# On the Physical Underpinnings of the Unusual Effectiveness of Probabilistic and Neural Computation

Sandip Tiwari, Damien Querlioz

# On the Physical Underpinnings of the Unusual Effectiveness of Probabilistic and Neural Computation

Sandip Tiwari
School of Electrical and Computer Engineering
Cornell University
Ithaca, New York

Damien Querlioz
Centre National de la Recherche Scientifique
Université Paris-Sud
Orsay Cedex, France

***Abstract*—Probabilistic and neural approaches, through their incorporation of nonlinearities and compression of states, enable a broader sampling of the phase space. For a broad set of complex questions that are encountered in conventional computation, this approach is very effective. In these patterns-oriented tasks a fluctuation in the size of data is akin to a thermal fluctuation. A thermodynamic view naturally applies to this computational style to information processing and from this reasoning one may estimate a variety of interesting consequences for computing: (a) efficiencies in energy, (b) complexity of tasks that can be tackled, (c) inaccuracies in inferences, and (d) limitations arising in the incompleteness of inputs and models. We employ toy model examples to reflect on these important themes to establish the following:**

- **A dissipation minimum can be predicted predicated on the averaged information being discarded under constraints of minimization of energy and maximization of information preservation and entropy. Analogous to the $k_BT\ln 2$ for the randomization of a bit, under biological constraints, the $\sim -70\ mV$ base and $\sim 40\ mV$ peak spike potential are then a natural consequence in a biological neural environment. Non-biological, that is, physical implementations can be analyzed by a similar approach for noisy and variability-prone thermodynamic setting.**
- **In drawing inference, the resorting to Occam's razor as a statistical equivalent to the choice of simplest and least number of axioms in developing of a theory conflicts with Mencken's rule—for every complex problem, there is an answer that is clear, simple and wrong—as a reflection of dimensionality reduction.**
- **Between these two factors, it is possible to make a measure of the error bound predicated on the averaged information being discarded and being filled in, and**
- **This lets one predict the upper limits of information processing rate under constraints.**

**These observations point to what may be achievable using neural and probabilistic computation through their physical implementation as reflected in the thermodynamics of the implementation of a statistical information mechanic engine that avoids computation via deterministic linear algebra.**

## I. Introduction

Probability estimations through the various styles—neural of the different varieties, Bayesian, et cetera—are estimations on limited samples using nonlinearities and compression. Finding maximum likelihoods, or minimizing energy consumption, and other objectives of a computation are subject to the problems of "fitting" that is natural to statistical model building. Entropy, energy, likelihood are a common theme here and therefore the thermodynamics of information and of physical implementations relevant. Employing the tools of neural and probabilistic computation and of statistical mechanics approaches, we use example toy models to point out a number of interesting conclusions and directions worthy of deeper exploration.

### A. Spiking action potentials

The action potential in spiking in the living biological world is viewable through the constraints of thermodynamics. The action potential moves down an axon whose simplest model is that of a capacitive membrane across which an electromotive potential exists due to the ion concentration differences and where conductance channels open and close. This signaling is dissipative. But, absence of signaling too is dissipative. So, the potentials, currents and times need to be consistent within the energetic constraints. The spiking and the noise together make this signaling mechanism effective. For this problem, the equilibrium potential is calculable from the ionic concentration across the membranes, with $K^+$ dominating, but $Na^+$, $Ca^+$ and $Cl^-$ also present. $K^+$ concentration is higher outside the tubule, while in case of $N^+$ it is higher inside. The biologically sustainable concentrations are in the order of few $mM$ to few $100s\ mM$. For example, for $K^+$, $Na^+$ and $Cl^-$ outside/inside these are $5/140$, $140/12$ and $20$.

