Neural Networks 22 (2009) 1093–1104

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Cortical circuits for perceptual inference

Karl Friston *, Stefan Kiebel

The Wellcome Trust Centre of Neuroimaging, University College London, Queen Square, London WC1N 3BG, United Kingdom

ARTICLE INFO

Article history:
Received 26 January 2009
Received in revised form 14 May 2009
Accepted 14 July 2009

Keywords:
Generative models
Predictive coding
Hierarchical
Dynamic
Nonlinear
Circuits
Variational
Birdsong
Free-energy

ABSTRACT

This paper assumes that cortical circuits have evolved to enable inference about the causes of sensory input received by the brain. This provides a principled specification of *what* neural circuits have to achieve. Here, we attempt to address *how* the brain makes inferences by casting inference as an optimisation problem. We look at how the ensuing recognition dynamics could be supported by directed connections and message-passing among neuronal populations, given our knowledge of intrinsic and extrinsic neuronal connections. We assume that the brain models the world as a dynamic system, which imposes causal structure on the sensorium. Perception is equated with the optimisation or inversion of this internal model, to explain sensory input. Given a model of how sensory data are generated, we use a generic variational approach to model inversion to furnish equations that prescribe recognition; i.e., the dynamics of neuronal activity that represents the causes of sensory input. Here, we focus on a model whose hierarchical and dynamical structure enables simulated brains to recognise and predict sequences of sensory states. We first review these models and their inversion under a variational free-energy formulation. We then show that the brain has the necessary infrastructure to implement this inversion and present stimulations using synthetic birds that generate and recognise birdsongs.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

This paper looks at the functional architecture of cortical circuits from the point of view of perception; namely, the fitting or inversion of internal models of sensory data by the brain. Critically, the nature of this inversion lends itself to a relatively simple neural network implementation that shares many formal similarities with real cortical hierarchies in the brain. The basic idea that the brain uses hierarchical inference has been described in a series of papers (Friston, 2005; Friston, Kilner, & Harrison, 2006; Mumford, 1992; Rao & Ballard, 1998). These papers suggest that the brain uses *empirical* Bayes for inference about its sensory input, given the hierarchical organisation of cortical systems. Here, we focus on how neural networks could be configured to invert these models and deconvolve sensory causes from sensory input.

This paper comprises three sections. In the first, we introduce a free-energy formulation of model inversion or perception, which is then applied to a specific class of models that we assume the brain uses — hierarchical dynamic models. An important aspect of these models is their formulation in generalised coordinates of motion. This lends them a hierarchical form in both structure and dynamics, which can be exploited during inversion. In the second section, we show how inversion can be formulated as a simple gradient descent using neuronal networks and relate these to cortical circuits in the brain. In the final section, we consider how evoked brain responses might be understood in terms of perceptual inference and categorisation, using the schemes of the preceding section.

2. The free-energy formulation

This section considers the problem of inverting generative models of sensory data and provides a summary of the material in Friston (2008). This problem is addressed using *ensemble learning* or *Variational Bayes*. These are generic approaches to model inversion that provide an approximation to the conditional density $p(\vartheta|\tilde{y})$ on some causes ϑ of generalised sensory input, $\tilde{y} = [y, y', y'', \ldots]^{\mathrm{T}}$. Generalised input (e.g., the intensity of photoreceptor stimulation) includes the input, its velocity, acceleration, jerk, *etc*. Causes are quantities in the environment that generate sensory input (e.g., the orientation of an object in the visual field). The approximation of the conditional density (i.e., the probability of a particular set of causes given sensory input) is achieved by optimising a recognition density $q(\vartheta)$ with respect to a bound on the surprise or negative log-evidence $-\ln p(\tilde{y})$ of the sensory input, as we will see next (Feynman, 1972; Friston, 2005; Friston et al., 2006; Hinton & von Cramp, 1993; MacKay, 1995; Neal & Hinton, 1998). This bound is called free-energy

