Bayesian mechanics of learning in the brain

Chang Sub Kim (cskim@jnu.ac.kr)
Department of Physics, Chonnam National University
Gwangju 61186, Republic of Korea

Abstract

The brain is a biological system orchestrating its embodied agent's perception, learning, and behavior in the environment. Recently, we elaborated on the brain theory of higher–order functions in a physics–guided manner. Our study revealed that the brain's Bayesian inference facilitates local, recurrent, and unsupervised neural dynamics, completing the perception–action closed loop. In this work, we extend our effort to incorporate 'learning' into the formulation by accounting for learning as inference as well and derive the governing equations unified with perception and motor behavior. Subsequently, using a parsimonious generative model, we show how the brain integrates the Bayesian mechanics of learning subject to a time-dependent sensory stream. As a result, attractors are manifested to form in neural phase space and make dynamic transitions during the learning period.

Keywords: Perception, learning, and behavior; Variational Bayesian inference; Attractor dynamics; Prediction-error representation

Motivation and Aim

Biological agents recognize and respond to the external world by calling forth internal models in the brain. In addition, learning constitutes the crucial brain function of consolidating memory, e.g., via Hebbian plasticity (Hebb, 2002). The brain's perception, body control, and learning are conjointly organized to ensure the agents' homeostasis and adaptive fitness in the environment.

Recently, we have pursued a brain theory reckoning perception and motor operation to be active inference (Kim, 2018, 2021, 2023). We postulated that the physical brain is in a nonequilibrium stationary state, continually aroused by sensory stimuli, and the functional brain is capable of performing variational Bayesian inversion of the sensory data. Consequently, we accounted for the brain's functional behavior as described by attractor dynamics in continuous neural manifolds; we developed the governing equations of attractor dynamics in the brain, which we termed *Bayesian mechanics* (BM).

In our previous studies, however, we did not accommodate the learning part in our formulation. This work incorporates the brain's *learning dynamics* into the BM. We aim to provide a simple but insightful model for learning as an inferential computation. In doing so, our agenda is that the biological brain operates continually, not discretely, using continuous environmental representations; also, the brain's learning is a macroscopic phenomenon that can be better understood when guided by statistical-physical laws.

Bayesian Inversion in the Brain

We denote the brain's representation of the sensory perturbation by s and that of the external cause by μ; the former is encoded at the brain-environment interface and the latter inside the brain, both by neurophysical variables. In addition, the synaptic connection between two neural states μ and s is potentiated by the weight variable w; namely, learning is also mediated by neurophysical variables, not arbitrary free parameters.

From the brain-centric view, the environment is external or 'hidden'. Accordingly, we shall focus on the brain's internal model without paying attention to how environmental processes induce sensory perturbation at the interface. Thus, the sensory data s are assumed to be given and are to be perceived by the neural observer μ using a generative model. Our Bayesian framwwork prescribes synaptic efficacy as that μ transmits a presynaptic signal, to which s responds postsynaptically, relayed by the synaptic weight w. Overall, the brain's functional behavior is attributed to inferring the external causes of sensory data by 'inversion', namely, the posterior.

Generative Learning Model

It is conceded that the brain is encoded with an internal, generative model of sensory arousal, 'perceptual element' in Hebb's term (Hebb, 2002). Apart from the likelihood $p(s|\mu)$ encoding the brain's prediction of sensory-data generation, the full probabilistic model $p(s,\mu,w)$ must include the prior over the hidden states $p(\mu)$ and sensory evidence $p(s)$ (Buckley, Kim, McGregor, & Seth, 2017). Here we further assume that the brain's belief about synaptic efficacy is encoded by the weight variable w as another prior $p(w)$. In this work, we shall adopt the following product rule,

$$p(s,\mu,w) = p(s|\mu,w)p(\mu|w)p(w). \tag{1}$$

Next, we must provide the purposeful objective function that can be defined in terms of only biophysical brain variables (Friston, 2010). For a static or instant sensory input, the Laplace-encoded, informational free energy (IFE) is the objective to be minimized for variational Bayesian inversion (Buckley et al., 2017). However, for time-dependent sensory events, the variational objective must be generalized over a temporal horizon (Kim, 2018, 2021), which we will investigate further in the present work. Furthermore, we suggest that all the involved probabilities must be specified as biophysical ensemble densities in the brain matter in nonequilibrium stationary states (Kim, 2023).

Continuous-State Bayesian Dynamics

We assume that the synaptic rate representing 'plasticity', i.e., $\dot{w} = dw/dt$, is described by a Langevin-type equation for the weight variable w:

$$\dot{w} = h(w, \mu, s) + \zeta, \tag{2}$$

where h is the biophysical force generating weight change, and ζ is the associated noise. By further assuming that the noise ζ is Gaussian-distributed about zero mean with a certain variance, we can specify the prior $p(w)$ in Eq. (1). The detailed consideration of the temporal correlation of the noise is an important biophysical issue.

Next, by practicing the principle of least action (Landau & Lifshitz, 1976), we will deliver the deterministic equations of motion for the latent variables, which perform Bayesian inversion incorporating both perception and learning. To this end, we need to build a Lagrangian, for which we hypothesize the aforementioned Laplace-encoded IFE serve. For the focused learning dynamics, we shall adopt the synaptic weight w and its conjugate 'momentum' denoted by p_w as the latent variables. These variables span neural 'phase space' where dynamic attractors form and make transitions. Here, we present the anticipated learning mechanics from our formulation:

$$\dot{w} = \frac{1}{m_w} p_w + h(w, \mu, s), \tag{3}$$

$$\dot{p}_w = -p_w \frac{\partial h}{\partial w} - m_z (s - g) \frac{\partial g}{\partial w}, \tag{4}$$

where g is the biophysical force of sensory-data generation. A simple specification is $g = w\mu$, which underlies the sensory prediction by the neural observer as the linear map.

