CellPress

Letter

Commentary on Sanborn and Chater: Posterior Modes Are Attractor Basins

Phillip M. Alday,[1] Matthias Schlesewsky,[1] and Ina Bornkessel-Schlesewsky[1,*]

Sanborn and Chater [1] propose an interesting theory of cognitive and brain function based on Bayesian sampling instead of asymptotic Bayesian inference. Their proposal unifies many current observations and models and, despite focusing primarily on cognitive phenomena, their work provides a springboard for unifying several proposed theories of brain function. It has the potential to serve as a bridge between three influential overarching current theories of cognitive and brain function: Bayesian models, Friston's [2–4] theory of cortical responses based on the free-energy principle, and attractor-basin dynamics [5,6]. Specifically, their proposal suggests a high-level perspective on Friston's theory, which in turn proposes a sampling procedure including appropriate handling of autocorrelation as well as a plausible neurobiological implementation. In turn, these two theories together link into attractor-basin dynamics at the level of networks (via Friston) as well as at the level of behaviour (via the relationship between the modes of prior and posterior distributions, as discussed by Sanborn and Chater). We will argue here that, by linking Sanborn and Chater's approach to neurobiological models based on the free-energy principle on the one hand, and attractor-basin dynamics on the other, the scope of their proposal can be broadened considerably. Moreover, a unified perspective along these lines provides an elegant solution to several of Sanborn and Chater's outstanding questions relating to the neural implementation of sampling.

Sanborn and Chater briefly touch upon the connection of their work to Friston's hierarchical model proposal, but only link it somewhat generally to the broad computational approach he has proposed for the representation and computation of these models in neural wetware [7]. Friston's other work, however, also describes and models the relationship between behaviour and the sampling procedure necessary for active Bayesian inference [8]. This is compatible with the phenomena that Sanborn and Chater describe as 'warm-up' tuning. Although Sanborn and Chater perhaps intentionally formulated their proposal in an implementation-agnostic manner, Friston's approach fills in the gaps regarding the neural implementation of sampling in an illuminating way that has been used to model a wide range of phenomena [3,4,8,9]. In particular, Friston's model provides large-scale suggestions – at the level of groups and networks of neurons – of how sampling is implemented (i.e., hierarchical structure [2,7] and active sampling [2]), and suggests that a simplified or indirect probability distribution is used, in other words free energy as a proxy for model evidence [8]. In this framework, autocorrelation is minimised via the active sampling procedure, but is also effectively handled by the iterative model updates – autocorrelated samples provide little additional information and thus small error signals. They therefore contribute in decreasing amounts to overall model convergence. This is related to the performance of particle (and the related Kalman) filters, which Sanborn and Chater also mention.

While Friston's model has been proposed as an overarching theory describing the brain as a whole, attractor basins have been proposed as an explanation of emergent classifier behaviour in dynamical systems such as neural networks (e.g., [5,6]). Attractor basins are steady states in a dynamical system that are associated with stable, high-firing rates in neural networks, whether computational or biological. In many ways this approach complements Friston's principles-first approach with an emergent, empirical observation, and indeed attractors can be seen as local optima (and hence stable states) in the free-energy landscape [10]. Sanborn and Chater's approach provides exactly this connection because the posterior modes in their sampling procedure are essentially attractor basins – areas of high probability density where a posterior belief tends to be drawn and to which estimates (beliefs) tend to converge (see their Figure 1). This observation goes beyond Sanborn and Chater's connection to the mechanisms of neural networks such as the Boltzmann machine and deep belief networks, and underscores the profound relationship between these two theories. The Bayesian combination of prior beliefs (including modes/pre-existing attractor basins) and likelihood (model based on current evidence), leading to a new set of modes when the evidence is strong enough but subject to bias from finite sampling, also provides a convenient explanation for the emergence of new attractor basins, in other words perceptual categories and decisions, over time. Because attractor networks provide a neurobiologically plausible way of modelling neural processes related to decision making and classification across a range of scales from perception [11,12] to more complex cognitive domains such as language processing [13], this isomorphism – both in sampling behaviour (as noted by Sanborn and Chater) and in large-scale behaviour via sampling modes and attractor basins – is particularly interesting.

Sanborn and Chater thus provide a compelling normative account of cognition based on Bayesian sampling that acquires a (neuronal) process theory under the free-energy principle.

Specifically, formulating Bayesian inference as gradient ascent (i.e., dynamical optimisation) on Bayesian model evidence(i.e., free energy) yields a plausible account of neuronal dynamics. These dynamics can either be regarded as a description of the average behaviour of a sampling-based scheme (e.g., particle filtering) or, equivalently, of the behaviour of a variational Bayesian filter (e.g., Kalman filter) of the type assumed by the predictive coding approach. This reformulation as a dynamical process theory enables a more intuitive expression of processing in terms of attractors, thus tying into the literature on dynamical systems and attractor basins as a description of mnemonic processes and decision-making. Indeed, the very existence of attractor states underwrites the free energy principle *per se* – and has some interesting connections with other normative approaches, such as value and reinforcement learning [10].