* Corresponding address: The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, Queen Square, London, WC1N 3BG, United Kingdom. Tel.: +44 207 833 7454; fax: +44 207 813 1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.07.023

$$F = C - \ln p(\tilde{y})$$
$$C = \left\langle \ln \frac{q(\vartheta)}{p(\vartheta|\tilde{y})} \right\rangle_q . \tag{1}$$

The free-energy comprises a cross-entropy or divergence term $C \geq 0$ and surprise. By Gibb's inequality, the divergence is greater than zero, with equality when $q(\vartheta) = p(\vartheta|\tilde{y})$; i.e., when the recognition density is the posterior or conditional density on the causes of sensory input. The recognition density can be optimised to minimise this bound and implicitly minimise the divergence between the recognition density and the conditional density we seek (Friston, 2005; Hinton & von Cramp, 1993; MacKay, 1995; Neal & Hinton, 1998). In summary, the recognition density, induces a free-energy bound, which converts a difficult integration problem (inherent in computing the exact conditional density) into an easier optimisation problem.

The bound can be evaluated easily because it is a function of $q(\vartheta)$ and some generative model $p(\tilde{y}, \vartheta)$ entailed by the brain

$$\begin{aligned} F &= \left\langle \ln q(\vartheta) - \ln p(\vartheta|\tilde{y}) - \ln p(\tilde{y}) \right\rangle_q \\ &= \langle \ln q(\vartheta) \rangle_q + \langle U(\vartheta) \rangle_q \\ U(\vartheta) &:= -\ln p(\tilde{y}, \vartheta). \end{aligned} \tag{2}$$

Here, we have expressed the free-energy in terms of the negentropy of $q(\vartheta)$ and an expected Gibb's energy $-U(\vartheta)$. This energy is usually specified in terms of a likelihood and prior; $U(\vartheta) = -\ln p(\tilde{y}|\vartheta) - \ln p(\vartheta)$, which define a generative model. This is important because it shows that we need a generative model in order to evaluate free-energy. The likelihood model just quantifies the probability of any sensations, given their cause; while the prior model encodes prior beliefs about the probability of those causes being present. It is fairly easy to show that minimising free-energy corresponds to finding a recognition density that predicts sensory input accurately, while suppressing its complexity.

If we assume that the recognition density $q(\vartheta) = N(\tilde{\mu}, \tilde{\Sigma})$ is Gaussian (the Laplace assumption), then we can express free-energy in terms of its sufficient statistics (i.e., its mean and covariance: $\tilde{\mu}, \tilde{\Sigma}$)

$$F = U(\tilde{\mu}) - \frac{1}{2} tr(\tilde{\Sigma} \nabla^2 U) - \frac{1}{2} \ln |\tilde{\Sigma}| - \frac{n}{2} \ln 2\pi e. \tag{3}$$

Here n is the number of unknown causes. We can now minimise free-energy w.r.t. the conditional covariances by finding the value that renders its gradient zero

$$\begin{aligned} &F_\Sigma = -\frac{1}{2}\tilde{\Pi} - \frac{1}{2}\nabla^2 U = 0 \Rightarrow \\ &\tilde{\Pi} = \nabla^2 U(\tilde{\mu}) \end{aligned} \tag{4}$$

where a subscript means differentiation; i.e., $F_\Sigma = \partial F / \partial \Sigma$ is the free-energy gradient w.r.t. the conditional covariance, Here, the conditional precision $\tilde{\Pi} = \tilde{\Sigma}^{-1}$ is the inverse covariance. Critically, the conditional precision is just a function of the mean and does not have to be encoded explicitly. This means we can simplify the expression for free-energy by eliminating the curvatures $\nabla^2 U$ of Gibb's energy

$$F = U(\tilde{\mu}) - \frac{1}{2} \ln |\tilde{\Sigma}| - \frac{n}{2} \ln 2\pi . \tag{5}$$

Now, the only unknown quantities are the conditional means of the causes, which only have to minimise Gibb's energy because this is the only term that depends on them. In this paper, we will focus on time-varying causes or states of the environment: $\tilde{u}(t) \subset \vartheta$. The values we seek are the solutions to the following differential equations.

