Formalizing the Use of the Activation Function in Neural Inference

Dalton A. R. Sakthivadivel
VERSES Research Lab, Los Angeles, CA, 90016, USA

Department of Mathematics
Department of Physics and Astronomy
Department of Biomedical Engineering
Stony Brook University, Stony Brook, NY, 11794, USA
dalton.sakthivadivel@stonybrook.edu

We investigate how the activation function can be used to describe neural firing in an abstract way, and in turn, why it works well in artificial neural networks. We discuss how a spike in a biological neuron belongs to a particular universality class of phase transitions in statistical physics. We then show that the artificial neuron is, mathematically, a mean-field model of biological neural membrane dynamics, which arises from modeling spiking as a phase transition. This allows us to treat selective neural firing in an abstract way and formalize the role of the activation function in perceptron learning. The resultant statistical physical model allows us to recover the expressions for some known activation functions as various special cases. Along with deriving this model and specifying the analogous neural case, we analyze the phase transition to understand the physics of neural network learning. Together, it is shown that there is not only a biological meaning but a physical justification for the emergence and performance of typical activation functions; implications for neural learning and inference are also discussed.

Keywords: neural networks; Ising model; phase transitions; perceptrons

1. Introduction

The perceptron learning algorithm, developed by McCulloch and Pitts in 1943, is one of the earliest applications of biological principles for computation to mathematics or to machines [1]. A simple model, the perceptron consists of a single logic gate and is only capable of classification using linearly separable functions, like AND and OR. Nonetheless, recent algorithms have deviated only slightly from the original developments by McCulloch and Pitts; in many cases, these simply stack perceptrons or add features onto the original algorithm, such as in deep neural networks or convolutional neural networks.

https://doi.org/10.25088/ComplexSystems.31.4.433

Clearly, the contribution of the single-layer perceptron remains relevant today.

Somewhat anomalous in the perceptron, and indeed in further models, is the critical importance of the activation function. McCulloch and Pitts recognized that neural firing occurs in an all-or-none fashion, and that any function with a rapid transition between two end behaviors would suffice to describe this phenomenon [2]. The same argument was later presented at the level of neural populations in [3], by showing that for a realistic distribution of single neuron activation, the ensemble activity necessarily looks like a sigmoid function. In other words, a specific class of functions is generally used for an activation function, which can be described as discontinuous or nearly discontinuous at a "switching point," vertically asymmetric about this point and bounded from below. Concrete examples include the Heaviside function originally used by McCulloch and Pitts, and Wilson and Cowan's sigmoid function. Interestingly, a class of activation functions that are bounded from below but exhibit asymptotically linear behavior for inputs greater than a critical threshold, such as ReLU, ELU, Mish and Swish, has been experimentally evaluated as providing the best performance for a large number of network architectures and tests [4–6]. Much like the Heaviside activation function, however, these functions are justified by their performance and are given heuristically.

While the use of an activation function is and has been justified by the biological facts, and its success is obvious, it is still assembled primarily phenomenologically. The activation function was certainly integral to the application of neural networks as logical devices that classify nonbinary variables—but the precise mechanism that justifies these functions' role in inference and the physiological relevance of this function both remain unclear. Most proofs of the previous statement also yield little insight into the relevance of the activation function and especially of the specific shape elaborated on earlier. These proofs often rely on what could be summarized as the power of nonlinearity, which allows the approximation of nonlinear or nonpolynomial functions. Consider that data-generating processes are governed by a dynamical system, which could be a high-dimensional stochastic system or partial differential equation, the solutions to which are typically nonlinear or nonpolynomial in character. Then the necessity of such a function becomes clear. In greater detail, a theorem offered in [7] states that the set of possible neural network configurations $\mathcal{N}$ is *dense* in the space of continuous real-valued functions, or that any real-valued function is contained in or is a limit point of $\mathcal{N}$, if and only if the activation function on $\mathcal{N}$ is nonpolynomial. In other words, given arbitrary width and depth, the property of being a "universal approximator" is precisely that of having a nonpolynomial activation function. Still, this proof leads to little insight about the biological plausibility of or physical motivation for the specific functions used.