The workings of the weight dynamics [Eqs. (3)-(4)] is subject to the choice of the generative function h. We shall explore the following learning function:

$$h(s, \mu, w) = \alpha s \mu - \gamma w, \tag{5}$$

where the first term accounts for the standard Hebbian plasticity; the second term describes the Miller-MacKay model (Miller & MacKay, 1994), which prevents unlimited growth in synaptic weight. The Miller-MacKay model may be replaced with Oja's rule $-\eta s^2 w$ to explore nonlinear learning (Oja, 1982).

Numerical Results

Here, we only present the outcome from a minimal model to elucidate the essential features; however, the model is scalable to incorporate both the multi-synaptic channels and the cortico-cortical architecture in the brain.

In Figure 1, we illustrate attractor dynamics, which shows the dynamics of phase trajectories in the course of the brain's integrating the BM of closing a perception-action loop [we repeated the calculation in (Kim, 2023) with modification]. The full latent dynamics is described in the 6-dimensional manifold in the studied model; however, we depicted the attractor dynamics in the projected 3 dimension for illustrational purposes.

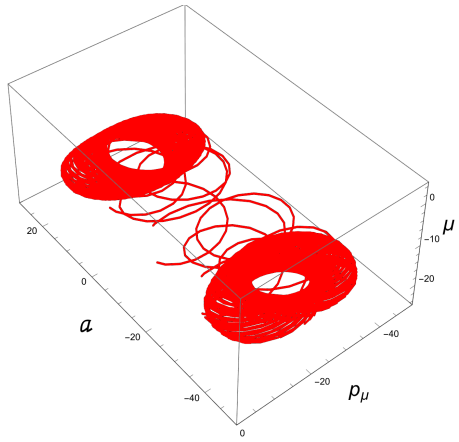

Figure 1: Dynamic transition of an attractor in neural phase space spanned by μ, a, and p_μ, which, respectively, represent the brain's expectation of sensory cause, motor variable, and prediction error in arbitrary units.

In the present work, we add the proposed learning dynamics [Eqs. (3)-(4)] to the BM; we are exploring the brain's perception and learning of 'nonstationary stimuli'. When a perceptual event elicits the sensory state s, the proposed BM activates attractor dynamics in the brain, which Hebb termed the *phase cycle* of a cell assembly (Hebb, 2002). Another sensory signal will launch a separable attractor with a distinctive phase cycle (Kim, in preparation). Furthermore, it would be of memory relevance to study temporal overlap and transience of independent attractors over multi-regions. In this respect, memories are generated via large-scale attractor dynamics.

Discussion

The momentum representation we unveiled [see the description above Eqs. (3)-(4)] matches with prediction error in predictive coding theory (Shipp, 2016). Recently, researchers found evidence of error neurons encoding prediction errors in the mouse auditory cortex (Audette & Schneider, 2023), which provides a neural base for our theory.

The neural circuitry proposed by our simple model accounts for the functional behavior of single cortical columns. We draw attention to an interesting report by others showing that every column in the neocortex behaves as an independent sensorimotor system, all performing the same intrinsic function (Hawkins, Ahmad, & Cui, 2017).

Finally, we recapitulate the premises of our theory: 1) the brain matter obeys physics laws and principles and affords the biological base for the emergent Bayesian mechanics; 2) the brain is a neural observer and capable of learning a model of the world, which reflects the subtle but inevitable top-down teleology in the current brain theory.

References

Audette, N. J., & Schneider, D. M. (2023). Stimulus-specific prediction error neurons in mouse auditory cortex. *Journal of Neuroscience*, *43*(43), 7119–7129. doi: 10.1523/JNEUROSCI.0512-23.2023

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55-79. doi: 10.1016/j.jmp.2017.09.004

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci*, *11*, 127–138. doi: 10.1038/nrn2787

Hawkins, J., Ahmad, S., & Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. *Frontiers in Neural Circuits*, *11*, 81. doi: 10.3389/fncir.2017.00081

Hebb, D. O. (2002). *The Organization of Behavior: A Neuropsychological Theory (1st ed.)*. Psychology Press. doi: 10.4324/9781410612403

Kim, C. S. (2018). Recognition dynamics in the brain under the free energy principle. *Neural Computation*, *30*(10), 2616–2659. doi: 10.1162/neco_a_01115

Kim, C. S. (2021). Bayesian mechanics of perceptual inference and motor control in the brain. *Biol Cybern*, *115*, 87–102. doi: 10.1007/s00422-021-00859-9

Kim, C. S. (2023). Free energy and inference in living systems. *Interface Focus*, *13*, 202200412. doi: 10.1098/rsfs.2022.0041

Landau, L. D., & Lifshitz, E. M. (1976). *Mechanics: Course of theoretical physics. Vol. 1 (3rd ed.)*. Elsevier. doi: 10.4324/9781410612403

Miller, K. D., & MacKay, D. J. C. (1994). The role of constraints in hebbian learning. *Neural Computation*, *6*(1), 100–126. doi: 10.1162/neco.1994.6.1.100

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267–273. doi: 10.1007/BF00275687

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, *7*, 1792. doi: 10.3389/fpsyg.2016.01792