Towards a Cross-Level Theory of Neural Learning

Anthony J. Bell

Redwood Center for Theoretical Neuroscience
University of California at Berkeley
3210F Tolman Hall, MC# 3192
Berkeley, CA 94720-3192

Abstract. This paper reviews ideas and results from unsupervised learning theory that have given the best explanation yet of how neural firing rates self-organise to code natural images in area V1 of visual cortex. It then discusses the generalisation of these ideas to self-organising spike-coding networks. A mismatch between the resulting spike-learning algorithm and the known physiological processes of synaptic plasticity is then used as a motivation to introduce the rather obvious idea that neurons are not sending their information to other neurons, but to synapses – more microscopic structures. This prompts a survey of other inter-level communications in the brain and inside cells. It is proposed on the basis of this that information flows all the way up and down the reductionist hierarchy – an idea that transforms many of our ideas about machine learning and neuroscience. What it transforms them into is not yet clear, but the remainder of the paper discusses this.

Keywords: Machine learning, Neuroscience, Spiking neurons, Reductionist hierarchy
PACS: 87.19.lo, 87.19.L-, 87.19.lv **Citation:** appears in ref. [31]. Slides on MaxEnt 2007 website.

THE NEED FOR PROGRESS

Despite many technical advances in probabilistic machine learning no-one has been able to connect its ideas convincingly with the learning processes occurring in the brain. At the same time, efforts to understand biological self-organisation with ideas from physics have not yielded as much as might be hoped. And to complete the triangle, the project to connect physical law to principles of information and computation is still a marginal activity, despite some fascinating results (for example [13] in this volume).

It is nonetheless anticipated that these 3 lines of enquiry (physics, biology, inference) will converge before long and a new science of complex systems will be invented. The mother-to-be of this invention is necessity. We face enormous challenges in climate, ecology, health and education – in the organisation of our societies and in their relationships to the biological systems that they contain and that contain them. At the same time, our communications and biotechnologies are transitioning to a new level of sophistication. It is hard to believe that we will be able to use our new technologies responsibly and find solutions to our problems without a better understanding of what life is, what learning is – for what characterises life perhaps above everything else, is its ability to adapt to and create new circumstances. We need to understand what gives biological systems their amazing adaptive abilities. This paper makes a serious attempt to propose a new line of thinking about this, using results and controversies in neuroscience and statistical learning theory as a guide.

MACHINE LEARNING AND THE BRAIN

The modern theories of statistical machine learning and probabilistic inference, although they emerged largely from the neural networks community, have little to say to an experimental neuroscientist. A comfortable and unexamined consensus seems to exist along the lines of “optimal perceptual inference to build representations, and optimal decision theory to choose actions”. This view is so common that it may be called ‘the Standard Model’. Its first component (inference) imagines the cortex utilising Bayesian procedures to estimate the values of variables in a scene (like depth). The theory is supported by the finding that humans can combine relevant information in Bayesianly optimal ways (for example visual and touch information [19]). But the question of which of the combinatorial number of possible ‘hidden variables in the world’ one should estimate is not answered by the Bayesian framework. The theory only works if it is already known what should be estimated. Our brains cannot always be estimating all variables of potential relevance. Ideas about attention, involving guiding feedback from higher cortical areas, have not yet matured into an accepted theory, and require the brain to have a motivation, to which we now turn.

Somewhere in the middle of the brain, the problem of representing turns into the problem of choosing and executing actions. At this point, the second stage of the Standard Model imagines cortex computing actions that maximise an abstract utility function, sometimes called ‘reward’, based on motivational information supplied by sub-cortical structures. The problem with this theory is that the circuits carrying this information themselves need to learn how to convert sensory input into motivational signals. There is no *given* reward function. The (rather elegant) mathematics of reinforcement learning, in which the reward signal comes in on a ‘special wire’ from outside unfortunately does not map onto the real situation where the neuromodulatory connections flooding cortex from subcortical structures are themselves plastic (or else they could not be altered by addictive substances). Much of what is rewarding is in constant flux as the needs of the physical organism change from moment to moment. The fact that reward is necessarily a plastic function within the system, and not a value judgement mystically arriving from outside is never more apparent than when one visits a robotics lab where reinforcement learning is being used. Like the undefined fitness functions of an Artificial Life simulation, the undefined reward functions of reinforcement learning signal the inadequacy of the theory underlying them to apply to the neurophysiological situation.

Even if we were to somehow blend the decision theoretic stage into the perceptual inference stage, so the Standard Model looked less like a two-stage homuncular hangover from a Cartesian worldview, we would still be stuck with the reward concept, defined when the rat is in the box, but much more elusive *in vivo*. Perhaps it is time to dispense with this concept and its attendant goal of explaining the complex social behaviours of N creatures as being that of the optimisation of N separate undefined scalar reward functions.

Aside from concerns about attention and reward, the Standard Model, with its focus on sensory and motor processes, has nothing to say about what the brain is doing when it is just thinking. Why for that matter do we need to sleep? All creatures with nervous systems above a certain complexity need to sleep or their nervous systems become epileptic, causing death. This leads many to believe that sleep is a neural requirement. It is unlikely that sleep exists just to conserve energy since we use as much when we are asleep as when we are awake and resting.

If reward is at best a learned function, then the brain's learning must be unsupervised. Unsupervised learning attacks a host of problems, including clustering, but in the absence of an *a priori* need to cluster, it is perhaps best viewed as simply density estimation. The line that I will follow in this paper is that density estimation is the correct way to think about learning in the brain. To summarise the theory in one sentence: the brain is saying 'how likely was that?' and adjusting itself accordingly. The density estimation occurs not between some external world input and some internal brain representation, but rather *across levels of the reductionist hierarchy*. But we are getting ahead of our story. First we must explain what density estimation is and why anyone would think it adequate to explain something so unsensory as behaviour.

SENSORY-MOTOR DENSITY ESTIMATION

The goal of density estimation is to fit a model probability density function (pdf) to data, as when we find the mean and variance that best fits a normal curve to a histogram of data values of, for example, peoples' heights. The objective that is minimised is called the Kullback-Leibler divergence between the data pdf p and the model pdf q, defined as $D[p|q] = \langle \log(p/q) \rangle_p$ where $\langle \cdot \rangle_p$ means the average over the pdf p.

Now if we had some arbitrary parameter w that we wanted to learn (say the strength of a synapse in the brain), we could learn it by computing the gradient of the KL divergence and running down it till we reach a minimum. A few lines of calculus shows this gradient to be:

$$\partial_w D[p|q] = \left\langle \left(1 + \log \frac{p}{q}\right) \partial_w \log p - \partial_w \log q \right\rangle_p \tag{1}$$

The second term is the gradient of the log likelihood of the data under the model, which (before we take the Bayesian path and talk about prior distributions over models) is the gradient normally followed in density estimation algorithms. This term is the *sensory* term, corresponding to changing ones model to fit the data. The first term, on the other hand, is a *motor* term: it corresponds to changing ones data to fit the model. While this would be a disreputable activity for a statistician, it is nonetheless a part of life, if we consider that a synaptic weight change may change the probability of future data.