To understand how activation functions arise in artificial neural networks and how they are connected to the fundamentals of biologically inspired computation, we employ a model from statistical mechanics called the Ising model. The Ising model was devised by Wilhelm Lenz and Ernst Ising in the 1920s to describe magnetism in metals and the loss of magnetization when magnets are heated [8]. The Ising model is a model of the atomic structure of a metal, where the nuclei of metallic atoms are defined with a property called "spin," pointed either up or down. When all spins are positively aligned or all lattice sites take values of $+1$, the system is magnetized. Like many systems in nature, it exhibits a phase transition at a critical temperature, in which the magnetization of a cooled metal is lost above a critical temperature or regained when cooled. In fact, many other phase transitions can be proven to take characteristics of this one—phase transitions lie in one of a few *universality classes*, meaning they show the same characteristics no matter what the underlying dynamics are. The Ising universality class contains such different phenomena as the liquid-gas critical point and the behavior of strings in string theory [9, 10]. Similar to this class of behavior, a neural spike is a sort of transition where, once past a critical point (a threshold potential), a spike is initiated, and the system goes from disordered uptake (random diffusion of ions across the membrane) to ordered uptake (uptake of Na^+ and other positive ions to initiate depolarization). We will use this correspondence to model the spiking behavior of the neural cell as an Ising model with an appropriate phase transition. In so doing, we recover expressions for both the hyperbolic tangent and unbounded-above linear activation function. This shows that an artificial neural network must be equipped with such an activation function if it is to be a meaningful approximation of a biological neural network, offering a new take on the manner in which activation functions lead to neural computation.

2. Main Results

2.1 Modeling Neural Dynamics

Neural spikes are both a regular phenomenon and a highly complex, nonequilibrium process. As an example of self-organization, neural firing emerges from complex but quantifiable dynamics, here involving ionic equilibria and membrane selectivity [11]. The neuron is surrounded by ions in its extracellular fluid, meaning it is subject to diffusion of these ions across its cell membrane through ion channels. It maintains a negative resting potential of around -70 mV, which requires active transport of positive Na^+ ions out of the cell. This results in a persistent concentration gradient, which is precisely what allows a spike to occur. When a critical voltage is reached, previously closed voltage-gated ion channels open. The positive Na^+ ions flow along this concentration gradient through these now-open channels, leading to an upward spike in voltage.

We have simplified this model of neural dynamics to be a stationary process coupled to a bath. We can then approximate the system as being in a local equilibrium, meaning we can use a simple equilibrium Ising model to describe the system. Consider a particular formulation of the Ising model coupled to a thermal bath, which undergoes a rapid quench and magnetizes in response to this sudden cooling. When the quench is removed, the Ising model heats up again. For periodic quenching, the dynamics themselves will be periodic, but will obey the typical $h = 0$ transition in magnetization. The phases induced by the quench can thus be made distinct, for example, alternating disordered phases with thermal fluctuations and ordered phases with positive magnetization. This provides a model of neural dynamics in which the crucial simplification in the model is ignoring the source of the external quench, thereby restricting us to only local (intracellular) interactions within the system. There is no consequence to the validity of our model, since firing is the organization of channel dynamics within the cell.

2.2 The Ising Case for Neural Firing

As stated, both systems are capable of exhibiting two differently stereotyped dynamics, or "phases." In the Ising model, one is a high-temperature *paramagnetic* phase, where the spins in the model are disorganized and unaligned, weakly correlated with one another and subject to random fluctuations. The magnetization m, or average spin, is zero in this case; this is a result of the random configuration of spins, such that approximately half of the spins should be occupying states -1 and half occupying $+1$, for $m = \langle s \rangle = 0$.

We also observe a *ferromagnetic* phase in which the spins are organized and aligned in one direction. Here, m is either -1 or $+1$, which correspond to anti-ferromagnetism and ferromagnetism, respectively. A quench is a decrease in thermal energy, which causes the spins to align with each other so that the model occupies a low energy state and $m = \langle s \rangle = 1$. We will see that the energy in the Ising model depends on the interactions between spins. As such, to get the total energy, we take the negative sum of spin states over all pairs of neighbors. Clearly, when pairs of neighboring spins are aligned in the positive direction, such that the sum of n spin pairs is $-n$, the energy is at a minimum. The converse is also true: when the energy in the system decreases, spins will align and take a lower energy configuration to satisfy this. The final state, -1 or $+1$, is "chosen" as the magnet cools toward the critical temperature, according to the boundary condition of the model.