$$\begin{aligned} \dot{\tilde{\mu}}^u &= D\tilde{\mu}^u - U_{\tilde{u}} \\ &\Leftrightarrow \\ \dot{\mu}^u &= \mu'^u - U_u \\ \dot{\mu}'^u &= \mu''^u - U'_u \\ \dot{\mu}''^u &= \mu'''^u - U''_u \\ &\vdots \end{aligned} \tag{6}$$

This solution (which is stationary in a frame of reference that moves with its generalised motion), minimises free-energy

$$\begin{aligned} &\dot{\tilde{\mu}}^u - D\tilde{\mu}^u = 0 \Rightarrow \\ &U_{\tilde{u}} = 0 \Rightarrow F_{\tilde{u}} = 0. \end{aligned} \tag{7}$$

This construction ensures that when Gibb's energy is minimised and $U_{\tilde{u}} = 0$, the mean of the motion is the motion of the mean; i.e., $\dot{\tilde{\mu}}^u = D\tilde{\mu}^u$. Here D is a derivative matrix operator with identity matrices along the first leading diagonal.

Eq. (7) prescribes recognition dynamics that track time-varying causes or states of the world and can be thought of as a gradient descent in a moving frame of reference. The recognition dynamics for time-invariant causes (i.e., parameters $\theta \subset \vartheta$, like rate constants) have a different form, because we know *a priori* their generalised motion is zero. In this paper, we will assume the parameters have already been learnt and focus on recognising hidden states of the environment. In summary, we have derived recognition dynamics for expected environmental states, which cause sensations. The solution to these equations minimise Gibb's energy and (under the Laplace assumption) free-energy, which is an upper bound on their surprise. Finding these solutions corresponds to perceptual inference. The precise form of Eq. (7) depends on the generative model that defines Gibb's energy. Next, we examine forms associated with hierarchical dynamic models.

2.1. Hierarchical dynamic models

This section introduces a general class of generative models that the brain may use for perception. We will start with simple dynamic models and then deal with hierarchical cases later. Consider a state-space model that describes the evolution of states in the world and how they map to sensory input

$$\begin{aligned} y &= g(x, v) + z \\ \dot{x} &= f(x, v) + w. \end{aligned} \tag{8}$$

Here, the functions f and g are parameterised by $\theta \subset \vartheta$ (which are omitted from the following expressions for clarity). These functions correspond to equations of motion and an observer function, respectively. The states $v \subset u$ are variously referred to as sources or causal states. The hidden states $x \subset u$ mediate the influence of causal states on sensory data and endow the system with memory. We assume the random fluctuations z are analytic, such that the covariance of $\tilde{z} = [z, z', z'', \ldots]^{\mathrm{T}}$ is well defined; similarly for state noise, $w(t)$, which represents random fluctuations on the motion of the hidden states. Under local linearity assumptions, the generalised motion of the data or response $\tilde{y} = [y, y', y'', \ldots]^{\mathrm{T}}$ is given by

$$\begin{aligned} y &= g(x, v) + z & \dot{x} &= x' = f(x, v) + w \\ y' &= g_x x' + g_v v' + z' & \dot{x}' &= x'' = f_x x' + f_v v' + w' \\ y'' &= g_x x'' + g_v v'' + z'' & \dot{x}'' &= x''' = f_x x'' + f_v v'' + w'' \\ &\vdots & &\vdots \end{aligned} \tag{9}$$

We can write this generalised state-space model more compactly as

$$\begin{aligned}\tilde{y} &= \tilde{g} + \tilde{z}\\ D\tilde{x} &= \tilde{f} + \tilde{w}\end{aligned} \tag{10}$$

where the predicted response $\tilde{g} = [g, g', g'', \ldots]^{\mathrm{T}}$ and motion $\tilde{f} = [f, f', f'', \ldots]^{\mathrm{T}}$ are

$$\begin{array}{ll} g = g(x, v) & f = f(x, v)\\ g' = g_x x' + g_v v' & f' = f_x x' + f_v v'\\ g'' = g_x x'' + g_v v'' & f'' = f_x x'' + f_v v''\\ \vdots & \vdots \end{array} \tag{11}$$