Unless the dynamics of the world's data-generation process is deterministic and known, it seems impossible to evaluate this first term. But its very existence wakes us up to the fact that density estimation need not be a sensory game alone: the gradient of the KL divergence and the log likelihood are different. Furthermore, the tables are turned on a common complaint against unsupervised or Shannon information based learning models: that they do not distinguish between meaningful information and noise. The appearance of a second term dependent on the motor-influence of w suffices to make some data more relevant than other data. This term has the potential to actually provide a foundation for the signal/noise distinction, containing, as it does, the subjectivity of how ones brain-state effects the world.

But how can this term do so in a way that is meaningful to a creature - things like finding food and keeping warm? We do not have an answer to this question. But if we can find a cross-level theory of learning such that even cells and molecules can be seen as modeling and changing their local environments and contributing to a global model, then when they are too cold or lack energy, they will display dynamics that, just like an agent, will give rise to emergent properties that cause macroscopic behaviour changes, such as eating or finding shelter.

Such a theory may be out of reach at the moment, but we have to start somewhere. We will start by exploring the technicalities of statistical density estimation in order to set the stage for the cross-level ideas introduced later on. This will necessitate a dip into the mathematics, but for the unitiated, hold on - the story gets better later.

If the sensory-motor problem does have the unsupervised structure argued for by eq.(1), then the introduction of the motor term, for all its analytic intractability, *cannot* make things worse. It is *easier* to find the hidden degrees of freedom of the world when we can manipulate objects than when we are just looking at pictures. Children learn by doing, not just observing. It is easier to understand a world that we are in the process of creating than one that is given to us. (Of course, we can always go too far in this, leading to solipsism. A question for later sections is: what force tends to keep an adaptive agent away from a solipsistic solution?)

Knuth et al ([32] and references therein) describe a different unsupervised approach to the sensory-motor problem. It uses the combined calculus of inference and inquiry to design instruments that perform actions to maximise expected information gain. This is eminently sensible, but biological systems do not appear to function this way. Rather, organisms seem to converge on *autopoetically* stable reverberations with partially self-constructed environments [36], and this is closer to the idea we are trying to reach. A robot maximising information gain would never stop to sing a song.

Another interesting unsupervised approach maximises the information flow in the sensory-motor loop [30], though it is subject to the this same criticism.

SENSORY DENSITY ESTIMATION

Density estimation learning methods have already provided our best computational model to date for how the brain might self-organise from experience. One can show tiny images or movies to a neural network and train it to have receptive fields similar to those measured by single-neuron recordings from area V1 of visual cortex of cats and monkeys. Fig. 1a viewable on the internet, shows a movie of an experiment done by Hubel and Wiesel on a cat in the 1960s, demonstrating how the notion of a receptive field arose.

Theoretical results of this kind were first obtained by Olshausen and Field [38] using the idea that neurons should try be sparse (fire rarely) and decorrelated. But the receptive fields can be obtained using density estimation alone. The results in Fig. 1b and 1c were obtained by an Independent Component Analysis (ICA) network [5] and a kind of *Dependent* Component Analysis (DCA) network related to Hyvärinen and Hoyer's 'Topographic ICA' [27, 39, 49]). These receptive fields are static. Dynamic (spatio-temporal) receptive fields were obtained by van Hateren and Ruderman [48], and are

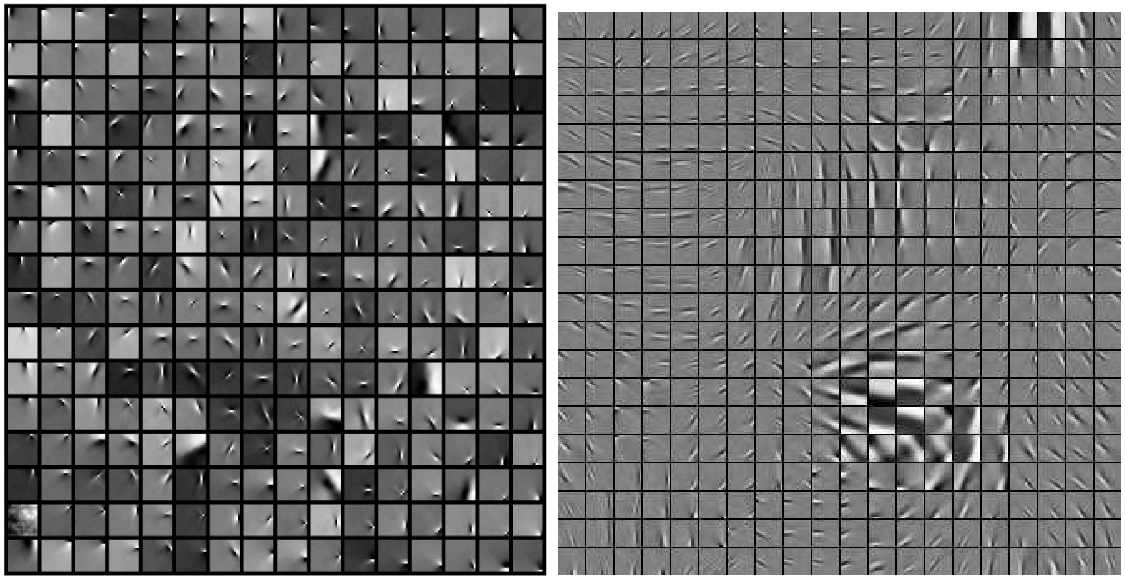


FIGURE 1. Receptive fields learned from natural image data. The full version of this figure is on the internet at www.snl.salk.edu/t̃ony/RecFields.html. (a) On web: A movie showing Hubel and Wiesel's discovery of visual receptive fields in cat. (b) Left above: ICA-learned image bases. Each picture is a learned axis in image space, corresponding to a column of $\mathbf{W}^{-1}$ (see text). (c) Right above: a typical basis set obtained with a model closely related to Topographic-ICA [49] [Thank you to Simon Osindero for permission to reprint his figure]. (d) On web: spatio-temporal receptive fields learned from natural movies by van Hateren and Ruderman [48]

shown on the web (Fig. 1d). In both cases shown here, receptive fields are 'Gabor-like': localised in space, orientation, spatial frequency and phase, like Hubel and Wiesel's 'simple cells' of V1. (Phase-invariant 'complex cells' may also be learned [27]). In the topographic ICA case, the neurons are also spatially ordered in a 2D grid, or 'map', very much as V1 cells are arranged across the sheet of cortex, ie: in a orientation 'column' where position, orientation and spatially frequency vary continuously across the map, except at discontinuities called pinwheels visible in Fig. 1c.

