Pathway for Energy Based Models

Germán González, gonzalezv.germanh@javeriana.edu.co

***Abstract*—**

***Index Terms*—Gabor filters, Modern Hopfield networks, Convolutional Networks, Associative memory.**

I. Introduction

One of the main field of research on AI is becoming the analysuis of Energy Based Models. The article "Attention is All You need" has had a tremendous impact in the advance of artificial intelligence. And the recent "Hopfield Networks is all you need" has shown that the Energy based models is the way to go in terms of modeling. However, the theory that backs them is not clearly understood, eventhough it promises that there is a resemblance between biology and the models that path is difficult to follow up. In order to cover that gap, this article will approach the models from two main stating points: The linear associative models that back the idea of attention mchanisms and energy models starting from the Ising model. In general article will be structured in the following mode:

1) **Previous work**
2) **Associative memories**
3) **Differential equations, eigenvalues and eigenvectors**
4) **Hamiltonian dynamics**
5) **Clasical Hopfield network**
6) **Modern Hopfield networks**

II. Previous work

Energy Based Models can be traced back to the clasical Hopfield networks [8]. The article "Hopfield Networks is all you need" treats an extension of the classical Hopfield networks allowing the storage of exponentially more records that the ones possible in the classical networks. This has led then to focus on the investigation about energy Based Models as depicted in the recent article by Yann LeCun "A path towards autonomous machile learning" [5].

III. Associative memories

It is know that the human memory is associative. That is to say, that the human needs to have a cue of information in order to correalte the rest of a recall. The memory is built up day after day from a colection of information that somehow gets store in the brain (Hebbian Learning).

The intuition of association can be explained as in the following figure. A person might have two recalls that are totally uncorrelated. However, the presence of a third recall might bring a link between the two so that getting to remeber one induces the recall of another one. When the process only leads to getting to get the full recovery of the recall, then we just have a completition process while, in the case of linking to another recall might lead to and induction process. However, in this article, the process of induction will not be analyzed.

To understand the process of associative learning it is necessary to go back to the theory of eigenvalues and eigenvectors. The reason why the theory was develped was help resolving systems of diferential equations of first order.

$$\dot{x} = \mathbf{A}x \tag{1}$$

The objective is to transform the system in a system where the **A** matrix is replaced with a diagonal **D** matris that is easy to solved.

$$\dot{z} = \mathbf{D}z \tag{2}$$

The eingenvector and eigenvalues comes from a change of coordinates as follows:

$$x = Tz \tag{3}$$

Deriving it:

$$\dot{x} = \mathbf{T}\dot{z} \tag{4}$$

From equation 1:

$$\mathbf{T}\dot{z} = \mathbf{A}\mathbf{T}\dot{z} \tag{5}$$

$$\mathbf{T'}\mathbf{T}\dot{z} = \mathbf{T'}\mathbf{A}\mathbf{T}\dot{z} \tag{6}$$

$$\dot{z} = \mathbf{T'}\mathbf{A}\mathbf{T}\dot{z} \tag{7}$$

In this last equation **T'AT** can by identified as:

$$\mathbf{T'}\mathbf{A}\mathbf{T} = \mathbf{\Lambda} \tag{8}$$

So, in the equation 2 we can identify Λ as the **D** matrix which is a diagonal matrix of the eigenvalues of **A**.
For example, consider the following system of non linear differental equations:

$$\begin{aligned} \dot{x} &= -0.4x + 0.02xy \\ \dot{y} &= 0.8y - 0.01y^2x - 0.1xy \end{aligned} \tag{9}$$

The equation has a complex eigenvalue 0.8-2,8i near the point (6,20) when the Jacobian of the equation 9 is evaluated in that point, meaning that any starting point in the right quadrant will converge to it.

Now that is clear that a system of differetial equiation with complex eigenvalues can represent memory let´s analyze the propertis of a matrix to store them. Consider a set of orthogonal vectors $v^{(j)}$, j = 1,2,3,..n where $v^{(j)T}v^j = 0$