Gaussian assumptions about the fluctuations $p(\tilde{z}) = N(\tilde{z} : 0, \tilde{\Sigma}^z)$ provide the form of the likelihood, $p(\tilde{y}|\tilde{x}, \tilde{v})$. Similarly, Gaussian assumptions about state noise $p(\tilde{w}) = N(\tilde{w} : 0, \tilde{\Sigma}^w)$ specify empirical priors, $p(\tilde{x}|\tilde{v})$ in terms of predicted motion

$$\begin{aligned} p\left(\tilde{y}, \tilde{x}, \tilde{v}\right) &= p(\tilde{y}|\tilde{x}, \tilde{v})p(\tilde{x}, \tilde{v})\\ p\left(\tilde{x}, \tilde{v}\right) &= p(\tilde{x}|\tilde{v})p(\tilde{v})\\ p\left(\tilde{y}|\tilde{x}, \tilde{v}\right) &= N(\tilde{y} : \tilde{g}, \tilde{\Sigma}^z)\\ p\left(\tilde{x}|\tilde{v}\right) &= N(D\tilde{x} : \tilde{f}, \tilde{\Sigma}^w). \end{aligned} \tag{12}$$

The covariances $\tilde{\Sigma}^z$ and $\tilde{\Sigma}^w$ or precisions $\tilde{\Pi}^z(\lambda)$ and $\tilde{\Pi}^w(\lambda)$ are functions of precision parameters, $\lambda \subset \vartheta$, which control the amplitude and smoothness of random fluctuations. Generally, these covariances factorise; $\tilde{\Sigma}^{\cdot} = \Sigma^{\cdot} \otimes R^{\cdot}$ into a covariance proper and a matrix of correlations $R^{\cdot}$ among generalised motion that encodes an autocorrelation function.

2.1.1. Hierarchical forms

Hierarchical dynamic models with m levels have the following form, which generalises the $m = 1$ model above

$$\begin{aligned} y &= g(x^{(1)}, v^{(1)}) + z^{(1)}\\ \dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + w^{(1)}\\ &\vdots\\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + z^{(i)}\\ \dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + w^{(i)}\\ &\vdots\\ v^{(m)} &= z^{(m+1)}. \end{aligned} \tag{13}$$

Again, $f^{(i)} = f(x^{(i)}, v^{(i)})$ and $g^{(i)} = g(x^{(i)}, v^{(i)})$ are continuous nonlinear functions of the states. The innovations $z^{(i)}$ and $w^{(i)}$ are conditionally independent fluctuations that enter each level of the hierarchy. These play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causal states $v = [v^{(1)}, \ldots, v^{(m)}]^{\mathrm{T}}$ link levels, whereas the hidden states $x = [x^{(1)}, \ldots, x^{(m)}]^{\mathrm{T}}$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. Inputs from higher levels can enter nonlinearly into the state equations and can be regarded as changing its control parameters to produce quite complicated generalised convolutions with deep (i.e., hierarchical) structure.

In summary, hierarchical dynamic models are about as complicated as one could imagine; they comprise causal and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. A key aspect of these models is their hierarchical form, which induces empirical priors on the causal states. See Kass and Steffey (1989) for a discussion of approximate Bayesian inference in conditionally independent hierarchical models of static data.

2.1.2. Energy functions

We can now write down Gibb's energy for these generative models, which has a simple quadratic form (ignoring constants)

$$\begin{aligned} U &= \ln p\left(\tilde{y}, \tilde{x}, \tilde{v}, \theta, \lambda\right) = \frac{1}{2}\ln|\tilde{\Pi}| - \frac{1}{2}\tilde{\varepsilon}^{\mathrm{T}}\tilde{\Pi}\tilde{\varepsilon}\\ \tilde{\Pi} &= \begin{bmatrix} \tilde{\Pi}^z & 0\\ 0 & \tilde{\Pi}^w \end{bmatrix}\\ \tilde{\varepsilon} &= \begin{bmatrix} \tilde{\varepsilon}^v = \tilde{y} - \tilde{g}\\ \tilde{\varepsilon}^x = D\tilde{x} - \tilde{f} \end{bmatrix}. \end{aligned} \tag{14}$$

