Article

On the Theory of Deep Learning: A Theoretical Physics Perspective (Part I)

Alejandro Chinea Manrique de Lara

Departamento de Fisica Fundamental, Facultad de Ciencias de la UNED, Paseo Senda del Rey No. 9, 28040 Madrid, Spain; alejandro.1138@gmail.com

Academic Editor: name

Abstract: Deep learning machines are computational models composed of multiple processing layers of adaptive weights to learn representations of data with multiple levels of abstraction. Their structure is mainly reflecting the intuitive plausibility of decomposing a problem into multiple levels of computation and representation since it is believed that higher layers of representation allow a system to learn complex functions. Surprisingly, after decades of research, from learning and design perspectives these models are still deployed in a heuristic manner. In this paper, deep learning feed-forward machines are modeled from a statistical mechanics point of view as disordered physical systems where its macroscopic behavior is determined in terms of the interactions defined between the basic constituent of these models, namely, the artificial neuron. They are viewed as the equilibrium states of a theoretical body that is subject to the law of increase of the entropy. The study of the changes in energy of the body when passing from one equilibrium state to another is used to understand the structure and role of the phase space of the system, the stability of the equilibrium states, and the resulting degree of disorder. It is shown that the topology of these models is strongly linked to their stability and resulting level of disorder. Furthermore, the proposed theoretical characterization permit to assess the thermodynamic efficiency with which information can be processed by these models, and to provide a practical methodology to quantitatively estimate and compare their expected learning and generalization capabilities. These theoretical results provides new insights to the theory of deep learning and their implications are shown to be consistent through a set of benchmarks designed to experimentally assess their validity.

Keywords: Deep Learning; Thermodynamics; Learning and Generalization; Diophantine equations

1. Introduction

Deep feed-forward neural networks, with multiple hidden layers, have achieved remarkable performance across many domains [35,56], perhaps also motivated by the fact that these methods require very little engineering by hand, and have also taken advantage of the increasing amount of computational resources and data. These kind of machine learning methods are principally characterized because of the fact that they use multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level, starting with the raw input, into a representation at a higher, slightly more abstract level [35] since it is believed that with the composition of enough such transformations, very complex functions can be learned. The problem of how well an artificial neural network may infer an unknown classification rule from input-output examples defines the generalization error [11,30,48]. It is a well-known fact in the machine learning community that the architecture of a neural network (number of units and topology of connections) can have a significant impact on its generalization performance in any particular application. At the same time, the random input-output rules learned by feedforward neural networks also defines the problem of storage capacity [28]. The Vapnik-Chevornenkis dimension (VC hereafter) is a combinatorial concept that relates both problems [68]. More specifically, it is a measure of the capacity or expressive power of the family classification functions realized by a statistical classification algorithm, that is, its expresses the size of the largest set of input patterns for which all 2^{VC} combinations of binary inputs can be

learned by the neural network. In other words, it is a measure of the complexity (capacity) of the space of functions that can be learned by a statistical classification algorithm.

In the last few decades a substantial effort has been devoted to the study of role of the VC dimension in the learning and generalization capabilities of neural networks [26,47]. In particular, it was shown why networks with large storage capacities and VC dimensions realize more complex operations compared to those with smaller capacities. Using statistical physics methods it has been shown that the phase space and the VC dimension can play independent roles in the process of generalization for a parity machine [47], that is, for a fixed phase space dimension, the VC dimension grows arbitrarily by increasing the number of hidden units. However, up to a critical number of training examples (growing with the VC dimension), generalization is impossible [47]. In the case of committee machines, the storage capacity and VC dimension converge to a finite value to the extent the number of hidden units tends to infinity [57].

In the case of Deep neural networks [35,56], it has been recently shown that deep networks with a hierarchical architecture can approximate the class of compositional functions with the same accuracy as shallow networks (i.e, one hidden layer networks) but with exponentially lower number of training parameters as well as VC dimension [51]. Similarly, a recent study has revealed the importance of the adaptive weights of early hidden layer for reliable computation in sparse and fully-connected feedforward neural networks [36,66].