$$\mathbf{W} = \sum_{j=1}^{r} v^{(j)T}v^j = \mathbf{V}\mathbf{V}^T \tag{10}$$

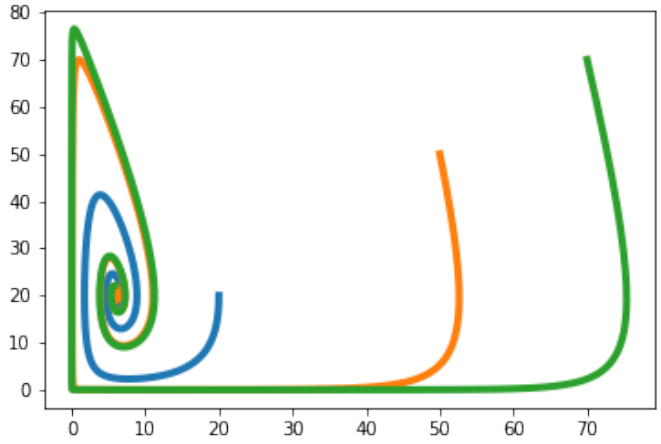

Fig. 1: Phase diagram of the system non linear differential equations 9. The lines blue, orange and green indicate the trayectories of three starting points converging to the sink point 6,20. This can be seen as an associative memory that once given certain inofrmation is able to get to the stored memory by itself.

And getting back to a linear system as in [12]:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \tag{11}$$

Where $\mathbf{x}=\mathbf{v}^k$, then:

$$\mathbf{y} = \mathbf{W}\mathbf{v}^{(k)} = \sum_{j=1}^{r} v^{(j)T} v^j v^{(k)} \tag{12}$$

$$= v^{(k)T} v^k v^{(k)} = c v^{(k)} \tag{13}$$

So, the output can be taken as a linear combination of the inputs. If the vector $\mathbf{v}^{(k)} = \mathbf{v}^{(k1)} + \mathbf{v}^{(k2)}$, where $\mathbf{v}^{(k1)} and \mathbf{v}^{(k2)}$ are orthogonal, presenting partial information in the form of $\mathbf{v}^{(k1)}$:

$$\mathbf{y} = \mathbf{W}\mathbf{v}^{(k1)} \tag{14}$$

Produces:

$$\mathbf{W} = \sum_{j=1}^{r} v^{(j)T} v^j v^{k1} \tag{15}$$

$$= v^k v^{(k)T} v^{k1} \tag{16}$$

$$= (v^{k1} + v^{(k2)})(v^{k1} + v^{(k2)})^T v^{k1} \tag{17}$$

$$= (v^{k1} + v^{(k2)}) v^{(k1)T})^T v^{k1} \tag{18}$$

$$= d v^{k1} \tag{19}$$

The equation (19) is what is called a Content addressable memory.

IV. Attention mechanism

As mentioned at the begining in [24], the attention mechanismo is derived from:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{20}$$

The equation above means that attention is a linear combination as expressed down below:

$$\mathbf{y} = \sum_{j=1}^{N} \alpha_i \mathbf{v}_i \tag{21}$$

Where:

$$\sum_{j=1}^{N} \alpha_i = 1 \tag{22}$$

To avoid scaling up the input. In equation 20 we see that there is a factor $\sqrt{(d)}$ that is similar to the factor d in equation 19, so it leads to consider also that the product $\mathbf{QK}^T$ can be replaced by $\mathbf{V}^T\mathbf{V}$ and finally multiplying by $\mathbf{V}$ and adding over all N we get to an equation simillar to 12 where the softmax accounts for the non-orthoganality of the v^k vectors.

V. Softmax interpretation

The intepretation of the softmax can be deduced from the entropy function [4]. Assuming the multiplicity W of a combinatorial event:

$$W = \frac{N!}{n1!n2!...n_t!} \tag{23}$$

Using the Stirling approximation $x! \approx (x/e)^x$:

$$W = \frac{(N/e)^N}{(n1/e)^{n1}...(n_t 1/e)^{n_t}} \tag{24}$$

$$= \frac{(N)^N}{(n1)^{n1}...(n_t 1)^{n_t}} = \frac{1}{p_{n_1}^1 ... p_{n_t}^t} \tag{25}$$

Then:

$$ln\, W = -\sum_{1}^{t} n_i ln\, p_i = \frac{S}{k} \tag{26}$$

The equallity (27) is the entrophy divided by the Bolztmann contant. In the mean time wil will no worry about it. What is important it to to find a probablity distribution that obeys that constrains of an energy system and the probabilty sum and also maximizes the entropy. That is to say:

$$\sum_{1}^{t} p_i = 1 \tag{27}$$

And the average energy:

$$\epsilon = \frac{E}{N} = \sum_{1}^{t} \epsilon_i p_i \tag{28}$$

The solution to that is:

$$p_i^* = \frac{e^{\beta \epsilon_i}}{\sum_1^t n_i e^{\beta \epsilon_i}} \tag{29}$$

Equation (29) shows that the softmax function is indeed the probability of and event that maximizes the entrophy. In other words is telling you what is the most probable macrostate possible(in terms of thermodynamics) taking certain input.