Both ICA and DCA are simple density estimation networks. They take a multivariate data distribution and find a new set of axes in it (just as a Fourier transform or Principal Component Analysis does). Unlike PCA, the new coordinate system is chosen entirely on the basis of the statistics of the data (PCA and Fourier bases impose the additional constraint that the axes be orthogonal in the original space). [It is important here not to confuse the orthogonality of the transform with the decorrelation of the resulting output variables.] The axes are found by training a complete set of filters (ie: a square matrix $\mathbf{W}$) to transform the data by $\mathbf{u} = \mathbf{W}\mathbf{x}$ into a new vector space $\mathbf{u}$ where the elements u_i are either as statistically independent as possible, or statistically dependent in some specified way. The training is done by presenting the images one at a time and changing the filter matrix according to one of the following equations:

$$\text{ICA}: \quad \Delta\mathbf{W} \quad \propto \quad \left(\mathbf{I} - \left\langle \mathbf{f}(\mathbf{u})\mathbf{u}^T \right\rangle_p\right)\mathbf{W} \tag{2}$$

$$\text{DCA}: \quad \Delta\mathbf{W} \quad \propto \quad \left(\left\langle \mathbf{f}(\mathbf{u})\mathbf{u}^T \right\rangle_q - \left\langle \mathbf{f}(\mathbf{u})\mathbf{u}^T \right\rangle_p \right) \mathbf{W} \tag{3}$$

where $\mathbf{I}$ is the identity matrix. The learned axes (or *basis functions*) actually correspond to the columns of the inverse of the filter matrix $\mathbf{W}^{-1}$.

Both algorithms linearly transform the data into a $\mathbf{u}$-space where a certain statistical model, $q(\mathbf{u})$, (a 'shaping density') is imposed. The optimisation is to fit the transformed data to this model by gradient ascent in the log likelihood of the data under this model via $\Delta\mathbf{W} \propto \langle \partial_{\mathbf{W}} \log q(\mathbf{x}) \rangle_p$, just as in eq.(1) without the motor term. The models on the input and output neurons are related by $q(\mathbf{x}) = q(\mathbf{u})\,|\mathbf{W}|$, where $|\cdot|$ means the absolute determinant.

In ICA, the model factorises: $q(\mathbf{u}) = \prod_i q(u_i)$, and the details of the univariate marginals, $q(u_i)$, may also be learned (though it is often un-necessary to do so). The vector of functions $f(\mathbf{u})$ has entries $f_i(\mathbf{u}) = -\partial_{u_i} \log q(\mathbf{u})$, and these are called the score functions. If these were linear, a condition satisfied by having a gaussian models on the $q(u_i)$, then ICA can be seen to stabilise on average ($\langle \Delta\mathbf{W} \rangle_p = 0$) when $\mathbf{I} = \left\langle \mathbf{u}\mathbf{u}^T \right\rangle_p$, in other words when the outputs are unit variance decorrelated. To make non-gaussian signals independent we need statistics higher than second-order and these are provided by the Taylor-expansion of the score functions.

Much more could (and has) been said about this, but the main point to make here is that DCA is the completely general form, turning into ICA when the model we impose is that of independence, ie: $\left\langle \mathbf{f}(\mathbf{u})\mathbf{u}^T \right\rangle_q = \mathbf{I}$. The DCA form can be derived by writing the model density in the completely general Gibbs' form:

$$q(\mathbf{u}) = \frac{1}{Z} e^{-E(\mathbf{u})} \tag{4}$$

involving an 'energy' $E(\mathbf{u})$ and a normaliser called the partition function Z [25]. The two averages over the model and data densities in eq.(3) are then seen to arise from the gradients with respect to $\mathbf{W}$ of the log partition function and the energy respectively. The learning equation for a single weight has exactly a Boltzmann machine structure [24] consisting of a Hebbian (correlational) term sampled over the data density and an anti-Hebbian (anti-correlational) term sampled over the model density. This is said by some to be accomplished by alternately learning from data in an awake phase, then unlearning from the model in an asleep phase. This idea, while intriguing, has yet to condense into a serious neurobiological theory of sleep, but it is one to which we will return. There are few applications of DCA, for the same reason that there are few for the Boltzmann Machine, namely that the training (sampling from q) is just too slow.

The topographic ICA results in Fig. 1c [27, 39] are actually obtained by a very simple DCA model[1], but more complicated models run into this need to integrate over the model density (the first term of eq.(3)). This integration, which is a universal bother in machine learning and statistical physics is only tractable in simple cases (like Gaussian or ICA models), and otherwise, as mentioned above, we must resort to sampling from

[1] overlapping neighbourhoods on a map: $q(\mathbf{u}) \propto \prod_K q(\mathbf{u}_K)$ where $\mathbf{u}_K$ means neighbourhood K of the map, and radially symmetric laplacian multivariate marginals, $q(\mathbf{u}_K) \propto \exp(-\|\mathbf{u}_K\|)$

the model density using one of many schemes (Monte-Carlo Markov Chain (MCMC) or Contrastive Divergence [26] being two such schemes).

Were we able to solve the model-selection problem (the choice of $q(\mathbf{u})$) and the gradient of the partition function, we would be in good shape to attempt the Holy Grail problem of building hierarchical representations just from data, as we could use the resulting groupings of variables (like the neighbourhoods of the topographic map) to non-linearly recoordinatise the data at each layer and then look for new structure 'unwrapped' by the non-linear recoordinatisation. An example would be to re-express data fitting a radially symmetric laplacian model in spherical coordinates (phases and an amplitude) and input this to a higher network.

Many have travelled this road (my attempt is in [7]) and few have emerged unbloodied and with meaningful results (the few are [29, 27, 39]). It is not a problem I would recommend to a graduate student unless he had a good new idea. Rather I would recommend stepping back to look at the problem afresh, and biology can be a great inspiration in redesigning ones question until the answer looks right. In other words, if one is struggling with the problems of model selection and partition function gradients then perhaps it is a good idea to ask how on earth these problems map onto the tissue inside our skulls. That is the track that we will follow in the next part of the story.

To conclude, this section has been a quite dense summary of much technical work by many people just to arrive at two equations. eq.(3) is Amari et al's Natural Gradient [2] transformation of Hinton et al's view [25] of the Infomax-ICA algorithm [4], which is identical to the maximum likelihood approach (see [12] for an explanation). The Natural Gradient concept (optimisation in the metric space of matrices) is explained in detail (without reference to Amari's Information Geometry) in [47, 37]. The ICA method in eq.(2) is also in the natural gradient form proposed by Amari et al where the weight space is given a Fisher metric based on reasons of information geometry [3]. A review of related sparse-coding techniques is found in [45].

SPIKING DENSITY ESTIMATION

Looking at biology, there is quite a variety of phenomena to draw inspiration from. Since it was not at all clear what model selection and the partition function gradient might mean in neural tissue, I decided to focus on a problem that had disturbed me for a long time: the issue of learning with neurons that spike. My earlier attempts on this had floundered (actually giving the Infomax-ICA algorithm as a by-product). The reason to tell the story is that it moves us close enough to biology that we can derive the *reductio ad absurdem* which sends us in a completely new direction. The story is interesting and I hope the reader will indulge me.