The diffusive and electrical flow balancing establishes the reversal potential, which is the voltage across the specific ion channel during its operation. This reversal potential, for $K^+$ flow in its channel, is

$$V_{rev} = \frac{RT}{zF}\ln\frac{[K^+]_{out}}{[K^+]_{in}} = \frac{8.314\times 310}{96845}\ln\frac{5}{140} = -88.7\ mV. \tag{1}$$

Here, $R$ is the gas constant ($8.314\ J/K.mole$), $T$ is the body temperature $310\ K$), $z$ is the ionicity (1 for $K^+$), $F$ is Faraday constant ($96485\ J/V.mole$) and concentrations of ions is a ratio in identical units. $Cl^-$, which is not actively pumped,

settles at a reversal potential close to the resting potential determined by other ions. Chlorine is also highly impermeable. This resting potential, absent any activity, is a balance of concentrations and the permeabilities of their channels. Again following the Nernst equation approach,

$$V_{rest} = \frac{RT}{F} \ln \frac{\sum_i \pi_i [A^+]_{out} + \sum_j \pi_j [B^-]_{in}}{\sum_i \pi_i [A^+]_{in} + \sum_j \pi_j [B^-]_{out}}, \quad (2)$$

where $\pi$s are permeabilities. The size of the ion ($Na^+ < K^+(0.138\ nm) < Ca^+$)[1], for example, and the size of that ion's pore matter for resting potential. $\pi_{Na}/\pi_K < 0.01$. The resting potential is maintained by active ion pumping to compensate for leakages. The pumps—marvels of near-ideal electrochemomechanical coupling—and the permeability lead to smaller resting potential, which for these parameters, including leakage of $K^+$, $Na^+$ and $Cl^-$ are $1 \times 10^{-6}\ cm/s$, $2 \times 10^{-8}\ cm/s$ and $5 \times 10^{-10}\ cm/s$[2].This resting potential is $-78\ meV$, and is in range that is measured across species and cell types.

Noise is essential in a driven nonlinear dynamic system[3].The noise is Poissonian—$1/f$-like as in many electronic physical systems. Fisher information[4]—as a measure of the ability to estimate a parameter as well as of state of disorder—gives us a tool to extracting the consequence that the high voltage of the the action is $\sim 40\ mV$ which is slightly above the thermal voltage to maintain a judicious signal-to-noise ratio where synchronization becomes achievable.

The $+40\ \ mV$ peak—of the order of $k_BT$—in action potential is significant in its role in how noise, spiking and information processing interact. And this spike contains $\approx 0.5 \times 110\ mV \times 10^{-3}\ s \times 3\ pA \approx 165\ aJ \approx 40000 k_BT$ of energy. The membrane signaling process is a capacitance-based, signaling is through conductance of channels that are manifested in the voltage spike.

So, noise and thermodynamics give a powerful example in the low energy processing of information in the living system.

## II. Mutual information, convolution, information aggregation and separation

Information is relative. It a measure of the difference between two levels of uncertainty. Reduction in this uncertainty is gaining of information. Shannon entropy is one of our tools to quantifying uncertainty over several states of variables (say $X$). The observation of other variables (say $Y$), informing us of the state $X$ is the mutual information. It is something in common. The mutuality of this information is also embedded in the convolutions—pairwise and higher—and this underlies the success of convolution networks as well as in the use of partition functions. Observations on $Y$, either increase or leave the information on $X$ unchanged. This lets us write the rules of aggregation of information.

With $H(\mathbf{X}) = -\sum_x \mathfrak{p}(x) \log_2 \mathfrak{p}(x)$ as the Sannon entropy, the mutual information between the two variables is

$$I(X;Y_i) = H(X) - H(X|Y_i). \quad (3)$$

Information is a differential so we may write this as $I(X;Y_i) = -\Delta H(X)/\Delta Y_i$. As measurements are accumulated over $Y$, it is now possible to account for the change of entropy—a discrete calculus of uncertainty. One can now see that this leads to

$$H(\mathbf{X}) \leq H(\mathbf{X}|\mathbf{Y}_{k-1}) \leq H(\mathbf{X}|\mathbf{Y}_k), \ \ and \quad (4)$$

$$I(\mathbf{X}|\mathbf{Y}_k) \geq I(\mathbf{X};\mathbf{Y}_{k-1}) \geq \cdots \geq I(\mathbf{X}|\mathbf{Y}_1) \quad (5)$$