We consider the neural membrane a two-dimensional lattice of channels wrapped around a cell body. The transition from order to disorder and subsequent ensemble action in spiking are now that of the Ising model. It is well known that a threshold potential exists, in which a neuron exhibits sudden spiking in response to a critical level of stimulation. In this case, the state of the neuron—the voltage—is the negative of Ising temperature, and as voltage increases toward the threshold, temperature decreases toward a critical temperature T_c. In this case, the adaptation of the standard Ising model applies well. While it is coupled to a thermal bath that increases its temperature, it remains disordered due to high energy fluctuations. Similarly, at rest, the neuron occupies a highly entropic and thus high energy state, given the open and closed channels in the membrane; it is also subject to fluctuations that destroy order in the system as a result of constant diffusion and pump dynamics. Together, the two maintain a disordered resting state, and along with the highly entropic channel states, a spike is impossible without cooling.

We will treat events discretely, as a single spike in response to an input. We will also consider the neuron as being in disequilibrium with its surroundings, with occasional field interactions. As stated, for modeling purposes, we simplify this as a local input effect. In the neural case, cooling comes in the form of the summation of inputs from other neurons in the network. This plays the role of a quench, in that it moves the system's state toward order.

2.3 The Transition to Magnetization

To connect macroscopic observables to microscopic state variables, we often rely on formalisms from statistical mechanics. One such technique used in study of phase transitions is a particular type of coarse graining called mean-field theory (MFT), which formulates a model of the macroscopic-level change that results from certain microscopic changes. MFT is quantitatively incorrect in two dimensions and is only an approximation; nonetheless, for both computational and pedagogical reasons, we will demonstrate this using the mean-field approach.

Here, we will briefly state the derivation for the phase transition in the Ising model. A spin lattice in zero field is described by its Hamiltonian $\hat{H}$ in the following way:

$$\hat{H} = -J \sum_{i,j} s_i s_j,$$

with $s_n \in \{-1, 1\}$. $\hat{H}$ gives the total energy of the system, which in turn gives its dynamics, as the sum over all neighboring spins. In our analogous neural model, a spin is a channel state, which at any time t either contains an Na$^+$ ion or does not.

MFT assumes that at large length scales, a system converges to its average dynamics, with only large fluctuations playing a role in the dynamics of the system. We use this to coarse-grain the model by removing second-order fluctuations, which are assumed to be vanishingly small. The local interaction term in $\hat{H}$ is

$$s_i s_j = (\langle s_i \rangle + \sigma s_i)(\langle s_j \rangle + \sigma s_j),$$

which is a product of two variables under the influence of a random displacement. We assert that in this system, the spins tend toward a similar mean and fluctuations necessarily decrease; therefore, in describing the dynamics leading to an ordered transition, we may assume the fluctuations become small. This means that we can rewrite $s_i s_j$ as the following:

$$\langle s_i \rangle \langle s_j \rangle + \langle s_j \rangle \sigma s_i + \langle s_i \rangle \sigma s_j + \sigma s_i \sigma s_j.$$

Since we have assumed the fluctuations are small, the final term will vanish. If we use this and an expansion of the random displacement into a fluctuation about a mean, this becomes

$$s_i s_j \approx \langle s_i \rangle \langle s_j \rangle + \langle s_j \rangle (s_i - \langle s_i \rangle) + \langle s_i \rangle (s_j - \langle s_j \rangle).$$

We also use the fact that as the phase transitions, the average spin value $\langle s_n \rangle$ will approach a magnetization value m, corresponding to the organization of spins needed to produce magnetization. Then, we can rewrite $\hat{H}$ as

$$-J \sum_{i,j} m(s_i + s_j) - m^2,$$

replacing spins with the mean field m. If we take spin states as being highly correlated, then the i's and j's become equal; in that case, the sum over neighbors will reduce to the number of connections, half the number of neighbors z, across all sites in the lattice. This gives a scaling factor of $z/2$:

$$-\frac{zJ}{2} \sum_{i=1}^{N} 2m s_i - m^2.$$

We can further simplify to

$$\hat{H}_{\mathrm{MF}} = -zJm\sum_{i=1}^{N} s_i + \frac{NzJm^2}{2}$$

by distributing the scaling factor into the sum.