The auxiliary variables $\tilde{\varepsilon}(t)$ comprise prediction errors for the generalised response and motion of hidden states, where $\tilde{g}(t)$ and $\tilde{f}(t)$ are the respective predictions, whose precision is encoded by $\tilde{\Pi}(\lambda)$. These prediction errors provide a compact way to express Gibb's energy and, as we will see below, lead to very simple recognition schemes. For hierarchical models, the prediction error on the response is supplemented with prediction errors on the causal states

$$\varepsilon^v = \begin{bmatrix} y\\ v^{(1)}\\ \vdots\\ v^{(m)} \end{bmatrix} - \begin{bmatrix} g^{(1)}\\ g^{(2)}\\ \vdots\\ 0 \end{bmatrix}. \tag{15}$$

Note that the data enter the prediction error at the lowest level. At intermediate levels, the prediction errors, $v^{(i-1)} - g^{(i)}$ mediate empirical priors on the causal states.

2.2. Summary

In this section, we have seen how the inversion of dynamic models can be formulated as an optimisation of free-energy. By assuming a Gaussian (Laplace) approximation to the conditional density, one can reduce optimisation to finding the conditional means of the unknown causes of sensory data. This can be formulated as a gradient ascent in a frame of reference that moves along the path encoded in generalised coordinates (Eq. (6)). The only thing needed to implement this recognition scheme is Gibb's energy, which is specified by a generative model. We have looked at hierarchical dynamic models, whose form provides empirical priors or constraints on inference at both a structural and dynamic level (Eq. (14)). The *structural* priors arise from coupling different levels of the hierarchy with causal states and the *dynamic* priors emerge by coupling different levels of generalised motion of the hidden states. We can now look at the recognition dynamics entailed by these models, in the context of neuronal processes in the brain.

3. Hierarchical models in the brain

A key architectural principle of the brain is its hierarchical organisation (Felleman & Van Essen, 1991; Maunsell & van Essen, 1983; Zeki & Shipp, 1988). This has been established most thoroughly in the visual system, where lower (primary) areas receive sensory input and higher areas adopt a multimodal or associational role. The neurobiological notion of a hierarchy rests upon the distinction between forward and backward connections (Angelucci et al., 2002; Felleman & Van Essen, 1991; Murphy & Sillito, 1987; Rockland & Pandya, 1979; Sherman & Guillery, 1998). This distinction is based upon the specificity of cortical layers that are the predominant sources and origins of extrinsic connections. Forward connections arise largely in superficial pyramidal cells, in supra-granular layers and terminate on spiny stellate cells of layer four in higher cortical areas (DeFelipe, Alonso-Nanclares, & Arellano, 2002; Felleman & Van Essen, 1991). Conversely, backward connections arise largely from deep pyrami-