Use of chain rule now lets us aggregate this information in the form

$$\begin{aligned} I(\mathbf{X}|\mathbf{Y}_k) &= H(X) - H(X|Y_k) \\ &= -\sum_{i=1}^{k} \frac{\Delta H(X)}{\Delta Y_i} - \sum_{i>j=1}^{k} \frac{\Delta^2 H(X)}{\Delta Y_i \Delta Y_j} - \cdots \end{aligned} \quad (6)$$

where the second term on the right is just $I(X;Y_j)|Y_i) - I(X;Y_j)$. Information aggregation produces both a synergy and a redundancy of aggregation and these relationships and their extensions let us formalize information gain and translate them to probability distributions. For example, information gain about $X$ from a pair ($Y_1$ and $Y_2$) is the sum of independent mutual information with $X$ and an additional term. The first here is independent mutual information with $X$ and the second is the correlations between the $Y_1$ and $Y_2$ variables.

One can see in this aggregation picture the ability to extract features as a synergy. The simplest examples are logical gates such as $NAND$, $NOR$, $XOR$. These can be viewed both as a neural network as well as feature extractor. They tell us about the synergy that exists at the input that the gate aggregates and where the mutual information and convolutions matter. The information theory view here has related to us the entropic notions about how this information aggregation has taken place and its informational features. The problem of statistical inference gains from the insights of this discussion.

## III. Thermodynamic implications for inference

For the difficult reverse problem of inference, machine learning approaches employ a number of thoughts that we have already mentioned. Fisher's measure—the maximum likelihood estimate—is one. Maximum entropy as a guide to prescribe initial probabilities is another one since it leads to inferences being dependent only on the data and the conditions for which there is no prior information. Introduction of noise—randomness—is a tool to faster computation and is also useful where priors cause problem.

Thermodynamics teaches us a number of important principles. One is the principle of maximum likelihood where the best estimators should most duplicate data. The second is the principle of maximum entropy which prescribes no bias to internal states of the variables. The entropies from these two views need balancing. The third important principle from thermodynamics is the need to balance minimum energy and maximum entropy at any temperature. This view can be integrated in the algorithms of machine learning[6].

For information, postulate the Kullback-Leibler divergence between distribution of a target and an empirical distribution

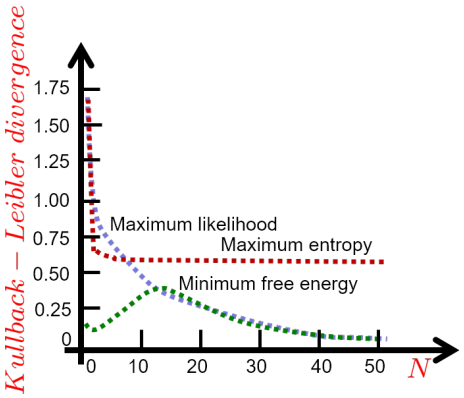


Fig. 1. The Kullback-Leibler divergences between estimated probability functions and true probability functions based on maximum likelihood, maximum entropy and minimum free energy estimate. The problem assumes an entropy of 0.5 with 3 internal states.

as the internal energy. By regarding fluctuation in data size as a thermal fluctuation, temperature can be postulated through the distortion between estimator and the probability function estimated where one data has been removed. The free energy (a Helmholtz energy, $F = U - H/\beta$, where $F$ is the free energy, $U$ is the internal energy and $\beta$ represents a data temperature) then follows from the internal information energy, the Shannon entropy and the temperature. And now one can apply the constraint that the probability functions are to arise from minimization of the free energy. In this formulation, the maximum likelihood is approached when the number of data points is high. This is the condition in the energy equation where $\beta \to 1$.

If the number of data points is low—the condition when many of machine learning algorithms have large error—the minimum free energy estimate can be more accurate.