This mean-field Hamiltonian describes what the total energy looks like at a phase transition, at a coarse scale. It does not, however, describe the transition itself. To achieve this, we will need a partition function, which uses a Hamiltonian to describe the statistical properties of a system, giving us crucial information about the system's dynamics. Using $\hat{H}_{\mathrm{MF}}$ with the canonical partition function yields

$$Z = e^{-\beta \hat{H}_{\mathrm{MF}}},$$

where β is a particular thermodynamical quantity, $(k_B T)^{-1}$. Expanding the site-wise partition function and using some trigonometric identities, we have

$$Z = e^{-\beta \frac{NzJm^2}{2}} 2\cosh(\beta zJm)^N.$$

Finally, to find our magnetization m, we must minimize the free energy of the system with respect to m. Using $F = -(\beta N)^{-1} \ln\{Z\}$,

$$F = \frac{1}{2} zJm^2 - \frac{1}{\beta} \ln\{2\cosh(\beta zJm)\}.$$

The m that minimizes this free energy, by solving $\partial F / \partial m = 0$, is

$$m = \tanh\left(\frac{zJm}{k_B T}\right),$$

where we have used the previous definition of the thermodynamic beta. If we define the critical temperature as $T_c = \mathrm{zJ} / k_B$, then this simplifies to

$$m = \tanh\left(\frac{T_c m}{T}\right), \tag{1}$$

the plot of which is contained in Figure 1.

Immediately we observe a hyperbolic tangent function arise in this mean-field model. The curve bifurcates at the critical value of T, showing the two possible magnetized states. We clearly see either $m = -1$ or $m = 1$, given by the ends of equation (1). We disregard the zero solution at $T < T_c$ as energetically unfavorable.

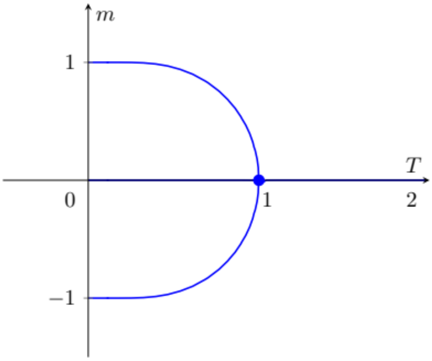

Figure 1. Mean-field magnetization m as a function of temperature. A phase transition is evident in the graph of equation (1) at $T = T_c$ (here set such that $T_c = 1$), where magnetization becomes nonzero.

Note that this tanh curve is a different sort of activation function; rather than determining magnetization, it determines either of the possible magnetized states. In principle, we could maintain this bifurcation: suppose we defined two different firing patterns, where, upon receiving an input and crossing a critical point, all channels either contained an Na^+ ion or all channels did not. This could represent the firing of an inhibitory neuron causing selective inactivity in a firing excitatory neuron, which would normally communicate a signal. Then, we would have something corresponding to a critical input creating a single spike according to the statistics of the input. In this case, the activation function determines whether a stimulus is likely to elicit excitatory or inhibitory neural spikes, perhaps comprising a different sort of classification.

However, much like this is not the main feature of the Ising model phase transition, this is not the major point of this paper. Instead, we restrict the magnetization to $m = 1$ and examine the resultant analogy to firing dynamics. This will also allow us to determine the salient features of the previously mentioned class of activation functions.

2.4 Recovering the Activation Function from the Ising Model

It is clear to see from Figure 1 that at temperatures above T_c, the only solution is $m = 0$. Below T_c the solution is $m = -1$ or 1, given by the two ends of equation (1). When restricted to $m = 1$, this curve behaves like a different hyperbolic tangent function, going from zero to one. So, for some parameter a, our function looks like

$$m(T) = -\frac{1}{2}\tanh(a(T - T_c)) + \frac{1}{2}, \tag{2}$$

which reverses when we set the temperature to neural state, as suggested earlier: recall the neural membrane voltage is itself negative, and so $T = -V$. As seen in Figure 2, this reproduces the neural activation function, where the threshold T_c is a bias and the switching behavior represents spiking or not spiking.

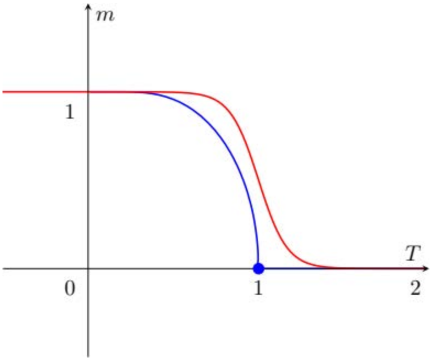

Figure 2. Magnetization fitted to a hyperbolic tangent. The previously defined magnetization curve can be fitted to another hyperbolic tangent curve, satisfying the typical firing or not-firing activation function with V increasing like $-T$. The hyperparameter a is set to $a = 6$, and $T_c = 1$. A second hyperparameter multiplying the critical temperature can be used to bring the fit even closer.