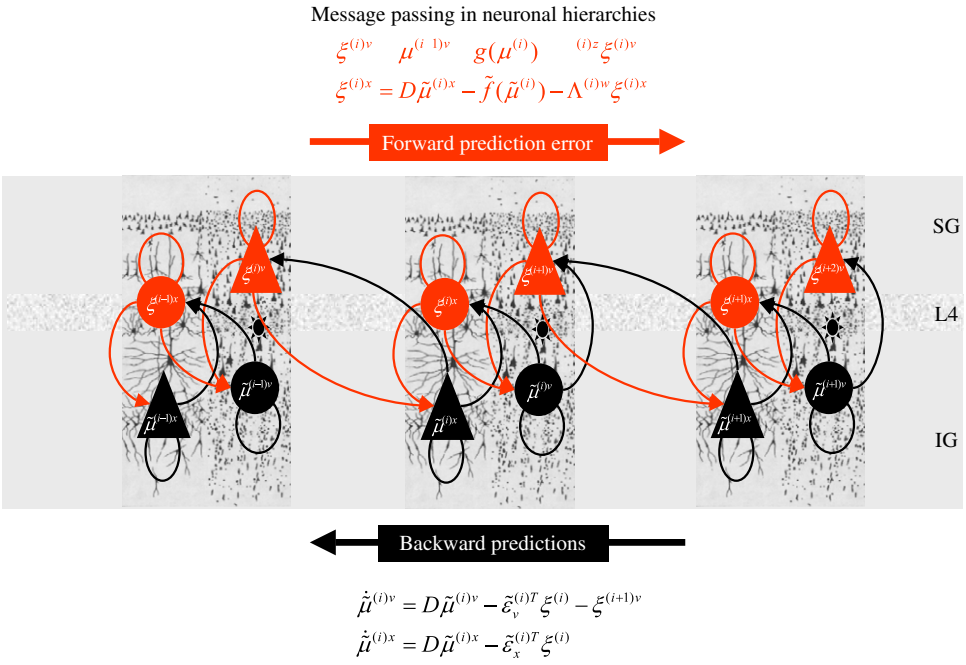

Fig. 1. Schematic detailing the neuronal architectures that encode a recognition density on the states of a hierarchical model. This schematic shows the speculative cells of origin of forward driving connections that convey prediction error from a lower area to a higher area and the backward connections that are used to construct predictions. These predictions try to explain away input from lower areas by suppressing prediction error. In this scheme, the sources of forward connections are superficial pyramidal cell populations and the sources of backward connections are deep pyramidal cell populations. The differential equations relate to the optimisation scheme detailed in the main text. The state-units and their efferents are in black and the error-units in red, with causal states on the right and hidden states on the left. For simplicity, we have assumed the output of each level is a function of, and only of, the hidden states. This induces a hierarchy over levels and, within each level, a hierarchical relationship between states, where causal states predict hidden states. This schematic shows how the neuronal populations may be deployed hierarchically within three cortical areas (or macro-columns). Within each area the cells are shown in relation to the laminar structure of the cortex that includes supra-granular (**SG**) granular (**L4**) and infra-granular (**IG**) layers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dal cells in infra-granular layers and target cells in the infra- and supra-granular layers of lower cortical areas. Intrinsic connections mediate lateral interactions between neurons that are a few millimetres away. There is a key functional asymmetry between forward and backward connections that renders backward connections more modulatory or nonlinear in their effects on neuronal responses (Sherman & Guillery, 1998; see also Hupe et al., 1998). This is consistent with the deployment of voltage-sensitive NMDA receptors in supra-granular layers that are targeted by backward connections (Rosier, Arckens, Orban, & Vandesande, 1993). Typically, the synaptic dynamics of backward connections have slower time constants. This has led to the notion that forward connections are driving and illicit an obligatory response in higher levels, whereas backward connections have both driving and modulatory effects and operate over larger spatial and temporal scales. This hierarchical aspect of the brain's functional anatomy speaks to hierarchical models of sensory input. We now consider how this functional architecture can be understood under the inversion of hierarchical models by the brain.

3.1. Perceptual inference

If we assume that the activity of neurons encode the conditional mean of external states causing sensory data, then Eq. (6) specifies the neuronal dynamics entailed by recognising states of the world from sensory data. Using Gibb's energy in Eq. (14) we have

$$\begin{aligned}\dot{\tilde{\mu}}^u &= D\tilde{\mu}^u - U_{\tilde{u}} \\ &= D\tilde{\mu}^u - \tilde{\varepsilon}_u^{\mathrm{T}}\xi \\ \xi &= \tilde{\Pi}\tilde{\varepsilon} = \tilde{\varepsilon} - \Lambda\xi \\ \tilde{\Pi} &= \begin{bmatrix}\tilde{\Pi}^z & \\ & \tilde{\Pi}^w\end{bmatrix}.\end{aligned} \tag{16}$$

Eq. (16) describes how neuronal states self-organise, when exposed to sensory input. Its form is quite revealing and suggests two distinct populations of neurons; causal or hidden *state-units* whose activity encodes $\tilde{\mu}^u := \tilde{\mu}(t)$ and *error-units* encoding precision-weighted prediction error $\xi = \tilde{\Pi}\tilde{\varepsilon}$, with one error-unit for each state. Furthermore, the activities of error-units are a function of the states and the dynamics of state-units are a function of prediction error. This means the two populations pass messages to each other and to themselves. The messages passed within the states, $D\tilde{\mu}$ mediate empirical priors on their motion, while $-\Lambda\xi$ mediates precision-dependent modulation of prediction errors. The matrix $\Lambda = \tilde{\Sigma} - 1$ can be thought of encoding self-inhibition, which is modulated by precision (where precision might be encoded by neuromodulatory neurotransmitters like dopamine and acetylcholine).