Surprisingly, after several decades of research and in spite of the fact that a theoretical characterization of these model have started to emerge [51–54,59,62–64,69] a detailed understanding of the principles underlying their functioning remains unclear. According to [52] there are three main sets of theory questions about Deep Neural Networks that are relevant in several fields such as machine learning, function approximation or statistics just to mention a few.

The first set of questions are related to the power of any given architecture. For example, it is not yet fully understood which classes of functions may it approximate or learn well any given architecture [50,51], or why hierarchical networks behave so well in many practical problems. Perhaps, the most outstanding open problem related to this set of theory questions is the development of quantitative methods to estimate and compare the learning and generalization capabilities of different architectures. Although a very recent approach have started to address this problem [3], the architecture selection procedure in these kinds of models is still practically carried out in a heuristic manner which usually involves an exhaustive search through a restricted class of network structures that may require significant computational effort and yet it only searches a very restricted class of network models.

The second set of questions is about the learning process. For example, it appears that Stochastic Gradient Descent algorithms (e.g.., Backpropagation algorithm) are unreasonably efficient in deep networks compared to shallow networks. Furthermore, it is not yet understood why networks with different topology but with an identical number of parameters (e.g., an identical number of adaptive weights) learn faster than others, that is, certain network architectures reach a prescribed learning and generalization error with a reduced number of epochs (i.e., the number of presentations of the entire training set to the network) compared to others.

The third, and perhaps the most important set of question is about generalization. For example, it is not yet fully understood why minima appear easier to be found in deep networks compared to shallow networks. Concerning this question some authors have pointed out that overparametrization might explain this phenomenon [51], however this does not explain why deep networks are less affected by overfitting compared to shallow networks, that is, networks learned with stochastic gradient descent techniques exhibit good generalization, even when the number of parameters is significantly larger than the amount of training data N [45,70]. Similarly, another question not yet fully answered concerns what are the complexity notions that control or ensure generalization [37,45]. For example, one of the parameters linked to the generalization error is the storage capacity of a network and its generalization error, both concepts related by the VC dimension. Therefore, a perfect understanding of these models at theoretical level would permit not only to fully answer and connect the set of theory question

described before but particularly the deployment of systematic and principled procedures for the design of Neural network architectures (shallow or deep) achieving the best attainable learning and generalization performance with very little computational effort.

This paper is the first part of a research divided in two parts that are aimed at presenting the different contributions that has been made to the theory of Deep Learning from a novel and unifying theoretical framework rooted in thermodynamics and methods of statistical physics of disordered systems. More specifically, deep learning fully connected feed-forward network architectures are modeled as disordered physical systems from a statistical mechanics point of view [15,67] where its macroscopic behavior is determined in terms of the interactions defined between the basic constituent of these models, namely, the artificial neurons. Particularly, they are viewed as the equilibrium states of a physical system that is subject to the law of increase of the entropy. The study of the changes in energy of the body when passing from one state to another is used to understand the structure and role of the phase space of the system, the stability of the equilibrium states as well as the resulting degree of disorder of the system. It is shown that the topological structure of these networks (i.e., the number of layers and the number of units per layer) as well as the input space dimension to the networks is strongly linked to their stability and resulting level of disorder but also to the thermodynamic efficiency with which information can be processed by these models.

This general study will shed light on the role of the topology of deep learning machines in their information storage capacity, the structure of its phase space together with their resulting degree of disorder, and stability but especially how these properties relate to their learning and generalization performance. The statistical physics-based formulation (yet to be described) is used to rephrase classic machine learning concepts such as overfitting, learning, and/or generalization performance in terms of thermodynamic concepts.

Summarizing, this first part is focused on presenting the theoretical framework together with some preliminary material that advance part of the results described in the second part of this study. The second part is aimed at validating through exhaustive benchmarks the principal theoretical predictions of the model providing an alternative interpretation of the main sets of theory question concerning recent advances in the theory of deep learning [52,59]. Thus, providing new insights to those advances which include challenging some of the interpretations and conjectures that have been proposed in the last few years. Furthermore , a quantitative method fundamented in the laws of thermodynamics to estimate and compare the capabilities of different deep learning machines is also presented but not only from the point of view of their capacities for storing information and/or their learning and generalization capabilities but also through novel dimensions such as their expected computational times in the learning and recall phases of these models, thereby offering an alternative perspective to the recent method proposed in [4]. Finally, cross-disciplinary applications of the proposed framework are also presented.