Since T, the temperature, depends on the quench, we can combine the temperature of the bath T with the quench, $T(t) = T - \delta(t - t_c)Q$. Here, t_c is a time when cooling is applied, and the delta function $\delta(x)$ returns one for an argument of zero, and zero everywhere else. Thus, cooling only acts on the temperature when $t = t_c$. Note that while an interaction with a thermal field could have been included in the Hamiltonian, we opt to couple it to the temperature later in the paper, for a variety of reasons—in particular to follow through on our approximation of anomalously introduced local effects, along with our desire to begin with a time-independent, and thus equilibrium, Hamiltonian.

We may now parameterize motion along this curve due to changes in temperature in time. Recall we have negated temperature, as it is equal to neural membrane voltage, $-V$. We have already coupled our quench to temperature as a subtractive element that restores it to order. Suppose the model heats up linearly. This is accurate with respect to the neuron, which uses pumps to eject positive ions at a constant rate. Indeed, we have previously defined a quench as a perturbation from equilibrium. In reality, it is the influx of positive ions due to some firing event adjacent to the neural cell. The model will heat up again as soon as it loses heat or pumps ions out of its cell body. We assume the neural pump acts with a constant speed, so that the time spent in the $m = 1$ regime can be parameterized as linear in time t. Suppose also that this rate is one degree per second, such that $T / t = 1$ and $T(t) = t - Q$ for $t_c < t \leq Q$. Then, the number of spikes emitted or the time spent in $m = 1$ is the integral of equation (2) from zero to Q. This is because, under our assumptions, the time to cool is equal to one kelvin per unit time. In that case, the time to cool back to zero perturbation when Q is subtracted from T is exactly Q.

Given the analogy drawn in Sections 2.1 and 2.2 holds, clearly the mean-field description of neural firing leads to a function that describes neural firing based on perturbations to equilibrium, as described earlier. Then, the number of spikes emitted in this time is the time spent in $m = 1$, or the time before cooling, which is the previously described integral of equation (2) with respect to temperature. This integral is evaluated as follows:

$$\frac{1}{2}\int_0^Q [\tanh(T) + 1]\mathrm{d}T = Q + \frac{1}{2}\ln\{2\} + \frac{1}{2}\ln\left\{1 + e^{-2Q}\right\} + C.$$

Clearly, we have the linear term dominating for $Q \gg 0$; in which case, for $C = 0$, we recover our ReLU function. This behavior also reproduces that of the exponential linear unit for $C = -\alpha$, differing by no more than α anywhere. In general, we have an expression for linear or approximately linear, unbounded-above activation functions.

If we choose to fit a sigmoid function to our magnetization, rather than the hyperbolic tangent, then we have a similar result:

$$\int_0^Q \frac{1}{1 + e^{-T}}\mathrm{d}T = Q + \ln\left\{1 + e^{-Q}\right\} + C.$$

The relevance of this with respect to neural firing is that the saturating activation function is only a binary classification case with one spike—a single logic gate. More complex learning, such as the encoding of complex stimuli, on the other hand, requires many spikes. Hard quenches, or strong inputs, mean more time spent in the $m = 1$ regime; thus, stronger inputs mean more spikes get emitted. We then recover ReLU and ELU as functions for firing rate, by counting spikes over time. Since time spent magnetized, or time before heating, corresponds directly to quench strength, so too does spike count.

3. Discussion

To summarize our argument, we have shown that, for underlying physical dynamics that are complicated enough to perform inference at some (coarse-grained) level, the sigmoidal nonlinearities of an activation function arise by simple scaling arguments from mean-field theory. Conversely, scaling arguments reveal that sigmoidal activation functions are necessary for a model of neural dynamics that is capable of inference. This suggests a principled reason for the ubiquity of these two classes of activation functions: artificial neural networks are physically realistic coarse-grainings of biological neural networks if and only if their firing dynamics look like sigmoid functions (or, when extended through time, rectifiers).

Given their comparatively small number of units and reduced computational power, we do indeed expect artificial neural networks to be mean fields of biological neural networks. In fact, it has been shown that one biological neuron can be modeled using multiple deep or artificial neural networks [12], suggesting that many interacting mean-field units are needed to recover the microscopic activity of a real neuron. We recall, for instance, [13], where a thermodynamic limit is postulated when making sense of the firing statistics of large neural networks (suggesting finite size effects in the absence of a large number of neurons); also [14], where it is shown that pairwise correlations contain important information in biological neurons. That the design of traditional artificial neural networks follows from mean-field analysis, and the consistency of this observation with previous results in the literature, confirm the general idea that biological neural networks are orders of magnitude more complex than their artificial counterparts. It remains to be seen how the analysis here lends itself to numerical estimates of the complexity of a biological neural network.