The problem was as follows: most real neurons communicate with each other not by sending real numbers (as in neural network models) but by sending pulses called spikes which last about 1ms. You can hear them crackle away in the Hubel & Wiesel movie in Fig. 1a (on the internet). Unresolved controversy has raged in the neuroscience community for decades about whether or not the timings of these spikes is meaningful since they sound so much like Geiger counters popping randomly. Cortical pyramidal cells in area V1 which are *not* driven by their preferred visual stimulus (so-called

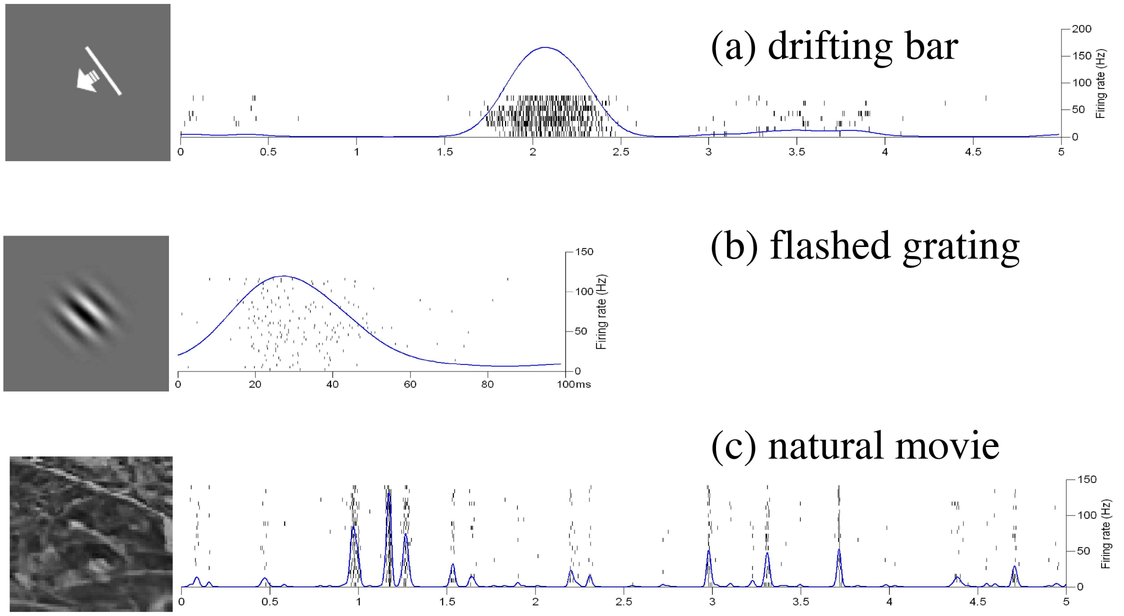


FIGURE 2. Recordings from an excitatory pyramidal cell in area V1 of an anaesthetised cat's visual cortex when shown (a) a drifting bar (see also the movie in Fig. 1a), (b) a flashed grating, and (c) a natural movie. In each case, each row of dots represents a single trial, and each dot is a neuron spiking. Data from Blanche et al [10] with permission.

spontaneously firing cells) have roughly Poisson firing statistics: that is - they look like completely random point processes. And when we repeatedly give a neuron its preferred stimulus, its rate repeatably goes up while the detailed structure of its spike timings are different on each trial, as can be seen in Fig. 2a and 2b.

There are two interpretations of this seemingly noisy Poisson-like firing. The dominant one has been that neurons are 'noisy rate coders' of their preferred stimuli. But a radically different explanation emerges if we consider that Poisson firing could also be the consequence of a neuron trying to maximise its information transmission rate. When we compress signals to maximise their information rate (as in image or video compression), the elements of the code become statistically independent (minimally redundant). If such an optimisation were to occur in spiking neurons, we would expect to see neurons firing with Poisson-like statistics.

The noisy rate coding idea is diametrically opposed to the idea that spike timings look noisy because they are highly informative. If one of these ideas is correct, the other one is wrong. Evidence for spike timing codes has built up over the years in studies of sensory neurons, but it is harder to demonstrate in cortical neurons because they are further from the sensory input and receive many unknown inputs from higher in the brain. The crucial breakthrough in studying this came when researchers started to record from cortical neurons while the animal was exposed to naturalistic stimuli instead of drifting gratings and bars designed to elicit maximum rate responses. An example of this is shown in Fig. 2c. In multiple presentations of a natural movie to an anaesthetised cat while recording from an area V1 cortical pyramidal cell, the spiking pattern was quite repeatable from trial to trial, and when a neuron fired, it fired 1-3 spikes reliably usually within a 50ms time window. Such responses to natural stimuli (which the neuron is presumably more used to) are not consistent with the noisy rate coding hypothesis, but they are consistent with a picture where individual spikes signal the precise timing of the perceptual events they encode (see also [17] for an example from rat auditory cortex).

Although this debate is by no means settled, it does stimulate the theorist to attempt a proof-in-concept that spike-timing codes can self-organise. I embarked upon this project, together with Lucas Parra and Jeff Beck [8, 40]. Our idea was to use the same density estimation learning described earlier, but where the elements of the neural code are spike-timings, not real numbers representing rates, as in the simpler neural networks trained by ICA or DCA.

The principles are the same, but the network is an integrate-and-fire network [22]. For the *i*th neuron, the time-dependent voltage is:

$$u_i(t) = \sum_j \mathbf{W}_{ij} \sum_k R_{ij}(t - t_k) \tag{5}$$

It sums over synaptic inputs j and spikes k arriving at that synapse at times t_k. The functions R_{ij} are the shapes of the potentials caused by the spikes, except R_{ii} which is the shape of the voltage reset after $u_i(t)$ reaches a threshold value and neuron i itself fires a spike. Our learning algorithm works by maximising the sensitivity of all output spike timings to input spike timings in a single-layer feedforward network. Without going into too much detail, there is a density model $q(\mathbf{t}_{in})$ ($\mathbf{t}_{in}$ being the vector of all input timings) and it is a function of the weight matrix $\mathbf{W}$ and the output timings $\mathbf{t}_{out}$. For every input spike l that helps cause an output spike k, the relevant synaptic weight $\mathbf{W}_{ij}$ changes according to:

$$\Delta \mathbf{W}_{ij} \propto \frac{\mathbf{T}_{kl}}{\mathbf{W}_{ij}} \left(\left[\mathbf{T}^{T\#}\right]_{kl} - \left[\mathbf{T}\mathbf{T}^{T\#}\right]_{kk} \right) - f(r_i) r_j \tag{6}$$

in which the matrix $\mathbf{T}$ is the spike-timing Jacobian (or sensitivity) matrix, having entries $\partial t_k / \partial t_l$ and the last term is a non-linear Hebbian term in the input (r_j) and output (r_i) spike rates, appropriately defined. As in eq.(2), f is again a score function.

We were very disappointed with this rule. It was a lot of work to find it, a lot of work to simulate it, and it is utterly biologically implausible. The simplicities of the ICA/DCA algorithms were not replicated in the spiking situation. The notation $\left[\mathbf{T}^{T\#}\right]_{kl}$, for example, represents the kl-th entry of the pseudoinverse of the transpose of the matrix representing the sensitivity of all output spike timings to all input spike timings, defined over all time and all neurons. The learning algorithm is horrendously non-local in space and time (meaning synaptic weight changes cannot be made using time-local pre- and post-synaptic information). Furthermore, the algorithm only works if there are more output spikes than input spikes (an *overcomplete* mapping being required to make a non-lossy map more probable). Were it not for Lucas Parra's persistence, this network learning rule would never have been derived, proven correct in simulations or published.