The rest of this paper is organized as follows: In the next section the structural basis of the model are introduced. In turn, based upon the observation that the basic building block of deep learning feed-forward architectures is the perceptron (i.e., a single artificial neuron), and the fact that the interactions between these building block are defined by the topology of these networks, a statistical physics formulation of deep networks is presented in section 3. In particular, the partition function of a generic feed-forward deep network is obtained. Afterwards, the thermodynamic potentials are derived and used to describe differences between different deep network architectures. It is shown that the thermodynamic formalism introduced is able to characterize not only networks with different topologies but also networks with different topologies under an identical set of adaptive parameters and/or units (e.g., shallow networks against deep networks). Furthermore, several hypotheses are formulated based on a preliminary interpretation of the derived thermodynamic potentials of feedforward deep networks.

Section 4 represents the main bulk of the paper and is focused on the presentation ans analysis of the theoretical results, to this end three illustrative scenarios imposing different restrictions on the

structural parameters of the networks are studied. Specifically, each of the restrictions considered together with the topology imposes important differences in the resulting phase space of the networks permitting to understand, according to the theoretical model presented, the particularities observed in the thermodynamic potentials. A special emphasis is put on determining the influence of the structural parameters of these models, namely, the dimension of the input space to the networks and the number of hidden layers (i.e., the depth of the networks) combined with the total number of units, and the total number of adaptive weights under different network topologies. Furthermore, a theoretical interpretation grounded in the second law of thermodynamics of the results provided by the model is also provided . In particular, deep learning machines are viewed as the equilibrium states of a body that is subject to the law of increase of the entropy. The study of the changes in energy of the body when passing from one state to another is used to understand the structure and role of the phase space of the system, the stability of the equilibrium states as well as the resulting degree of disorder of the system. Section 5 is focused on the thermodynamic efficiency with which information can be processed by these models. In section 6 the theoretical results are discussed together with its preliminary implications with respect to the main sets of theory question. Finally, section 7 provides a summary of the present study and some concluding remarks.

2. Model Description

From a theoretical perspective deep neural networks can be considered as physical systems composed by a relatively large number of constituents (or subsystems) represented by the number of artificial neurons (i.e., information-processing units) that it contains which are fundamental to the operation of these machine learning models. Furthermore, from a statistical mechanics perspective, the "macroscopic" behavior of these computational models may be defined in terms of their resulting learning and generalization capabilities. It is important to note that these macroscopic properties are strongly influenced by the size of the system (i.e., the total number of neurons), and the set of interactions defined between their "microscopic" constituents (i.e., artificial neurons). Particularly, the interactions between the constituents of deep learning models are naturally expressed by the topology of these networks (i.e., the number of layers, the particular arrangement of neurons per layer, and the connectivity between neurons). For a fixed network architecture "spontaneous" fluctuations of the aforementioned macroscopic properties are the result of the particularities of the training set and/or the learning algorithms used for training these computational models. Clearly, the number of constituents that compose current state of the art deep learning networks (i.e., typically networks composed of 9 or more layers, where the number of units per layer ranges from hundreds to thousands) are not comparable with the size of the systems typically studied in statistical mechanics nor the number of possible interactions between those constituents. However, the interest here is to stress the idea whether the performance of these models measured in terms of their learning and generalization capabilities (i.e., its macroscopic properties and their fluctuations) can be interpreted in terms of the thermodynamic formalism of statistical mechanics. More specifically, the main assumption that is exploited by the model (yet to be described) is the idea that the interactions defined between the neurons in a network due to the particularities of its topology have a profound influence in the resulting learning and generalization performance of these models. As a matter of fact there is a large body of empirical evidences from biological neural networks and statistical mechanics of complex networks suggesting that the structural aspects of these networks (i.e., the topology of these networks) shape their functional dynamics [10,14,22]. Thus, one of the principal goals of the model is to assess if the aforementioned fluctuations of the macroscopic properties under study are (or not) principally due to the topology of these networks rather than a consequence of the particularitities of the training set and/or the learning algorithms used.