Only when the number data points is high does $\beta \to 1$, and maximum likelihood is approached. However, this minimum free energy estimate continues to be applicable to the small data limit. Figure 1 shows the divergences for the different approaches and the efficacy of this minimum free energy approach is quite noticeable. For $N < 15$ the three approaches give widely different estimates with the minimum free energy approach being the most accurate.

This figure also shows through the divergence at the low $N$ count, the issue of errors. Techniques such as resorting to Occam's razor in drawing inference can produce serious errors. In the physical world and this machine world, one also encounters Mencken's rule, which states that for every complex problem, there is an answer that is clear, simple and wrong. Simplest explanations do not always comport.

This broad outline describes a mapping of the statistical physical techniques to information data, and similar properties as those derived in statistical physics, such as energy fluctuations, variances, et cetera, may be found. The approach shows that ranging over small to large sample sizes, the approach gives more accurate estimations compared to maximum likelihood or maximum entropy approaches alone. It reduces error bounds.

## IV. Neural networks as physical models

Forms of neural networks have now shown a variety of properties that correspond to those of physical systems. A stochastic neural network is a Boltzmann machine, that is, a thermodynamic model[7]. The example in section III was a specific example of use of thermodynamic principles as an approach to understanding the statistical consequences. Variational renormalization groups have been mapped on the networks [8] and have been tested with two-dimensional Ising models. Features such as entanglement entropy are apparently inherent outcomes of the coarse grained organization in the networks. It has also been argued[9] that the deeper networks are an ersatz learning with exponentially fewer parameters because of how the general physical functions arising in physical laws can be translated onto a network.

Physics favors simple probability distributions and this is what the networks also prefer. In the correspondence between the physical and the network, the correspondence between a quadratic Hamiltonian to a Gaussian probability distribution, locality to sparsity, translational symmetry to convolution, and free energy difference to the Kullback-Leibler divergence (we employed this in section III), can be easily seen.

An illustration of the success of such a correspondence is the observation on phase transitions where Ising models are useful toy models. The neural network—simple or multi-layered—are a corresponding convenient tool to observe the noise, and the breaking of symmetries in regions of transition as the network evolves from a random state to a phase transitioned state with a change in temperature.

Figure 2(a) and 2(b) show results—of order and of correlation/randomization—of two-dimensional Ising simulations through a restricted Boltzmann machine by sweeping temperature. One can see, as the temperature parameter $T$ is lowered, the appearance of magnetization in (a) and the frequency of the weight amplitudes in (b) in a restricted Boltzmann machine. When $T$ is high, the spins are random, with either polarity of spins evenly distributed, and the weight histogram is strongly peaked. The hidden layer is "decoupled." When $T$ is low, spin is polarized and the frequency is flat.

## V. Information, errors, and energy

The challenge of rebooting computing and the applicability of the probabilistic and neural techniques to rebooting computing is multifold.

Machine learning techniques employing neural techniques in all their forms, probabilistic versions, together with Bayesian approaches without overfitting constraints are all still affected by two considerations. The first is that, while any new observation that fits with the model already learned is a major success, any observation that does not fit with the model leads to an error. So correlations or information synergy not observed before, or fewer data, both lead to more significant error. The second consideration is that in any network where the calculation of probabilities, weights,

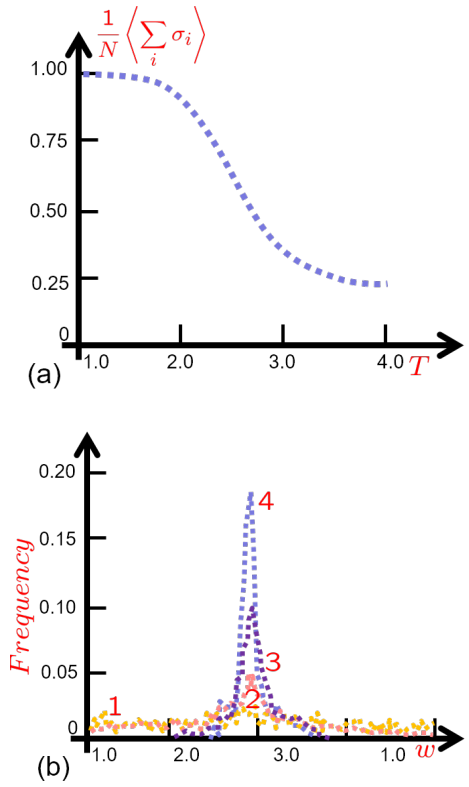


Fig. 2. (a) shows "magnetization" of the Ising model and (b) shows the weight distributions as simulations proceed through the point of criticality.