The answer was so complicated that the question had to be wrong. Referring again to the neurophysiology (always a good idea) soon revealed why. In focusing all our attention on the mapping between input spikes and output spikes, we had treated the dendrites of a neuron as if they were simple feedforward functions, designed to get information to the next neuron. In reality, there is also feedback from the output of the cell back to the synapses (called the back-propagating action potential), as well as electrical communication between synapses in the dendrites. In addition, the growing experimental literature on spike-based synaptic learning (called Spike Timing-Dependent Plasticity, or STDP), clearly showed that this information fed back to the synapse was implicated in synaptic plasticity. The physiology of synaptic plasticity is inordinately complex and controversial [15, 44], but one common theme emerges from the literature: calcium converts electrical signals into the molecular changes required to alter synapses. There are at least two distinct calcium currents operating in and around synapses to do this. The first enters through ion channels opened by neurotransmitter (the various kinds of NMDA receptor). The second enters through ion channels which are opened by changes in voltage internal to the cell (the NMDA receptor also has a voltage dependency). The other kind of receptor at excitatory synapses, the AMPA receptor, does not let in calcium and thus cannot drive plasticity directly. Details aside, what this means is that the synapse integrates activity external and internal to the cell to determine how it should change.

We had used the mapping from input spikes to output spikes as our trainable density model, ignoring the backward and sideways information pathways in the dendrites. Our learning rule was clearly unbiological, also in the way it required the feedforward neural mapping to be non-lossy.

The conclusion was obvious: when we added the other information pathways in, the mapping relevant for the purpose of learning was not from input spike to output spike, *but from input spike to synaptic readout*. This readout was done by calcium currents local to the synapse, not a thresholding mechanism at the axon hillock. And what was read out were three kinds of spiking activity: spikes arriving at that synapse, spikes arriving at other synapses and spikes propagating back from the cell body, the latter two signalling to the local synapse through graded potentials in the dendrites.

This may not sound startling to a physiologist, but from the direction we were approaching, the implications were startling indeed. Firstly, since there are roughly 1000 times as many synapses in the brain as neurons, the neuron-to-synapse mapping was 1000 times overcomplete, easily solving the problem that an invertable mapping was required for the density estimation maths to work. In fact the state variables at the neural level (ie: spikes) could now be as lossy as they liked, because they no longer had to model the statistics of other spikes - this job could be done by new state variables (driven by calcium) operating at a different level of the system: the synaptic level. Suddenly neural information was preserved in the map to *synaptic* readout, while (in all likelihood) thrown away in the map to *neural* readout. This bypassed the main criticism of information theoretic neural learning algorithms: that they could not throw away information, as neurons clearly did. It also made sense to have the informational readout at the site of learning, rather than the output of the cell, thus decoupling the circuit's statistical model from its feedforward computation, two essentially different tasks which were conflated in the ICA/DCA case which had no synaptic state variables, operating only at a single level.

To make the model concrete, it is proposed that synaptic plasticity (at excitatory glutamatergic synapses anyway) operates roughly within the following framework. The neuron is a network of protein complexes (post-synaptic densities and the axon hillock), communicating similarly to eq.(5):

$$u_a(t) = \sum_b w_b \sum_k R_{ab}(t - t_k, u_b) \tag{7}$$

except that now the indices a and b refer to these sites on the membrane. Each site has a learnable synaptic weight w_b and the transfer functions R_{ab} represent the effects that spikes k at site b can have on the voltage at site a (R_{aa} is the local synaptic response). This is essentially just the cable equation for linear electrical communication in dendrites, with a non-linear voltage-dependence added to account for the NMDA receptor voltage-dependency and conductance effects. At each site a, there are two calcium readouts, the first being synaptic (NMDA receptor) calcium c_a^+ and the second being intrinsic (voltage-dependent) calcium c_a^- carrying information from the rest of the cell:

$$c_a^+(t) = \lambda_a^+ w_a \sum_k R_{aa}(t - t_k, u_a) \tag{8}$$

$$c_a^-(t) = \lambda_a^- \sum_{b \neq a} w_b \sum_k R_{ab}(t - t_k, u_b) \tag{9}$$

The new ‘plasticity parameters’ λ_a^+ and λ_a^- represent the fraction of the local synaptic and intrinsic ionic currents which are calcium-carrying and thus available to drive molecular change. (Hippocampal excitatory synapses, for example, are much more plastic than cortical synapses, having much higher NMDA receptor counts. So-called ‘silent synapses’, common in developing nervous systems, and largely lacking AMPA receptors, would have λ_a^+ close to 1.)

These two kinds of calcium drive a first-order kinetic scheme involving a phenomenological variable y_a which is the ‘readout’:

$$\dot{y}_a = e^{c_a^+}(1 - y_a) - e^{-c_a^-} y_a \tag{10}$$

You may ask: where did these equations come from? The answer is that they are pure guesses based on intuition, a reading of the literature on the physiology of synaptic plasticity and a desire to simplify things. They are included here merely to illustrate what may be the essential features of a calcium-based synaptic readout: a dynamic computation that compares external input with internal activity to determine how a weight should change. The real situation is much more complex [15], and varies greatly with synapse-type. However the kinetic scheme is not a complete fantasy: the push-pull of c_a^+ and c_a^- is meant to represent the actions of calcium-driven kinase and phosphotase proteins (like CAM-K2 and calcineurin) which activate opponent processes controlling the delivery to and recycling from the membrane of AMPA receptors, or the alteration of their sensitivity through phosphorylation.

It is useful to try to make a concrete model that shows information flowing from the neural to the synaptic level. But we have no learning rule here, just some equations for synaptic readout that suggest that macroscopic (neural) activity might be statistically modeled by a more microscopic set of dynamic variables located at synapses. What is missing is an understanding of this kind of inter-level communication in general. It is to this that we now turn.

LEVELS IN BIOLOGY

As scientists, we usually like to believe that the level at which we work is the important level for understanding more macroscopic phenomena, more microscopic phenomena being irrelevant. This is understandable because science is the search for lawful behaviour – this search involves adjusting experimental conditions (macro-variables) until the things which are measured (meso-variables) behave deterministically. Micro-variables are then not needed to explain these cases, and are then often regarded as merely implementation detail for the observed laws, or if they interfere with the lawful behaviour, they are "noise". In other words, the scientific method, for all its successes, creates a series of self-reinforcing parochialisms, each centred on a certain level of description, each behaving deterministically largely only under the experimental conditions imposed.

No-one is specially to blame here. A molecular biologist who regards the quantum level as irrelevant cannot criticise a social psychologist for whom the skull is a reflecting barrier. These points may seem obvious, but think how often we read phrases like "the genetic basis of behaviour" or "the social basis of religion". The word 'basis' betrays a fundamentalism that seeks to diminish the importance of ordered emergence from the microsphere. And the notion that higher laws (ie: more compact determinisms) in the macrosphere are not much better ways of talking is also implicit here.

We have already seen two examples of problematic thinking in neuroscience that can arise from this: the rate-coding neuron which disappears when we show the cat a natural movie, and reward-maximising behaviours that become elusive when the rat escapes from the box. Rewards and rates are not defined under more natural conditions.