We now dissect some consequences of this argument.

3.1 Firing Rates and Sparse Neural Codes

Evidence suggests that neurons rely on sparse coding to efficiently communicate stimuli, especially in high-noise or high-dimensional environments. In fact, many separate neural coding schemes have been considered to emerge from sparsity, which neural networks employ due to energy constraints and to cope with dimensionality [15]. Broadly, sparse coding states that different firing rates, which contain representations of information by encoding features of a stimulus, will be sparsely distributed in a neural network. In large neural populations, key neurons will be firing at various rates and most other neurons will not be firing at all. Such a sparse code is advantageous for efficient learning by decorrelating inputs, which allows features to be coded independently. Crucially, this leads to a robust representation and is equivalent to reducing the coding of redundant features while preserving coded information [16].

The rectifier, or the unbounded-above activation function we discuss, has indeed been shown to improve representation in deep neural networks by precisely these mechanisms [17]. Some neurons are firing with a particular rate, lying on the linear portion of the curve, and others are resting, lying on the portion of the curve valued at zero. The coding benefits highlighted are exactly those found in sparse representations in biological neural networks, where disentangling is referred to as decorrelating inputs, which assists in learning high-dimensional data. These networks also utilize sparsity as a rich but energy-efficient coding scheme, showing that in deep neural networks, sparse representations take fewer computational resources while showing high training accuracy.

In our model, this sparsity is reproduced by local effects such as quenches of different magnitudes acting on particular neurons in the network. It can be shown that, for Gaussian distributed quenches, our results imply the results in [3]: calculating the transfer function of a network of neurons under quenches fluctuating about a mean of zero follows the argument in [3] exactly.

3.2 Energy-Based Learning

Recently, a broad theory of machine learning and inference has been formulated in an energy-based framework—in particular, a paradigm based on energy minimization has been proposed in [18], where choosing a network configuration that minimizes energy is equivalent to finding an output that minimizes loss. These follow on older ideas where free energy minimization is employed in statistical learning, such as the Boltzmann machine or spin glass models. Here, the energy of a configuration is used as a penalty, following the idea that physical systems seek to minimize free energy and that this underlies the stability of a given state. This appeals to statistical mechanical ideas about energy minimization, which we have already used in discussing the Ising model—the configuration chosen by a system always obeys a minimization principle. As such, this can be used as a measurement of error, where we designate high energy states as being incorrect in both the physical and statistical sense.

A useful way of thinking about the idea of free energy minimization is that free energy is defined as

$$F = E - TS.$$

For clamped energy levels, clearly, maximizing entropy is equivalent to minimizing free energy, since

$$\Delta F = -T\Delta S$$

for constant energy and temperature. Then, free energy minimization is a natural consequence of the second law of thermodynamics, which states that systems will always produce greater entropy. In the information-theoretic sense, defined in [19] as essentially equivalent to the thermodynamical sense, maximizing entropy is choosing the best model of observed variables. Thus, we have a direct application to our inference or learning process.

Following this, we examine why a neural Ising model spikes. Clearly, when the temperature decreases, the entropic contribution to free energy decreases as well. Hence, minimization of free energy occurs when total energy is minimized. We observed this happen when the Hamiltonian was in a magnetized state. In the sense of an error signal, when an input—a temperature-lowering quench—arrives, the error in the system is high as long as the Ising model occupies a high energy state, which is unlikely given the physical and statistical scenario. By magnetizing, or spiking, the system decreases this error through responding to the input, which is equivalent to choosing a free-energy-minimizing stable state. In the energy-based learning scheme, loss functions are often arrived at by explicitly considering the marginalized Gibbs distribution over the inputs to the system, and learning is performed by minimizing the resultant free energy in the zero-temperature limit.

This accords with other, more biological ideas concerning energy minimization in learning, wherein neural spikes learn the relationship between stimulus and evoked response, and minimization of energy underlies learning. It has been found that real neural networks, in vitro, minimize *variational* free energy when learning representations of stimuli [20]. Variational free energy is an information-theoretic notion closely related to the thermodynamical Helmholtz free energy, although whether only by statistical mechanical analogy or also by physical principles remains controversial [21, 22].

