ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

A tutorial on the free-energy framework for modelling perception and learning

Rafal Bogacz *

MRC Unit for Brain Network Dynamics, University of Oxford, Mansfield Road, Oxford, OX1 3TH, UK
Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK

HIGHLIGHTS

- Bayesian inference about stimulus properties can be performed by networks of neurons.
- Learning about statistics of stimuli can be achieved by Hebbian synaptic plasticity.
- Structure of the model resembles the hierarchical organization of the neocortex.

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

This paper provides an easy to follow tutorial on the free-energy framework for modelling perception developed by Friston, which extends the predictive coding model of Rao and Ballard. These models assume that the sensory cortex infers the most likely values of attributes or features of sensory stimuli from the noisy inputs encoding the stimuli. Remarkably, these models describe how this inference could be implemented in a network of very simple computational elements, suggesting that this inference could be performed by biological networks of neurons. Furthermore, learning about the parameters describing the features and their uncertainty is implemented in these models by simple rules of synaptic plasticity based on Hebbian learning. This tutorial introduces the free-energy framework using very simple examples, and provides step-by-step derivations of the model. It also discusses in more detail how the model could be implemented in biological neural circuits. In particular, it presents an extended version of the model in which the neurons only sum their inputs, and synaptic plasticity only depends on activity of pre-synaptic and post-synaptic neurons.

1. Introduction

The model of Friston (2005) and the predictive coding model of Rao and Ballard (1999) provide a powerful mathematical framework to describe how the sensory cortex extracts information from noisy stimuli. The predictive coding model (Rao & Ballard, 1999) suggests that visual cortex infers the most likely properties of stimuli from noisy sensory input. The inference in this model is implemented by a surprisingly simple network of neuron-like nodes. The model is called "predictive coding", because some of the nodes in the network encode the differences between inputs and predictions of the network. Remarkably, learning about features present in sensory stimuli is implemented by simple Hebbian synaptic plasticity, and Rao and Ballard (1999) demonstrated that the model presented with natural images learns features resembling receptive fields of neurons in the primary visual cortex.

Friston (2005) has extended the model to also represent uncertainty associated with different features. He showed that learning about the variance and co-variance of features can also be implemented by simple synaptic plasticity rules based on Hebbian learning. As the extended model (Friston, 2005) learns the variance and co-variance of features, it offers several new insights. First, it describes how the perceptual systems may differentially weight sources of sensory information depending on their level of noise. Second, it shows how the sensory networks can learn to recognize features that are encoded in the patterns of covariance between inputs, such as textures. Third, it provides a natural way to implement attentional modulation as the reduction in variance of the attended features (we come back to these insights in Discussion). Furthermore, Friston (2005) pointed out that this model can be viewed as an approximate Bayesian inference based on minimization of a function referred to in statistics as free-energy. The

* Correspondence to: Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK.
E-mail address: rafal.bogacz@ndcn.ox.ac.uk.

http://dx.doi.org/10.1016/j.jmp.2015.11.003

free-energy framework (Friston, 2003, 2005) has been recently extended by Karl Friston and his colleagues to describe how the brain performs different cognitive functions including action selection (FitzGerald, Schwartenbeck, Moutoussis, Dolan, & Friston, 2015; Friston et al., 2013). Furthermore, Friston (2010) proposed that the free-energy theory unifies several theories of perception and action which are closely related to the free-energy framework.

There are many articles which provide an intuition for the free-energy framework and discuss how it relates with other theories and experimental data (Friston, 2003, 2005, 2010; Friston et al., 2013). However, the description of mathematical details of the theory in these papers requires a very deep mathematical background. The main goal of this paper is to provide an easy to follow tutorial on the free-energy framework. To make the tutorial accessible to a wide audience, it only assumes basic knowledge of probability theory, calculus and linear algebra. This tutorial is planned to be complementary to existing literature so it does not focus on the relationship to other theories and experimental data, and on applications to more complex tasks which are described elsewhere (Friston, 2010; Friston et al., 2013).

In this tutorial we also consider in more detail the neural implementation of the free-energy framework. Any computational model would need to satisfy the following constraints to be considered biologically plausible:

1. Local computation: A neuron performs computations only on the basis of the activity of its input neurons and synaptic weights associated with these inputs (rather than information encoded in other parts of the circuit).
2. Local plasticity: Synaptic plasticity is only based on the activity of pre-synaptic and post-synaptic neurons.