The other error mentioned above is to assume that deviations from deterministic behaviour are a consequence of 'noise'. The inability of an experimenter to predict a phenomenon does not mean that there are not meaningful hidden variables producing it. (You may not expect to read this sentence but that doesn't mean I am noisy!) Thus it is a mistake to talk about synaptic transmission as unreliable just because an experimenter cannot predict if a spike arriving at a presynaptic bouton will cause a vesicle release or not. Presynaptic filtering of spikes based on a bouton's internal state is probably an intelligent process.

Similar mistakes are made at the cellular level. Those studying bulk ion channel kinetics regard the motions of individual channels as noisy. Yet for molecules in the neighbourhood of single channels, these motions are a signal, particularly if any calcium flows through the channel. Calcium concentrations vary meaningfully over distances of 10nm – the width of a membrane protein. Calcium ions address individual molecules. Calcium may be to the post-synaptic density (PSD) what voltage is to the neuron as a whole: a spatially varying field communicating between proteins the way voltage communicates between the protein complexes called synapses. ([44] is an up-to-date review of the amazing structure of the PSD.)

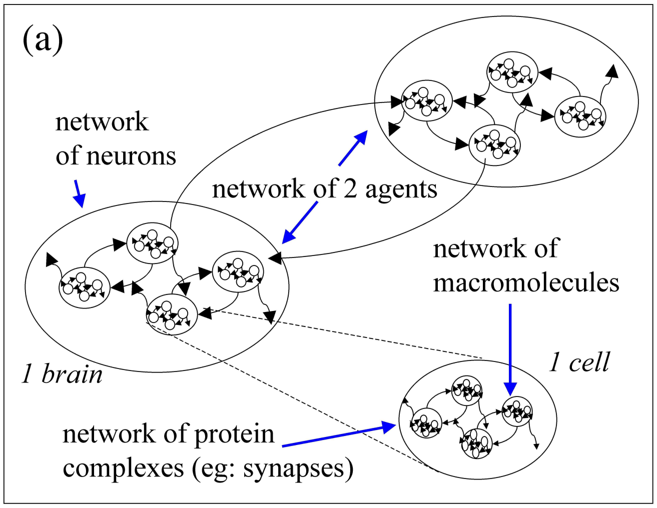
We turn now to the cytoplasm. For the classical biochemist, enzyme reactions occur as they do *en masse* in a thermal aqueous medium, molecules bumping into each other randomly. But a more modern picture of the cytoplasm reveals that there is 'macromolecular crowding' (in which up to 50% of the cell's volume may be taken up by mostly immobilised proteins and polynucleotides). This has opened the fascinating possibility that the remaining space (water) is *completely ordered* into a 3-dimensional protein-gated switching network, lined by charged hydration shells. These water pathways continuously semi-conduct ions and small molecules differentially based on their size, valence and shape [46, 18]. The cellular metabolism, argue some, is “vectorial”. The same seems true of the membrane, a packed and organised 2-dimensional fluid in which cholesterol-based structures of all sizes (called lipid rafts) control the positions and motions of every integral protein [23, 28]. The quiet revolution in cell biology that has produced this new picture of ordered and meaningful multiscale micro-organisation mirrors the developments in systems neuroscience away from preoccupation with rate-coding feature detectors and towards an understanding of cross-level interactions involving spikes, and correlated activity on all scales. We turn now to the latter.

At the level above that of the neuron, many new lawful relations are coming to light. (This area of research was largely pioneered by Walter Freeman). Global brain oscillations in the delta range (1-4 Hz) seem to constrain less global oscillations in the theta range (4-8 Hz) [34], just as theta oscillations seem to constrain oscillations in the even more spatio-temporally local gamma range (40-200 Hz) [14], and just as gamma oscillations seem to determine when spikes are most likely to occur [20, 21]. In each case, it is at a particular phase of the lower frequency, more global, oscillation that the higher frequency, more local, effect is likely to increase its amplitude (or probability of occurence, in the case of gamma-to-spike coupling). The timing of spikes relative to the gamma oscillation is emerging as an important information-carrying factor in visual coding [33] and elsewhere [21]. See also [11] for information on this and other related topics, including important findings about how theta-to-gamma coupling in the hippocampus codes information about whereabouts a rat is in its environment (place-coding). A mini-review on these topics appears in [43]. It is noteworthy that the structured relations between spike timings and oscillation phases can be disrupted for several minutes by zapping the local tissue with a magnetic field [1].

All of this indicates a very structured organisation of the brain’s oscillatory activity across space and time. Some neuroscientists have argued that these oscillations are ‘epiphenomena’, since they are just the result of many synchronised membrane currents. The computation, goes the argument, is occuring at the synapses – the junction points of the spiking computer. Others disagree, arguing that these oscillating electrical fields *can* influence spike timings directly [41], so-called ephaptic interactions. But even if this is not the case, describing these fields as epiphenomena is like saying that the US government is an epiphenomenon as it is just the result of the activity of many people talking to each other. The law-like influences of higher-order structures on lower is not an indicator of mystical downward causality but an epistemic statement about which groupings of variables are sufficient to summarise causal dependency. When we converse, we hear each others words, not each others spikes filtered through the air.

PUTTING IT TOGETHER?

We have mused on the problem of self-organising sensory-motor systems, explained some statistical learning theory, dived into the physiology of synaptic plasticity and outlined the flows of information across scales in the nervous system and inside cells.



(b) LEVEL	UNIT	INTERACTIONS	LEARNING
ecology	society	predation, symbiosis	natural selection
society	organism	behaviour	sensory-motor learning
organism	cell	spikes	synaptic plasticity
cell	synapse	voltage	bulk molecular changes
synapse	macromolecule	direct, Ca	molecular changes
macro-molecule	amino acid	molecular forces	gene expression, protein recycling

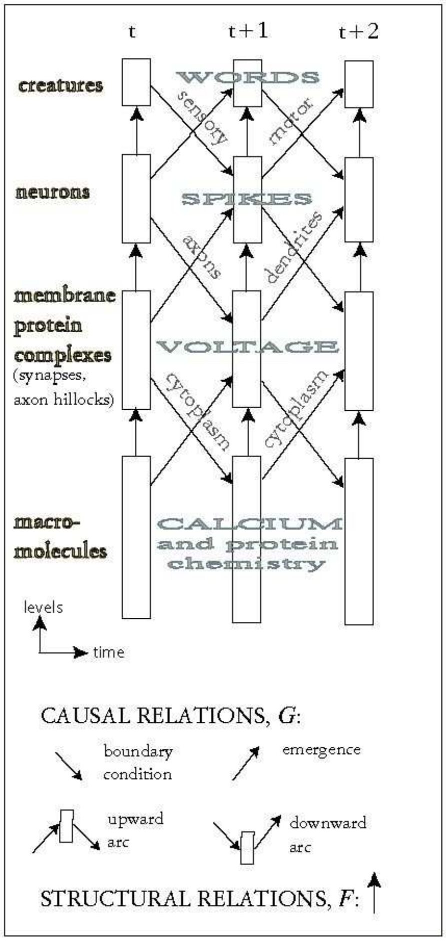


FIGURE 3. Three schematic views of the multi-level organisation found in biology. (a) Nervous systems viewed as networks of networks. (b) The units at each level are the networks at the level beneath. Adaptivity is given a different name at each level, but it is really a single unitary process, viewed at different scales, like an image viewed at different resolutions in space and time. Natural selection and self-organisation are thus not different processes, or competing explanations for biological adaptivity, but different spatio-temporal accounts of the same underlying thing. (c) A discrete-time 'cartoon' (for didactic purposes only) of the temporal evolution of a multiresolution state vector. Structural relations reduce the state vector as we ascend the reductionist hierarchy. Causal relations are thus similarly transformed. Objects at a given level do not communicate directly with each other, but rather through downward and upward arcs. For example, spikes can influence other spikes through dendrites and axons or through agents communicating.