VI. Lagrangian and Hamiltonian dynamics

In order to study certain Energy based models it is necessary to understand the foundation of classical mechanics. The idea is that knowing the differential equations of a neural networks it is then possible to derive an energy model that can help to model a system.

A. Lagrangian mechanics

The root of Lagrangian mechanics come from the minimization of the atcion. Assuming that a system is decribed in generalize coordinates. The action is expressed as in equation (30)

$$S = \int_{x_1}^{x_2} L(\mathbf{q}, \dot{\mathbf{q}}, t)dt \tag{30}$$

The value of S is maximized or minimized when the Euler-Lagrange (31) holds:

$$\frac{\partial L}{\partial q} - \frac{d}{dx}\frac{\partial L}{\partial \dot{q}} = 0 \tag{31}$$

B. Hamiltonian mechanics

As known from clasical mechanics, the Hamiltonian is the total energy of a system as in equation (32).

$$H = T + V \tag{32}$$

And the Lagrangian is:

$$L = T - V \tag{33}$$

The Legendre transformation allows us to move from the Lagrangian to the Hamiltonian as follows:

$$E = \left(\sum_{n}^{i=1} \dot{q}_i \frac{\partial L}{\partial \dot{q}}\right) - L \tag{34}$$

VII. Biological Model

The article from Hodkin-Huxley describes the dinamic equations of the action potencial [7]. The figure 2 shows the diagram that resembles the model. The equations that govern the model are the folllowing:

$$I = C_m \frac{dV}{dt} + I_{input} \tag{35}$$

The total input current is:

$$I = I_{NA} + I_K + I_l \tag{36}$$

The Currents as in equations (35) are the Sodium (Na), Potasium (K) and a Leakage current (l). Those current have

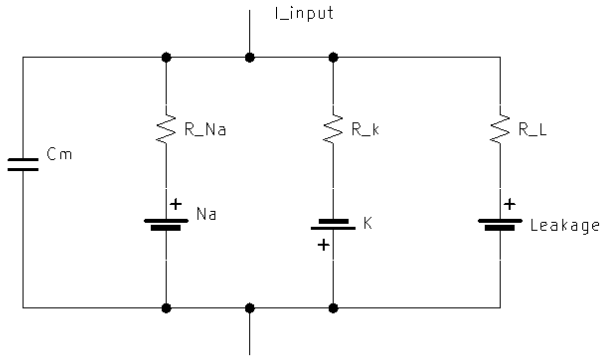

Fig. 2: The electrical equivalent circuit of the Hogdkin Huxley Model [**?**].

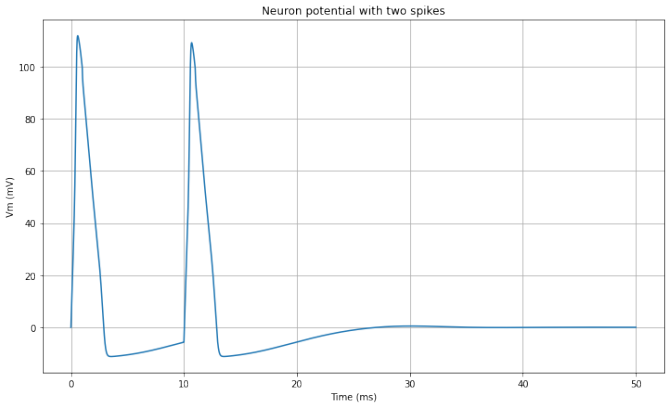

Fig. 3: The electrical response of a Hogdkin Huxley Model.

associated conductances g_{Na}, g_K and g_l. However, the conductances are non-lineal fuctions as they obey diffusion laws. Therefore:

$$I = C_m \frac{dV}{dt} + \overline{g}_K n^4 (V - V_K) + \overline{g}_{Na} m^3 (V - V_{Na}) + \overline{g}_l (V - V_l) \tag{37}$$

Where:

$$\frac{dn}{dt} = \alpha_n (1 - n) - \beta_n n \tag{38}$$

$$\frac{dm}{dt} = \alpha_m (1 - n) - \beta_m m, \tag{39}$$

$$\frac{dh}{dt} = \alpha_h (1 - n) - \beta_h h \tag{40}$$

In the above equations, the fact that they vary acccording to the voltage, allows the spikes to present transient forms different from the ones present in an electric RC circuit as in the figure 3.