et cetera, are performed through deterministic Boolean algebra operations, the calculation is traditionally performed to many significant bits and is therefore very energy consuming.

Section IV noted a direct correspondence between the physical and the network. Since any neural and probabilistic computation is going to be subject to the inherent errors arising in the model that has been learned, it is not unreasonable to speculate that employing the physical for creation of the machine learning primitives would be of use. Among these, low energy probability generations, convolutions, Gaussian probability distributions, and stochastic noise are all quite foreseeable. Multiple possibilities exist for each one of these.

The details of how much energy is consumed for a given task certainly depends on the scale and the nature of task at hand. Boltzmann entropy change is high at the point of phase transition in the physical world. The same will also hold true for the data world's machine learning. If $\eta$ is a scaling factor for $k_BT$ that is reflective of the technology's per bit manipulation energy, and $N$ bits are being manipulated, the energy consumed is $\eta k_BTN\log_2 N$. Section I had one example of this energy need in the presence of noise in a biological spiking arrangement. Physical implementations will be different. But, this energy is quite small even if $\eta$ were to be $10^5$. The challenge is to find the model physical embodiments for these primitives that will perform the function desired in the machine learning system.

## VI. Conclusion

In this work, we have explored a few themes to build an argument related to rebooting of computing. Section I showed how thermodynamic principles suffice in predicting action potentials while utilizing noise in a driven nonlinear dynamic system. Neural networks and probabilistic systems compress states, employ nonlinearities, and work with noise of data size with many equivalences to the physical thermodynamic systems. This thermodynamic basis was employed to explore the question of information aggregation and separation in machine learning. As an example, we showed the increased accuracy of of minimum free energy estimation approach. We extended this thermodynamic discussion to show the equivalences at work in neural network environment for a model physical system. This bringing together of the physical and the machine approaches have let us make remarks on errors and energies that are potentially the limits for the set of problems for which the machine learning approaches are most suitable. We remain far away from these limits.

## Acknowledgment

Sandip Tiwari acknowledges the hospitality of Université Paris-Sud during summers. Damien Querlioz acknowledges the support of European Research Council, BAMBI EU collaborative FET project, ANR, and the French Ministère de l'écologie, du développement durable et de l'énergie.

## References

[1] B. Hille, *Potassium channels in myelinated nerve. Selective permeability to small cations*, J. Gen. Physiology, **61**, 669–686(1973)

[2] P. Ronald and J. MacGregor, *Theoretical mechanics of biological neural networks*, Publisher.

[3] A. R. Bulsara and L. Gammaitoni, *Tuning in to noise*, Physics Today, **49**, 3, 39(1996)

[4] For a discussion of Fisher information, see B. R. Frieden, *Science from Fisher information: A unification*, Cambridge, ISBN 0-521-00911-1(2004)

[5] Luís M. A. Bettencourt, *The rules of information aggregation and emergence of collective intelligent behavior*, Topics in Cognitive Science, 598–620(2009)

[6] T. Isozaki, *A thermodynamical approach to probability estimation*, arXiv:1201.1284v2, 12 Dec(2012)

[7] G. Torlai and R. G. Melko, *Learning thermodynamics with Boltzmann machines*, arXiv:166006.02718v1, 8 June(2016)

[8] P. Mehta and D. J. Schwab, *An exact mapping between the variational normalization group and deep learning*, arXiv:1410.3831v1, 14 Oct(2014)

[9] H. W. Lin and M. Tegmark, *Why does deep and cheap learning work so well?*, arXiv:1608.08225v2, 28 Sep(2016)