The model of Rao and Ballard (1999) fully satisfied these constraints. The model of Friston (2005) did not satisfy them fully, but we show that after small modifications and extensions it can satisfy them. So the descriptions of the model in this tutorial slightly differ in a few places or extend the original model to better explain how the proposed computation could be implemented in the neural circuits. All such differences or extensions are indicated by footnotes or in text, and the original model is presented in Appendix A.

It is commonly assumed in theoretical neuroscience, (O'Reilly & Munakata, 2000) that the basic computations a neuron performs are the summation of its input weighted by the strengths of synaptic connections, and the transformation of this sum through a (monotonic) function describing the relationship between neurons' total input and output (also termed firing-Input or f-I curve). Whenever possible, we will assume that the computation of the neurons in the described model is limited to these computations (or even just to linear summation of inputs).

We feel that the neural implementation of the model is worth considering, because if the free-energy principle indeed describes the computations in the brain, it can provide an explanation for why the cortex is organized in a particular way. However to gain such insight it is necessary to start comparing the neural networks implementing the model with those in the real brain. Consequently, we consider in this paper possible neural circuits that could perform the computations required by the theory. Although the neural implementations proposed here are not the only possible ones, it is worth considering them as a starting point for comparison of the model with details of neural architectures in the brain. We hope that such comparison could iteratively lead to refined neural implementations that are more and more similar to real neural circuits.

To make this tutorial as easy to follow as possible we introduce the free-energy framework using a simple example, and then illustrate how the model can scale up to more complex neural architectures. The tutorial provides step-by-step derivation of the model. Some of these derivations are straightforward, and we feel that it would be helpful for the reader to do them on their own to gain a better understanding of the model and to "keep in mind" the notation used in the paper. Such straightforward derivations are indicated by "(TRY IT YOURSELF)", so after encountering such label we recommend trying to do the calculation described in the sentence with this label and then compare the obtained results with those in the paper. To illustrate the model we include simple simulations, but again we feel it would be helpful for a reader to perform them on their own, to get an intuition for the model. Therefore we describe these simulations as exercises.

The paper is organized as follows. Section 2 introduces the model using a very simple example using as basic mathematical concepts as possible, so it is accessible to a particularly wide audience. Section 3 provides mathematical foundations for the model, and shows how the inference in the model is related to minimization of free-energy. Section 4 then shows how the model scales up to describe the neural circuits in sensory cortex. In these three sections we use notation similar to that used by Friston (2005). Section 5 describes an extended version of the model which satisfies the constraint of local plasticity described above. Finally, Section 6 discusses insights provided by the model.

2. Simplest example of perception

We start by considering in this section a simple perceptual problem in which a value of a single variable has to be inferred from a single observation. To make it more concrete, consider a simple organism that tries to infer the size or diameter of a food item, which we denote by v, on the basis of light intensity it observes. Let us assume that our simple animal has only one light sensitive receptor which provides it with a noisy estimate of light intensity, which we denote by u. Let g denote a non-linear function relating the average light intensity with the size. Since the amount of light reflected is related to the area of an object, in this example we will consider a simple function of $g(v) = v^2$. Let us further assume that the sensory input is noisy—in particular, when the size of food item is v, the perceived light intensity is normally distributed with mean $g(v)$, and variance Σ_u (although a normal distribution is not the best choice for a distribution of light intensity, as it includes negative numbers, we will still use it for a simplicity):

$$p(u|v) = f(u; g(v), \Sigma_u). \tag{1}$$

In Eq. (1) $f(x; \mu, \Sigma)$ denotes the density of a normal distribution with mean μ and variance Σ:

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{(x-\mu)^2}{2\Sigma}\right). \tag{2}$$

Due to the noise present in the observed light intensity, the animal can refine its guess for the size v by combining the sensory stimulus with the prior knowledge on how large the food items usually are, that it had learnt from experience. For simplicity, let us assume that our animal expects this size to be normally distributed with mean v_p and variance Σ_p (subscript p stands for "prior"), which we can write as:

$$p(v) = f(v; v_p, \Sigma_p). \tag{3}$$

Let us now assume that our animal observed a particular value of light intensity, and attempts to estimate the size of the food item on the basis of this observation. We will first consider an exact solution to this problem, and illustrate why it would be difficult to compute it in a simple neural circuit. Then we will present an approximate solution that can be easily implemented in a simple network of neurons.

Please cite this article in press as: Bogacz, R., A tutorial on the free-energy framework for modelling perception and learning. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.11.003