3.3 Self-Similarity and Criticality

We note one final implication by suggesting a relationship between this result and MFT applied to neural populations. In particular, we note that to recover nonlinear firing statistics, the collective dynamics of neural populations are almost ubiquitously described using a sigmoid function [23]. The importance of the sigmoid function in statistical approximations of neural population dynamics—especially mean-field models—was first suggested in [3], wherein it was shown that for a realistic model of population firing, the proportion of firing cells naturally followed a sigmoid. In following work by Amari, nonlinear functions were also necessary to model the collective dynamics of a neural field as a self-organized pattern [24]. More recently, transfer functions in biologically realistic mean-field models have taken the form of a rectifier [25], corresponding to our own unification of unbounded-above functions with sigmoid-type functions as a firing rate. The results in this paper undoubtedly extrapolate to the case of neurons as a subunit and neural populations as a mean field, a relationship consistent with the approximate self-similarity observed in the human cortex. We note that self-similar systems, including the Ising model around T_c, are generally in a state of criticality; this is the so-called "edge of chaos" close to a phase transition. Signatures of criticality have been observed in the brain [26, 27], and critical dynamics are known to be important for computation in both biological and artificial neural networks, which have been shown to perform best at criticality [6, 13, 28, 29]. This congruence adds a dimension to these results, as they capture the dependence of ideal computation on the scale invariance of certain expressions.

4. Conclusion

We have shown that, as a model of the key features of real neural dynamics, an artificial neural network is a mean-field model of biological neural networks. This model falls in the Ising universality class, and thus an artificial neural network naturally exhibits a sigmoidal or tanh-like switching behavior between firing and not firing. Various conventional activation functions can be easily arrived at from this mean-field model; as such, we have motivated the designs of historical and modern artificial neural networks, and in particular, the concept and typical form of the activation function. In so doing, we have also examined how ideal learning necessarily invokes the nonlinear processes in the neuron and utilizes energy minimization, by modeling this process with an Ising model and applying other statistical mechanical ideas.

References

[1] T. H. Abraham, "(Physio)logical Circuits: The Intellectual Origins of the McCulloch–Pitts Neural Network," *Journal of the History of the Behavioral Sciences*, **38**(1), 2002 pp. 3–25. doi:10.1002/jhbs.1094.

[2] W. S. McCulloch and W. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *The Bulletin of Mathematical Biophysics*, **5**(4), 1943 pp. 115–133. doi:10.1007/BF02478259.

[3] H. R. Wilson and J. D. Cowan, "Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons," *Biophysical Journal*, **12**(1), 1972 pp. 1–24. doi:10.1016/S0006-3495(72)86068-5.

[4] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS’12)*, Vol. 1, Lake Tahoe, NV (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.), Red Hook, NY: Curran Associates, Inc., 2012 pp. 1097–1105. dl.acm.org/doi/10.5555/2999134.2999257.

[5] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” in *Proceedings of the 27th International Conference on Machine Learning (ICML’10)*, Haifa, Israel, (J. Fürnkranz and T. Joachims, eds.), Madison, WI, Omnipress: 2010 pp. 807–814. dl.acm.org/doi/10.5555/3104322.3104425.

[6] S. Hayou, A. Doucet and J. Rousseau, “On the Impact of the Activation Function on Deep Neural Networks Training,” in *Proceedings of the 36th International Conference on Machine Learning (PMLR 97)*, Long Beach, CA (K. Chaudhuri and R. Salakhutdinov, eds.), 2019 pp. 2672–2680. proceedings.mlr.press/v97/hayou19a/hayou19a.pdf.

[7] M. Leshno, V. Y. Lin, A. Pinkus and S. Schocken, “Multilayer Feedforward Networks with a Nonpolynomial Activation Function Can Approximate Any Function,” *Neural Networks*, **6**(6), 1993 pp. 861–867. doi:10.1016/S0893-6080(05)80131-5.

[8] S. G. Brush, “History of the Lenz–Ising Model,” *Reviews of Modern Physics*, **39**(4), 1967 pp. 883–893. doi:10.1103/RevModPhys.39.883.

[9] A. D. Bruce and N. B. Wilding, “Scaling Fields and Universality of the Liquid-Gas Critical Point,” *Physical Review Letters*, **68**(2), 1992 pp. 193–196. doi:10.1103/PhysRevLett.68.193.

[10] A. Sedrakyan, “3D Ising Model as a String Theory in Three-Dimensional Euclidean Space,” *Physics Letters B*, **304**(3), 1993 pp. 256–262. doi:10.1016/0370-2693(93)90291-O.

[11] A. L. Hodgkin and A. F. Huxley, “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve,” *The Journal of Physiology*, **117**(4), 1952 pp. 500–544. doi:10.1113/jphysiol.1952.sp004764.