Moreover, it is important to emphasize that deep learning models typically operates under two phases: Namely, the training or learning phase, and the recall phase. Usually, the learning phase learning involves modification of the adaptive parameters (i.e., adaptive weights) of the deep network

by applying a set of labeled training samples (i.e., a set of known input-output patterns). Each example consists of a unique input signal and a corresponding desired response. The network is presented with an example picked at random from the training set, and the adaptive weights of the network are modified to minimize the difference between the desired response and the actual response of the network produced by the input signal in accordance with an appropriate statistical criterion. The training of the network is repeated for many examples in the set until the network reaches a steady state where there are no further significant changes in the adaptive weights.

During the recall phase the adaptive weights of the network have been fixed and the network is subject to patterns it has not seen before, but whose outputs are known and the performance of the network is monitored. It is important to emphasize that the goal of the training phase is not to learn an exact representation of the training data itself, but rather to build a statistical model of the process which generates the data. This is important if the network is to exhibit good generalization, that is, to make good predictions for new inputs during the recall phase. A deep network designed to generalize well will produce a correct input-output mapping even when the input is slightly different from the examples used to train the network. When, however, a neural network learns too many input-output examples, the network may end up memorizing the training data. Such a phenomenon is referred to as overfitting or overtraining. When the network is overtrained, it loses the ability to generalize between similar input-output patterns.

Having said this, it is important to observe that when considering deep networks as physical systems, taking a momentary "photograph" of this system, it is found that the inputs to the network, the adaptive weights of the network, and also the outputs of the network are fixed to specific real number values. Furthermore, this fact is independent of the regimen of operation of the network (i.e., training or recall phase). Thus, from a modelization point of view it can be assumed that the adaptive weights of the networks are constrained to be in a set of discrete states. A similar assumption may also be used to describe the inputs and outputs of these networks. In other words, it can be assumed that both, the adaptive weights, and the inputs/outputs of these networks can be represented by a finite set of energy states. The most important aspect of this description (yet to be described) is that it permits to abstract the two phases of operation associated to the development of these machine learning models together with the particularities of the training set and/or the learning algorithms used for training these computational models, allowing at the same time to capture the geometry of the information and the information-processing properties associated to these models that are mainly due to the interactions defined between the artificial neurons (i.e., the topology) [33,61].

In order to more formally investigate the ideas introduced so far in the following a statistical mechanics description of deep network architectures is introduced. Particularly, deep networks are viewed as physical systems where its macroscopic states are the result of the many degrees of freedom represented by the local energy configurations that result from considering the adaptive weights, the inputs and outputs to these networks as discrete random variables but following a specific energy model that takes into account the computational scheme of the constituents of these networks (i.e., the artificial neurons). Afterwards, the partition function of these networks is formulated as a generating function of energies to permit its calculation using techniques of advanced combinatorics. This formulation is aimed at permitting to characterize at theoretical level the thermodynamic behavior of such systems in terms of the properties that emerge from these random combinatorial structures resulting from the combinations of the artificial neurons that compose those networks following a predetermined topology.

3. Statistical Mechanics Formulation

From a Statistical Physics perspective the statistical description of a system with a very large number of particles in equilibrium starts first with the determination of the microscopic states accesible by the N-particles system. Secondly, with the evaluation of the probability for the system to be in a particular microstate, in the specified physical conditions [31]. This kind of problem is solved by

expressing that in equilibrium the statistical entropy is maximum under the constraints defined by the physical situation under study. Thus, in a first approach feed-forward fully connected deep learning machines can be considered as thermally isolated physical systems composed of a relatively large number of artificial neurons.

Two structural parameters are mainly responsible of the degrees of freedom of these models. Namely, the total number of particles, that is, the total number of artificial neurons, and the effective complexity of the network, that is to say the total number of adaptive weights. In other words, the degrees of freedom of these models are a function of the degrees of freedom of their constituents (the artificial neurons). These models are constructed from multiple layers of interacting artificial neurons. Particularly, the interactions between the neurons comprising the network are expressed by the topology of the network. From a theoretical perspective the topology of these networks can be interpreted in terms of the existence of a special kind of external field. More specificaly, in a similar way as occurs in the general theory of relativity where the metrical properties of space-time are regarded as the external conditions in which the universe is situated (i.e., the universe is regarded as a body situated in a variable gravitational field), the topology of these networks can be regarded as the "external conditions" or external field in which the artificial neurons of the network are situated. The effect of the external field is to alter the degrees of freedom of the artificial neurons depending on their position, that is, the layer to which the neuron belongs. In other words, the artificial neurons adquire a potential energy that is stored in these internal degrees of freedom and whose magnitude is linked to its position within the network.

