The sampling brain

Adam N. Sanborn
University of Warwick, Coventry, United Kingdom

Nick Chater
Warwick Business School, Coventry, United Kingdom

*Correspondence: a.n.sanborn@warwick.ac.uk (A.N. Sanborn)

Alday, Schlesewsky & Bornkessel-Schlesewsky (ASB) [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. Here, we trace 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 showed, too, 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---just as the dynamics of many physical systems can be modelled as descending in an energy landscape.

The Boltzmann machine [4], mentioned by ASB, 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 a 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"---

corresponding to Bayesian conditionalization. A learning rule carries out gradient ascent in the "likelihood" of the data presented at the visible units. All of this is achieved with no explicit representation of probability, but merely simple, distributed "neural" computations.

The Boltzmann machine does not scale-up well. But related ideas have evolved in a variety of directions. One approach focusses on representing complex probability distributions through sparse and structured "graphical models" which implicitly capture dependencies between variables (e.g., [6]). Indeed, general purpose programming languages for compositionally specifying and sampling from arbitrary probability distributions have been created (e.g., [7]).

A different development de-emphasizes compositional representation, and focusses on learning, typically with richly connected networks without a transparent interpretation. For example, "restricted" Boltzmann machines can be "stacked" into multiple layers (e.g., in deep belief networks; [8]). More broadly, deep learning has scaled up to achieve state-of-the-art machine learning performance [9].

More neurobiologically realistic implementations of sampling algorithms have recently been developed, some of which implement sampling for discrete variables on networks of spiking neurons (e.g., [10]). Other schemes for sampling continuous variables build on the link between energy and probability, producing dynamics in networks of excitatory and inhibitory neurons that implement an advanced sampling algorithm (e.g., [11]).

In contrast to our sampling proposal, Friston's (e.g., [12]) free energy approach does not treat the entire state of the brain as a single sample from a posterior probability distribution. The free energy approach also does not implicitly represent the probability of every possible hypothesis – far from it. The true posterior distribution is approximated by a simpler distribution, and minimizing free energy brings this simpler distribution into approximate correspondence with the true posterior. In Friston's model, neurons encode the parameters of this approximating distribution (cf. [13]), often a simple Gaussian distribution, which yields an elegant neurobiological implementation of the free energy approach.

We argued [2] that sampling will produce reasoning errors such as the unpacking effect and the conjunction fallacy if the sampler only samples a single mode in a multimodal distribution. Perhaps approximating a multimodal posterior distribution with a single (e.g., Gaussian) mode may be a different route to producing these same errors. Thus these various approximations to Bayesian inference may provide competing explanations of observed fallacies and biases observed in explicit reasoning with probabilities.

References

1. Alday, P. M. *et al.* (2017). Posterior Modes Are Attractor Basins. *Trends Cogn Sci*, 21, XXXX-XXXX.
2. Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, 20(12), 883-893.
3. Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554-2558.
4. Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1), 147-169.
5. Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721-741.
6. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.
7. Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4), 392–424.
8. Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313, 504-507.
9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
10. Pecevski, D., Buesing, L., & Maass, W. (2011). Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLoS Computational Biology*, 7(12), e1002294.
11. Aitchison, L., & Lengyel, M. (2016). The Hamiltonian Brain: Efficient Probabilistic Inference with Excitatory-Inhibitory Neural Circuit Dynamics. *PLoS Computational Biology*, 12(12), e1005186.
12. Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
13. Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11), 1432-1438.