Taking into account the equation (37). The model can be rewritten as:

$$I = C_m \frac{dV}{dt} + \sum_{i=1}^{N} M_{\mu i} g_i \tag{41}$$

Reweriting it:

$$\tau \frac{dV}{dt} = \sum_{i=1}^{N} \xi_{\mu i} g_i - I \tag{42}$$

VIII. Large Asociative Memory Problem

In the article [17]. Krotov and Hopfield address the problem of creating a plausible energy model. The model is based on RBM[1] and deduce the equations for a feature and hidden layers. The equatiosn are:

And get transformed into Hamiltonian energy models via a Legendre transformation.

References

[1] Brain model. (2021, 11 25). *Brain Model*. [Notebook]. Available at $https://github.com/ghgv/Brain_model$

[2] H. K. Hartline, 'The response of single optic nerve fibers of the vertebrate eye to illumination of the retina', *Am J Physiol* vol 121, pp :400–415, 1938

[3] D. H. Hubel, T. N. Wiesel,'Receptive fields of single neurones in the cat's striate cortex' *J. Physiol*,,pp 574-591,1959

[4] Ken A. Dill, Sarina Bromberg,'Molecular Driving Forces, Statistical Thermodynamics in Biology Chemistry, Physics, and Nanoscience, 2nd ed.' ,,pp 574-591,1959

[5] LeCun, Yann, 'A Path Towards Autonomous Machine Intelligence' , 2022.

[6] McCulloch, Warren S and Pitts, Walter,'A logical calculus of the ideas immanent in nervous activity' *The bulletin of mathematical biophysics.*,vol. 5, no. 4, pp. 115-133,1943.

[7] A. L. Hodgkin and A.F. Huxley,'A quantitative description of membrane current and its application to conduction and excitation in nerve' *Physiological Laboratory, University of Cambridge.*,1952

[8] Hopfield, J J.,'Neural networks and physical systems with emergent collective computational abilities' *Proceedings of the National Academy of Sciences* ,pp 2554-2558, 1982

[9] Hubert Ramsauer, Bernhard Schaefl, Johannes Lehner, Philipp Seidl ,Michael Widrich , Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, Victor Greiff, David P. Kreil, Michael Kopp,Genter Klambauer,Johannes Brandstetter,Sepp Hochreiter, 'Hopfield Networks is all you need',,2021

[10] Donald O. Hebb, 'The Organization of Behavior',,1949

[11] Fukushima, K..,'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position' *Biological Cybernetics.*,pp. 193-202,1980

[12] Oh, Cheolhwan and Stanisław H. Żak. "The Generalized Brain-State-in-a-Box (gBSB) Neural Network : Model , Analysis , and Applications." (2005).

[13] B. Olshausen, D. Field,'Emergence of simple cell receptive field properties by learning a sparse code for natural images' *Nature*, 1996

[14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, ' Gradient-Based Leardning Applied to Document Recognition' *Proc. of the IEEE*, 1998

[15] Seppo Linnainmaa, 'Taylor expansion of the accumulated rounding error'BIT ,*Numerical Mathematics*, vol 16, pages 146–160, 1976

[16] Peter H. Schiller, 'Parallel information processing channels created in the retina',*Proceedings of the National Academy of Sciences* , 2010

[17] D. Krotov, J.J. Hopfield, 'LARGE ASSOCIATIVE MEMORY PROBLEM IN NEUROBIOLOGY AND MACHINE LEARNING', *ICLR 2021*, 2021

[18] N. V. Medathati,H. Neumannb, G.S.Masson,P. Kornprobst, 'Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision', *Elsevier*,2016

[19] R. Rojas, 'Neural Network: A Systematic Introduction', *Springer*,1996

[20] A.L. Yuille, Anand Rangarajan, 'The Concave-Convex Procedure',*Neural Comput*,2003

[21] '$http://vision.psych.umn.edu/users/kersten/kersten-lab/courses/Psy5036W2017/Lectures/17_PythonForVision-/Demos/html/2b.Gabor.html$'

[22] M. Fee, '$https://Introduction-to-neural-computation-spring-2018/lecture-notes/MIT9_40S18_Lec02.pdf$'

[23] Johannes Brandstetter, '$https://ml-jku.github.io/hopfield-layers/$'.

[24] Attention is all you need, A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Kaiser, and I. Polosukhin.*Advances in Neural Information Processing Systems* , page 5998–6008. 2017

[1] Boltzmann Machines