statistical mechanics of learning from examples, *Phys. Rev. A* **45,** :6056–6091 (1992).
- *D. J. C. MacKay, Bayesian interpolation, *Neural Comp.* **4,** 415–447 (1992).
- *D. Haussler, M. Kearns, H. S. Seung, and N. Tishby, Rigorous learning curve bounds from statistical mechanics, *Machine Learning* **25,** 195–236 (1996).
- *W. Bialek, C. G. Callan, and S. P. Strong, Field theories for learning probability distributions, *Phys. Rev. Lett.* **77,** 4693–4697 (1996).
- *V. Balasubramanian, Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions, *Neural Comp.* **9,** 349–368 (1997).

Lecture 8 (15 Jan): **Predictability, complexity, and meaning (maybe).**

- *C. E. Shannon, Prediction and entropy of printed English, *Bell Sys. Tech. J.* **30,** 50–64 (1951).
- *P. Grassberger, Toward a quantitative theory of self-generated complexity , *Int. J. Theor. Phys.* **25,** 907–938 (1986).

– C. H. Bennett, How to define complexity in physics, and why, in *Complexity, Entropy and the Physics of Information,* W. H. Zurek, ed., pp. 137–148 (Addison-Wesley, Redwood City, 1990).

– M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York, 1993).

– A. S. Weigend and N. A. Gershenfeld, eds., *Time series prediction: Forecasting the future and understanding the past* (Addison–Wesley, Reading MA, 1994).

– *W. Bialek, Notes on predictive information, unpublished.