[12] D. Beniaguev, I. Segev and M. London, “Single Cortical Neurons as Deep Artificial Neural Networks,” *Neuron*, **109**(17), 2021 pp. 2727–2739. doi:10.1016/j.neuron.2021.07.002.

[13] G. Tkacik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry II and W. Bialek, “Thermodynamics and Signatures of Criticality in a Network of Neurons,” *Proceedings of the National Academy of Sciences*, **112**(37), 2015 pp. 11508–11513. doi:10.1073/pnas.1514188112.

[14] E. Schneidman, M. J. Berry II, R. Segev and W. Bialek, “Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population,” *Nature*, **440**(7087), 2006 pp. 1007–1012.
doi:10.1038/nature04701.

[15] M. Beyeler, E. L. Rounds, K. D. Carlson, N. Dutt and J. L. Krichmar, “Neural Correlates of Sparse Coding and Dimensionality Reduction,” *PLOS Computational Biology*, **15**(6), 2019 e1006908. doi:10.1371/journal.pcbi.1006908.

[16] P. Földiák, “Forming Sparse Representations by Local Anti-Hebbian Learning,” *Biological Cybernetics*, **64**(2), 1990 pp. 165–170. doi:10.1007/BF02331346.

[17] X. Glorot, A. Bordes and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL (G. Gordon, D. Dunson and M. Dudík, eds.), 2011 pp. 315–323. proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

[18] Y. LeCun, S. Chopra, R. Hadsell, M. A. Ranzato and F. J. Huang, “A Tutorial on Energy-Based Learning,” *Predicting Structured Data* (G. Bakir, T. Hofman, B. Schölkopf, A. Smola, B. Taskar and S. V. N. Vishwanathan, eds.), Cambridge, MA: MIT Press, 2007.

[19] E. T. Jaynes, “Information Theory and Statistical Mechanics,” *Physical Review*, **106**(4), 1957 pp. 620–630. doi:10.1103/PhysRev.106.620.

[20] T. Isomura and K. Friston, “In Vitro Neural Networks Minimise Variational Free Energy,” *Nature Scientific Reports*, **8**(1), 2018 16926. doi:10.1038/s41598-018-35221-w.

[21] A. B. Kiefer, “Psychophysical Identity and Free Energy,” *Journal of the Royal Society Interface*, **17**(169), 2020 20200370. doi:10.1098/rsif.2020.0370.

[22] M. Andrews, “The Math Is Not the Territory: Navigating the Free Energy Principle,” *Biology & Philosophy*, **36**(3), 2021 30. doi:10.1007/s10539-021-09807-0.

[23] G. Deco, V. K. Jirsa, P. A. Robinson, M. Breakspear and K. Friston, “The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields,” *PLOS Computational Biology*, **4**(8), 2008 e1000092. doi10 .1371/journal.pcbi.1000092.

[24] S. Amari, “Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields,” *Biological Cybernetics*, **27**(2), 1977 pp. 77–87. doi:10.1007/BF00337259.

[25] M. di Volo, A. Romagnoni, C. Capone and A. Destexhe, “Biologically Realistic Mean-Field Models of Conductance-Based Networks of Spiking Neurons with Adaptation,” *Neural Computation*, **31**(4), 2019 pp. 653–680. doi:10.1162/neco_a_01173.

[26] Z. Ma, G. G. Turrigiano, R. Wessel and K. B. Hengen, “Cortical Circuit Dynamics Are Homeostatically Tuned to Criticality *In Vivo*,” *Neuron*, **104**(4), 2019 pp. 655–664. doi:10.1016/j.neuron.2019.08.031.

[27] L. Cocchi, L. L. Gollo, A. Zalesky and M. Breakspear, "Criticality in the Brain: A Synthesis of Neurobiology, Models and Cognition," *Progress in Neurobiology*, **158**, 2017 pp. 132–152. doi:10.1016/j.pneurobio.2017.07.002.

[28] G. Yang and S. Schoenholz, "Mean Field Residual Networks: On the Edge of Chaos," *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.), Curran Associates, Inc., 2017 pp. 7103–7114. proceedings.neurips.cc/paper/2017/file/81c650caac28cdefce4de5ddc18befa0-Paper.pdf.

[29] S. Landmann, L. Baumgarten and S. Bornholdt, "Self-Organized Criticality in Neural Networks from Activity-Based Rewiring," *Physical Review E*, **103**(3), 2021 032304. doi:10.1103/PhysRevE.103.032304.