Having said this, it is important to remember that the degrees of freedom of a physical system are the number of independently variable factors affecting the range of states in which such system may exist. Furthermore, if a physical system has s degrees of freedom, the positions in space of the points of the system are described by s coordinates. Different states of the system can be mathematically decribed by points in phase space (which is a purely mathematical concept) [31]. Thus, the degrees of freedom of an artificial neuron (considered as an isolated system) are constituted by their free parameters. More specifically, the total number of adaptive weights, the set of values of its inputs, both defining its state (active or not active), that also constitutes *per se* a degree of freedom (i.e., the output of the neuron).

An artificial neuron is the basic constituent of these networks and it represents an information-processing unit that is fundamental to the operation of a deep network. Three basic elements can be identified in the neuronal model, namely, a set of synapses or connecting links, each of which is characterized by an adaptive weight whose values may lie in a range that includes negative as well as positive values (unlike real brain synapses), an adder or linear combiner for summing the input signals weighted by the respective adaptive weight of the neuron, and an activation function for limiting the amplitude of the output of a neuron to some finite value.

Moreover, the structure of these models cannot be considered hierarchical, however, it does exists an implicit compositional hierarchy when considering the different subsystems comprising the structure of these networks. Namely, a deep network is composed of layers, and in turn layers are composed of artificial neurons. In other words, a layer is a subsystem that is composed of smaller subsytems represented by the neurons that it contains. Furthermore, the energy levels used to represent both, the adaptive weights, and the inputs/outputs of these networks can be mathematically modeled as a set of discrete random variables. The principal advantage of this particular formulation is that firstly, it permit to abstract the underlying phase of functioning of the network (remember the illustrative example of the photography), secondly, the spontaneous fluctuations in learning and generalization resulting from the particular training set and/or algorithm used for the learning phase of the network are implicitly embedded in the randomness introduced by the model. Furthermore, the magnitude of these fluctuations at macroscopic level can be quantified averaging out the disorder of the thermodynamic potentials.

The left part of figure (1) shows the common architectural building block of a generic feedforward neural network, that is, the artificial neuron. The right most part of the figure shows the procedure of discretization carried out over the artificial neuron. Specifically, the adaptive weights that feed the transfer function, and also the inputs to the perceptron are modeled as multi-state variables (i.e., discrete random variables). Each adaptive weight is represented by a multi-state variable that can be in a finite number of states (indicated within curly brackets). In other words, both the inputs and the adaptive weights are represented by a discrete number of energy levels.The cardinality of the set of states is in direct correspondence with their corresponding local energy values. Furthermore, the higher the energy of the state, the higher its energy value (i.e., the associated integer value). Similarly, the inputs and outputs of the network are also represented by discrete random variables. The next step is to define an energy model that captures the particularities associated to the genre of computations carried out by these models. More specifically, the energy model must take into account the given basic rules for assembling the elements conforming an artificial neuron (e.g., the set of adaptive weights, the adder for summing the input signals, and the activation function), and especially the interactions between such discrete elements which are defined in terms of the scheme of computations carried out by the neuronal model (e.g., the operation carried out by the linear combiner, the squashing effect performed by the activation function on the output signal of the neuron etc).

3.1. The Energy Model

The main idea in statistical mechanics is that every microscopic configuration C, is assigned a probability $p(C)$ which depends on its energy $H(C)$ and is given by the Boltzmann-Gibbs distribution. The partition function is defined as the normalization factor of the distribution and it can be interpreted as a generating function of energies [29]:

$$Z = \sum_{C} e^{-\beta H(C)} = \sum_{n} M(n) e^{-\beta n} \tag{1}$$

In the previous expression β is equal to the inverse of the temperature T and $M(n)$ is the number of configurations C having exactly energy equal to n, i.e., $n = H(C)$. Here the energy $H(C)$ is defined in terms of the particular arrangement of local energy states that result from the discretization procedure described herafter that is carried out over the entire structure of the network. The fact of expressing the partition function of the system as a generating function of energies permits to work under the constraints defined by both commonly used statistical ensembles: the Canonical ensemble (the physical system is in contact with a heat reservoir and only interchanges energy) and/or the Grand Canonical ensemble (the system is in contact with a heat reservoir and a particle reservoir and interchanges both energy and particles). In other words, it permits to abstract the underlying ensemble used when calculating the partition function of the system.