3.2. Hierarchical message-passing

If we unpack these equations, we can see the hierarchical nature of the implicit message-passing

$$\begin{aligned}\dot{\tilde{\mu}}^{(i)v} &= D\tilde{\mu}^{(i)v} - \tilde{\varepsilon}_v^{(i)T}\xi^{(i)} - \xi^{(i+1)v} \\ \dot{\tilde{\mu}}^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{\varepsilon}_x^{(i)T}\xi^{(i)} \\ \xi^{(i)v} &= \tilde{\mu}^{(i-1)v} - \tilde{g}(\tilde{\mu}^{(i)}) - \Lambda^{(i)z}\xi^{(i)v} \\ \xi^{(i)x} &= D\tilde{\mu}^{(i)x} - \tilde{f}(\tilde{\mu}^{(i)}) - \Lambda^{(i)w}\xi^{(i)x}.\end{aligned} \tag{17}$$

This shows that error-units receive messages from the states in the same level and the level above, whereas states are driven by error-units in the same level and the level below (see Fig. 1). Critically, inference requires only the prediction error from the lower level $\xi^{(i)}$ and the level in question, $\xi^{(i+1)}$. These provide bottom-up and lateral messages that drive conditional expectations $\tilde{\mu}^{(i)}$

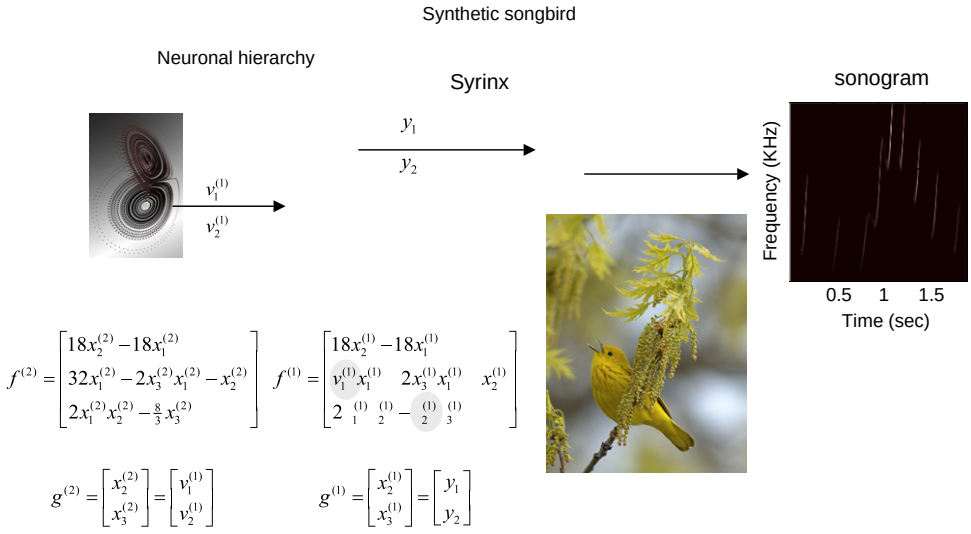

Fig. 2. Schematic showing the construction of the generative model for birdsongs. This comprises two Lorenz attractors where the higher attractor delivers two control parameters (grey circles) to a lower-level attractor, which, in turn, delivers two control parameters to a synthetic syrinx to produce amplitude and frequency modulated stimuli. This stimulus is represented as a sonogram in the right panel. The equations represent the hierarchical dynamic model in the form of Eq. (13).

towards a better prediction, to explain away the prediction error in the level below. These top-down and lateral predictions correspond to $\tilde{g}^{(i)}$ and $\tilde{f}^{(i)}$. This is the essence of recurrent message-passing between hierarchical levels to optimise free-energy or suppress prediction error; i.e., recognition dynamics. In summary, all connections between error- and state-units are reciprocal but the only connections that link levels are forward connections conveying prediction error to state-units and reciprocal backward connections that mediate predictions. This sort of scheme is referred to as predictive coding (Rao & Ballard, 1998).