But we have been dancing around our main theme: that of uniting learning theory with this multi-level picture. This is because when there is no good theory (note the first word of this paper's title!), it is important to to survey all clues, empirical and theoretical, as well as to identify the shortcomings in existing ideas.

But there is no shying away any more. The central observation so far, that neurons talk to synapses not neurons, may be trivial. But if we repeat it at different scales, it naturally leads to to a view of the nervous system which is at odds with most of the main strands of current computational and experimental thinking, yet this view is consistent with all the empirical data and throws the concepts of machine learning into new unexplored inter-level scenarios. This view is presented in Fig. 3. It has the following characteristics:

- Biological organisation consists of networks within networks (Fig. 3a). Microscopic networks are inside the nodes of macroscopic networks, like the network of synapses (eq.(7)) inside a neuron. The outputs and inputs of a network are signals to and from the network above in the hierarchy. Phases of groups of neurons, timings of individual neurons, flows of voltage and flows of calcium carry the network information at the cell assembly, neural, synaptic and molecular levels respectively.
- There is thus no ‘functionalist cut-off level’ [2] anywhere in the biological hierarchy [6]. Nature does not seem to shield the macro from the micro in the way that a computer shields bits from electrons. A single photon can save a cat’s life in the dark [42], and this kind of structured amplification (called emergence) from microscopic matter is happening continually all through biological tissue. There can thus be no “machine code of the brain”. The ordered flow of information from the microscopic is represented in Fig. 3c by the upward diagonal arrows. Memory recall and ‘spontaneous thought’ are seen as such upward cross-level emergences.
- This information flow is bidirectional. As we communicate with each other, we change each other’s gene expression through the downward diagonal arrows in Fig. 3c. Words (for example) cause spikes, spikes cause calcium flows and these send a message to the nucleus to make different proteins. See [9] for a mechanism by which this can be accomplished.[3] ‘One-shot’ learning is storage in a massively overcomplete distributed microscopic state space, and memories are accessed associatively through downward arrows from macroscopic neural patterns.
- The sensory-motor loop, properly considered, is an inter-level interaction. Going upward, behaviours like words emerge from spikes, and going downward, they are the (social) boundary condition for a listener’s neural activity, just as spikes are the neural boundary condition on the membrane for voltage flows in the dendrites.
- The sensory-motor loop is just one stage in a multi-resolution hierarchy of nested similar inter-level dynamics. It is not special, just as “agents” are not a special stage in this hierarchy, but merely a level of description.
- Spikes (for example) communicate with each other by two different mechanisms: an upward arc via the social network, and a downward arc via the dendritic network (see Fig. 3c). Viewed this way, there are no horizontal arrows, just flows of information up and down. Even messages between synapses in dendrites may be differently processed by different voltage or calcium-sensitive macromolecules in the post-synaptic density of the receiving synapse (the lowest downward arc in Fig. 3c).

[2] Functionalism is the idea that you could build a computer out of beer cans and string.

[3] The mechanism is the Endoplasmic Reticulum. Quoting Berridge: “the ER and plasma membrane form a binary membrane system that functions to regulate a variety of neuronal processes including excitability, associativity, neurotransmitter release, synaptic plasticity and gene transcription”.

- These processes are *all the same thing expressed at different resolutions*. The state vectors are transformed by dimensionality-reducing structural relations in 3c, and the causal relations are transformed with them.

A cross-level theory of learning would be a theory taking place in a causal structure like that of Fig. 3c. Normally in (Bayesian) learning theory, we separate the data $\mathbf{x}$ from the model (the learned weights $\mathbf{W}$), and define quantities like the likelihood $q(\mathbf{x}|\mathbf{W})$, the prior $q(\mathbf{W})$, and the posterior $q(\mathbf{W}|\mathbf{x})$. In a cross-level theory, the data (for example spike timings) is just the network traffic which flows through connections which are nodes in the network below, as described. Thus quantities like the likelihood and the posterior are actually conditional distributions of groupings of variables across levels. It is a tantalising task to connect the framework known as 'hierarchical Bayes' to the hierarchy of matter observed in experiments.

The key to this may lie in generalising to the learning-scenario a framework in physics known as the Renormalisation Group (RG) [50]. RG has been called the most important new mathematical idea discovered in the 20th century. It is what enabled the creation of quantum field theory from generic quantum mechanics. It is also the idea used to understand equilibrium phase transitions in certain physical systems (like the 2D Ising model). It allows one to investigate the changes of a physical system as one views it at different spatial scales. It applies to scale-invariant systems, the correlation functions of which change in simple ways as the resolution is altered. It does not apply in biology because biology has special scales, permitting the separation of nested networks that we have described.

As mentioned, the statistics of the activity in scale-invariant systems transform in simple and lawful ways as the scale is varied. RG provides equations that capture these constraints. These equations are used to solve for the partition function of the system, leading to the accurate prediction of important physical quantities. This partition function, Z, is exactly the same mathematical quantity we encountered before in eq.(4).

In machine learning, as indicated earlier, it is not Z but its *gradient* which matters (remember the first term in eq.(3)). Could it be that something like RG could enable the calculation of an appropriate learning gradient for the multiresolution state vectors shown in Fig. 3c?

The empirical relations between statistics of activity at different scales of the brain's material hierarchy (like the dendritic currents, spikes and local field potentials we have discussed) are only beginning to be unveiled by an array of sophisticated new multiresolution probes. The principles underlying the brain's multi-level statistical model of its sensory-motor world may even be identified by experimenters before the theorists get their act together.

Physics, biology, inference – the strings are not yet tied, the questions not even really formalised. In physics, even reductionism is not necessarily on solid ground [35]. Noise, signal, control, reward, agency, the brain as a logic device – the ideas that lay behind the cybernetics movement in the 1950s, and that have so heavily influenced our thinking today, are starting to look like they may not be the primitives of a future emergent understanding. What kind of statistical self-modelling is occuring in matter? How can we draw these loose strands together? The game is open. There's everything to play for.

ACKNOWLEDGMENTS

Thanks to Kilian Koepsell and Lucas Parra for many discussions on this, to Tim Blanche and Simon Osindero for permission to use figures from their work (in Tim's case of data not yet published in a journal [10]), to Geoff Hinton (and CIAR) and Terry Sejnowski for ideas and support, and to MaxEnt '07, CosNet/Complex '07 in Australia, and Okinawa Institute of Science and Technology in Japan for chances to air these issues The work was funded by the Swartz Foundation and the NSF Science of Learning Center (TDLC). As well as providing the primary funding, Jerry Swartz's encouragement, enthusiasm and perception have been a crucial help. Finally, thank you for countless discussions with Paul Bush, Rita Venturini and Zach Mainen and many others, especially my colleagues at the Redwood Center.