Moreover, the formulation of the partition function of a deep network as a generating function of energies permits to use the machinery of advanced combinatorics [19,27] to calculate the counts $M(n)$ according to a specific energy model (yet to be defined). Particularly, it permits to express the partition function of the entire network as a function of the generating functions of its elementary constituents, that is, layers and artificial neurons. Thus, the hierarchical structure of deep networks (in terms of its compositional elements) can be exploited by decomposing the calculation following the described compositional hierarchy. If we denote as $L_k(n)$ the counting sequence of the combinatorial class L_k (see appendix A for definitions and notations used hereafter) which accounts for the energy values associated to layer k of a network of L layers, i.e., $1 \leq k \leq L$, it is easy to deduce that the counts $M(n)$ associated to the combinatorial class describing the energy values of the entire network can be expressed by the convolution:

$$M(n) = \sum_{n_1=0}^{n} \sum_{n_2=0}^{n-n_1} \cdots \sum_{n_L=0}^{n-n_1-n_2-\ldots-n_k\ldots-n_{L-1}} L_1(n_1) L_2(n_2) \ldots L_k(n_k) \ldots L_L(n_L) \tag{2}$$

In turn, assuming that layer L_k of the network contains g_k units (i.e, artificial neurons), and denoting as X_i the counting sequence of the combinatorial class describing the energy values of unit i belonging to layer k, where $1 \leq i \leq g_k$. Then, the counts associated to the combinatorial class describing the energy values of this layer can be expressed by the convolution:

$$L_k(n) = \sum_{n_1=0}^{n} \sum_{n_2=0}^{n-n_1} \cdots \sum_{n_k=0}^{n-n_1-n_2-\ldots-n_{k-1}} X_1(n_1)X_2(n_2)...X_{g_k}(n_k) \tag{3}$$

Specifically, the counts representing the energy values associated to layer k are expressed in terms of the counts associated to the energy model (yet to be defined) of the artificial neurons comprising the structure of the layer.

3.1.1. The Neuronal Model: An Advanced Combinatorics Formulation

The neuronal model of an artificial neuron computes a weighted sum of the inputs (see equation (4)) to the unit that are later limited to the permissible amplitude ranges allowed by the particular activation function used. The neuronal model also includes an externally applied bias. The bias has the effect of increasing or lowering the net input of the activation function, depending on wether it is positive or negative, respectively. According to the discretization procedure described before, both the input variables and the adaptive weights of the artificial neuron are modeled in terms of multi-state random variables. Thus, if the set of states associated to the input variables and the adaptive weights are modeled respectively by the random multi-state variables ξ_{x_i} and and ξ_{w_i}, where $1 \leq i \leq n$, then the resulting product is also a multistate random variable whose set of states are bounded by the product mp. Here, for simplification purposes, the effect of the bias is not considered.

$$y = f\Big(\sum_{i=1}^{n} w_i x_i\Big) \tag{4}$$

The computational operation performed by the linear combiner of the artificial neuron can be interpreted from a combinatorics point of view as an integer composition with restricted summands and with a fixed number of parts (see appendix A). Specifically, the summands (or parts) of the composition are fixed to the number of inputs of the artificial neuron but at the same time they are only allowed to be taken from the integer set $\{1, 2, 3, \ldots, mp\}$. The elements of this set are in direct correspondence with the number of states resulting from the products $u_i = \xi_{x_i}\xi_{w_i}$ (see figure (1)). This simplification is justified by the fact that the goal is to capture the computational scheme of the perceptron, but abstracting at the same time the particularities of the training set and/or the learning algorithm employed. Thus, the i-esim summand of the composition denoted as u_i is a multi-state variable whose local energy values are drawn from the integer set $1, 2, 3, \cdots, mp$. The resulting integer composition takes the form $u_1 + u_2 + u_3 + \cdots + u_n = l$ where l is interpreted from a physical point of view as an energy value. Specifically, the ground state of the physical system modeled by the integer composition has an energy value equal to $l = n$ (i.e., the total number of inputs to the unit), whereas the maximum energy value is reached for $l = nmp$. From a combinatorics point of view, the arithmetic stucture of the composition described before can be described in terms of a combinatorial class whose generating function reads (see appendix B for the entire derivation):