We can identify error-units with superficial pyramidal cells, because the only messages that pass up the hierarchy are prediction errors and superficial pyramidal cells originate forward connections in the brain. This is useful because it is these cells that are primarily responsible for electroencephalographic (EEG) signals that can be measured non-invasively. Similarly, the only messages that are passed down the hierarchy are the predictions from state-units that are necessary to form prediction errors in lower levels. The sources of extrinsic backward connections are deep pyramidal cells; suggesting that these encode the expected causes of sensory states (see Mumford, 1992 and Fig. 1). Critically, the motion of each state-unit is a linear mixture of bottom-up prediction error (see Eq. (17)). This is exactly what is observed physiologically; bottom-up driving inputs elicit obligatory responses that do not depend on other bottom-up inputs. The prediction error itself is formed by predictions conveyed by backward and lateral connections. These influences embody the nonlinearities implicit in $\tilde{g}^{(i)}$ and $\tilde{f}^{(i)}$. Again, this is entirely consistent with the nonlinear or modulatory characteristics of backward connections.

It has been shown recently that hierarchical architectures (cf, Fig. 1) can be reformulated as a specific type of biased competition, where state-units receive messages from lower-level error-units and direct inputs from higher-level state-units (replacing lateral inputs from error-units in the original predictive coding scheme based on Kalman filtering; Rao & Ballard, 1998). It has been argued that this architecture provides a more realistic model of backward connections in cortex (Spratling, 2008a, 2008b) and usefully connects predictive coding, Kalman filtering and biased competition.

A related Bayesian algorithm called belief-propagation (Dean, 2006; Hinton, Osindero, & Teh, 2006; Lee & Mumford, 2003; Rao, 2006) also rests on message-passing. In these schemes, the messages are not prediction errors but, like prediction errors, are defined self-consistently, in terms of likelihoods and empirical priors. Critically, the belief-propagation algorithm can be derived by minimising free-energy (Yedidia, Freeman, & Weiss, 2005); for example, it can be shown that the Kalman filter is a special case of belief-propagation. This speaks to formal similarities between predictive coding, Bayesian filtering and belief-propagation, which could be implemented by recursive message-passing in the brain and understood in terms of free-energy optimisation.

3.3. Summary

In summary, we have seen how the inversion of a generic hierarchical and dynamical model of sensory inputs can be transcribed onto neuronal quantities that optimise a free-energy bound on surprise. This optimisation corresponds, under some simplifying assumptions, to suppression of prediction error at all levels in a cortical hierarchy. This suppression rests upon a balance between bottom-up (prediction error) and top-down (empirical prior) influences. In the final section, we use this scheme to simulate neuronal responses. Specifically, we look at the electrophysiological correlates of prediction error and ask whether we can understand some common phenomena in event-related potential (ERP) research in terms of the free-energy formulation and message-passing in the brain.

4. Birdsong and attractors

In this section, we examine a system that uses hierarchical dynamics as a generative model of sensory input. The aim of this section is to provide some face-validity for the functional deconstruction of extrinsic and intrinsic circuits in the previous section. To do this, we try to show how empirical measures of neuronal processes can be reproduced using simulations based on the theoretical analysis above. The example we use is birdsong and the empirical measures we focus on are local field potentials (LFP) or evoked (ERP) responses that can be recorded non-invasively. The material in section is based on the simulations described in Friston and Kiebel (2009).

We first describe our model of birdsong and demonstrate the nature and form of this model through simulated lesion experiments. We then use simplified versions of this model to show how attractors can be used to categorise sequences of stimuli quickly and efficiently. Throughout this section, we will exploit the