In brief, Sanborn and Chater's proposal provides deep connections to leading theories of neural organisation and their emergent dynamical and behavioural properties. Given its direct application to cognitive phenomena, their proposal thus provides a potential missing link in a 'trinity' of models of the brain and behaviour from the lowest levels of organisation (small-scale networks) all the way to its highest levels (cognition and behaviour).

Acknowledgments

I.B.S. acknowledges the support of an Australian Research Council Future Fellowship (FT160100437). The authors would like to thank an anonymous reviewer for helpful comments on a previous version of the manuscript.

[1]Cognitive Neuroscience Laboratory, School of Psychology, Social Work and Social Policy, University of South Australia, Adelaide, Australia

*Correspondence:
Ina.Bornkessel-Schlesewsky@unisa.edu.au (I. Bornkessel-Schlesewsky).

http://dx.doi.org/10.1016/j.tics.2017.04.003

References

1. Sanborn, A.N. and Chater, N. (2016) Bayesian brains without probabilities. *Trends Cogn. Sci.* 20, 883–893
2. Friston, K. (2005) A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836
3. Friston, K. (2009) The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* 13, 293–301
4. Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
5. Deco, G. *et al.* (2009) Stochastic dynamics as a principle of brain function. *Prog. Neurobiol.* 88, 1–16
6. Deco, G. *et al.* (2013) Brain mechanisms for perceptual and reward-related decision-making. *Prog. Neurobiol.* 103, 194–213
7. Friston, K. (2008) Hierarchical models in the brain. *PLoS Comput. Biol.* 4, e1000211
8. Friston, K. *et al.* (2012) Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3, 151
9. Garrido, M.I. *et al.* (2009) The mismatch negativity: a review of underlying mechanisms. *Clin. Neurophysiol.* 120, 453–463
10. Friston, K. and Ao, P. (2012) Free energy, value, and attractors. *Comput. Math. Methods Med.* 2012, 937860
11. Heekeren, H. *et al.* (2004) A general mechanism for perceptual decision-making in the human brain. *Nature* 431, 859–862
12. Basten, U. *et al.* (2010) How the brain integrates costs and benefits during decision making. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21767–21772
13. Alday, P.M. *et al.* (2014) Towards a computational model of actor-based language comprehension. *Neuroinformatics* 12, 143–179

Letter

The Sampling Brain

Adam N. Sanborn[1,*] and
Nick Chater[2]

Alday, Schlesewsky, and Bornkessel-Schlesewsky [1] provide a stimulating commentary on the issues discussed in our paper [2], highlighting important connections between sampling, Bayesian inference, neural networks, free energy, and basins of attraction. We trace here some relevant history of computational theories of the brain.

Consider the Hopfield network [3], a 'neural network' with symmetrical connections between binary neural 'units'. Hopfield showed how such a network could learn: patterns were 'imposed' on the network, and connections modified by local Hebbian learning. Remarkably, the network could 'fill in' patterns from fragments, providing a form of 'content-addressable memory'. Hopfield also showed that the 'free-running' of such a network minimized an 'energy function' across the entire network, measuring the coherence of the pattern with respect to the connection weights (roughly, coherence involves positive weights between units with the same value; negative weights between units with different values). The behavior of the network as it falls into a stable pattern can be viewed as falling into an attractor basin – exactly as the dynamics of many physical systems can be modeled as descending an energy landscape.

The Boltzmann machine [4], mentioned by Alday *et al.* [1], extends the Hopfield model in a variety of ways. Crucially, it can learn from patterns presented on subsets of 'visible' units, employing freely varying 'hidden' units which allow more-complex relationships between the visible units to be expressed. As before, the binary states of the 'neural' units in the Boltzmann machine can be assigned an energy function; but in the Boltzmann machine the units are stochastic. Thus, the network 'settles' not into a fixed pattern but rather into a probability distribution across patterns. Each 'update' of a new unit corresponds to drawing a new sample from the probability distribution using the technique of Gibbs sampling [5], first developed in computer vision, and now widely used in statistics and machine learning. Moreover, the Boltzmann machine can be trained to model a probability distribution presented over the visible units via Hebbian learning during a 'wake' phase, and anti-Hebbian learning during a 'sleep' phase, where no input is presented and the system runs freely.

This exciting constellation of ideas illustrates that a system of interconnected neuron-like units can learn to sample from a complex probability distribution from experience; and, indeed, sample from conditional distributions where some of the visible units are 'clamped' –