REFERENCES

1. Allen E.A et al. 2007. Transcranial magnetic stimulation elicits coupled neural and hemodynamic consequences. *Science*, **317**, 1918-21
2. Amari S.-I., Cichocki A. and Yang H.H. 1996. A New Learning Algorithm for Blind Signal Separation. *Advances in Neural Information Processing Systems 8*, 757-763, MIT Press, Cambridge MA
3. Amari S-I. 1997. Natural gradient works efficiently in learning. *Neural Computation*, **10**, 251-276
4. Bell A.J. and Sejnowski T.J. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, **7**, 6, 1129-1159
5. Bell A.J. and Sejnowski T.J. 1997. The 'independent' components of natural images are edge-filters. *Vision Research*, **37**, 3327-3338
6. Bell A.J. Levels and loops: the future of Artificial Intelligence and Neuroscience. 1999. *Phil. Trans. R. Soc. Lond. B* **354**, 2013-2020
7. Bell A.J. 2002. The co-information lattice. *4th International Symposium on Independent Component Analysis and Blind Source Separation*, Nara, Japan
8. Bell A.J. and Parra L.C. 2005. Maximising sensitivity in a spiking network. *Advances in Neural Information Processing Systems 17*, MIT Press
9. Berridge M.J. 1998. Neuronal calcium signaling. *Neuron*, **21**, 13-26
10. Blanche T.J., Freiwald W.A. and Swindale N.V. 2006. Neural sparseness in cat and monkey visual cortex studied with silicon polytrode arrays. *Society of Neuroscience Abstracts*, Atlanta, Georgia
11. Buzsaki G. 2006. *Rhythms of the brain.* Oxford Univ. Press
12. Cardoso J.-F. 1997. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, **4**, 4, 112-114
13. Caticha A. and Cafaro C. 2008. From information geometry to Newtonian dynamics. *In [31] (this volume)*
14. Canolty R.T. et al. 2006. High Gamma Power Is Phase-Locked to Theta Oscillations in Human Neocortex. *Science*, **313**, 5793, 1626-28
15. Dan, Y. and Poo, M.-m. 2004. Spike timing-dependent plasticity of neural circuits. *Neuron* **44**, 23-30
16. Dewar R.C. 2005. Maximum entropy production and the fluctuation theorem. *J. Phys. A: Math. Gen.*, **38**, L371-381
17. DeWeese M.R., Wehr M. and Zador A.M. 2003. Binary spiking in auditory cortex. *J. Neurosci.*, **23**, 7940
18. Ellis R. J. 2001. Macromolecular crowding: obvious but underappreciated. *Trends in Biochem. Sci.*, **26**, 10, 597-604
19. Ernst M.O. and Banks M.S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 6870, 429-33
20. Fries P. 2005. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.*, **9**, 474
21. Fries P., Nicolic D. and Singer W. 2007. The gamma cycle. *Trends Neurosci.*, **30**, 7, 309-16

22. Gerstner W. and Kistner W.M. 2002. *Spiking neuron models.* Cambridge University Press
23. Hancock J.F. Lipid rafts: contentious only from simplistic standpoints. 2006. *Nature Rev. Molec. Cell Biol.*, **7**, 456-462
24. Hinton G.E. and Sejnowski T.J. 1986. Learning and relearning in Boltzmann machines. In Rumelhart D.E. and McClelland J.L. (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, MIT Press
25. Hinton G.E. et al. 2001. A New View of ICA, *Proceedings of ICA 2001*, San Diego, CA.
26. Hinton, G.E. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, **14**, 1771-1800
27. Hyvärinen A. and Hoyer P. 2001. A Two-Layer Sparse Coding Model Learns Simple and Complex Cell Receptive Fields and Topography from Natural Images. *Vision Research*, **41**, 18, 2413-23
28. Jacobson K., Mouritsen O.G. and Anderson R.G.W. 2007. Lipid rafts: at a crossroads between cell biology and physics. *Nature Cell Biol.*, **9**, 1, 7-14
29. Karklin Y. and Lewicki M.S. 2005. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, **17**, 2, 397-423
30. Klyubin A.S., Polani D. and Nehaniv C.L. 2007. Representations of space and time in the maximization of information flow in the perception-action loop. *Neural Computation*, **19**, 2387-2432
31. Knuth K.H. et al. 2007. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, NY, USA*, AIP Conf. Proc., Melville NY:AIP
32. Knuth K.H., Erner P.M. and Frasso S. 2008. Designing intelligent instruments. *[31] (this volume)*
33. Koepsell K. et al. 2008. Retinal oscillations carry visual information to cortex. *under review*
34. Lakatos P. et al. 2005. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.*, **94**, 1904-11
35. Laughlin R.B. 2006. *A different universe: reinventing physics from the bottom down.* Basic Books, NY
36. Maturana H. and Varela F. 1973. *Autopoiesis and cognition: the realization of the living.* Dordecht: D. Reidel
37. Moon T.K. and Gunther J.H. 2002. Contravariant adaptation on structured matrix spaces. *Signal Processing*, **82**, 10, 1389-1410
38. Olshausen B. and Field D.F. 1997. Emergence of simple-sell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607-609
39. Osindero S., Welling M. and Hinton, G.E. 2006. Topographic Product Models Applied To Natural Scene Statistics. *Neural Computation*, **18**, 2
40. Parra L.C., Beck J. and Bell A.J. 2008. On the maximisation of information flow between spiking neurons *under review*
41. Radman T. et al. 2007. Spike timing amplifies the effect of electric fields on neurons: implications for endogenous field effects, *J. Neurosci.*, **27**, 3030-3036
42. Rieke F. and Baylor D.A. 1998. Single-photon detection by rod cells of the retina. *Rev. Mod. Phys.* **70**, 1027-36
43. Sejnowski T.J. and Paulsen O. 2006. Network oscillations: emerging computational principles. *J. Neurosci.*, **26**, 6, 1673-76
44. Sheng M. and Hoogenraad C.C. 2007. The postsynaptic architecture of excitatory synapses: a more quantitative view, *Annu. Rev. Biochem.*, **76**, 823-47
45. Simoncelli E.P. and Olshausen B.A. 2001. Natural image statistics and neural representation, *Annu Rev Neurosci.* **24**, 1193-216
46. Spitzer J.J. and Poolman B. 2005. Electrochemical structure of the crowded cytoplasm. *Trends in Biochem. Sci.*, **30**, 10, 536-541
47. Theis F.J. 2005. Gradients on matrix manifolds and their chain rule. *Neural Information Processing Letters*, **9**, 1, 1-13
48. van Hateren J.H. and Ruderman D.L. 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, **265**, 2315-2320
49. Welling, M., Hinton, G.E. and Osindero, S. 2003. Learning Sparse Topographic Representations with Products of Student-t Distributions. *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA
50. Zinn-Justin J. 2002. *Quantum field theory and critical phenomena.* Oxford Univ. Press