$$Cp(z) = \frac{z^n}{(1-z)^n}(1 - z^{mp})^n \tag{5}$$

Moreover, from a physical point of view the squashing effect of the activation function of the neuron can be viewed as an energetic barrier. The sigmoid function is by far the most common form of activation function used in the construction of deep neural networks. Two commonly used forms of which are the nonsymmetric logistic function, and the antisymmetric hyperbolic tangent function [30]. For example, if the non-linear activation function used is an hyperbolic tangent function the resulting

energy value of the composition is approximately mapped onto two possible energy states -1 or +1 (or to 0 and 1 if the logistic function is used).

This effect is equivalent to the reassigment of the energy values associated to the set of combinatorial objects of the integer composition (see the example of figure 2) onto two arbitrary energy levels μ and $\mu + \Delta$ that would model the effect of the energetic barriers as well as the corresponding energetic gap (i.e., Δ). Specifically, from a combinatorics point of view, the combinatorial objects of the integer composition whose size (i.e., the energy value according to the modelization) is bigger or equal to a predetermined threshold δ are assigned the highest of the two possible designated values, otherwise the lowest value is used. To take into account this effect the generating function (5) can be decomposed as follows:

$$C_p(z) = \sum_{k\geq 1}^{+\infty} C_p(k) z^k = \sum_{k\geq 1}^{\delta} C_p(k) z^k + \sum_{k=\delta+1}^{+\infty} C_p(k) z^k = C_p(z)^{[1\ldots\delta]} + C_p(z)^{[1+\delta\ldots+\infty]} \tag{6}$$

The aforementioned coefficient δ that appear in equation (A13) is a threshold that models the effect of the activation function of the neuron under the assumption that a sigmoid function is used. Specifically, the generating function $C_p(z)$ is decomposed in two functions, namely, the function $C_p(z)^{[1..\delta]}$ to gives account for the objects whose size is less or equal than the value defined by the threshold, and the function $C_p(z)^{[\delta+1..+\infty]}$ for the objects whose size is bigger than the value defined by the threshold. To ensure an approximate equipartition of the set of possible energy values that are present before the energetic barrier the threshold δ is defined as follows:

$$\delta = n_{min} + \frac{(n_{max} - n_{min}) - (n_{max} - n_{min}) \bmod 2}{2} \tag{7}$$

The variables $n_{min} = n$ and $n_{max} = nmp$ in equation (7) correspond respectively to the minimum (i.e., the ground state) and maximum of the energy values that result from the integer composition at the input of the artificial neuron. It is important to remember that the summands of the composition lie in the set $\{1, 2, 3, \cdots, mp\}$ and the number of parts is restricted to the value n that corresponds to the number of inputs to the neuron. Thus, assuming a sigmoid activation function, and denoting as $X(z)$ the generating function of the artificial neuron it finally reads:

$$\begin{aligned} X(z) &= C_p(1)^{[1..\delta]} z^{\mu} + C_p(1)^{[1+\delta..+\infty]} z^{\mu+\Delta} && (8)\\ &= \left(\sum_{k=n}^{\delta} C_p(k) \right) z^{\mu} + \left(\sum_{k=1+\delta}^{nmp} C_p(k) \right) z^{\mu+\Delta} && (9)\\ &= \lambda^1_{m,p,n} z^{\mu} + \lambda^2_{m,p,n} z^{\mu+\Delta} && (10) \end{aligned}$$

where the coefficients $) = \lambda_{m,p,n}$ express the counts resulting from the reassignment of energy values commented before. In other words, the combinatorial class corresponding to the generating function $X(z)$ has objects (or atoms) of sizes μ and $\mu + \Delta$ respectively and the expression for its coefficients $\lambda^1_{m,p,n}$ and $\lambda^2_{m,p,n}$ reads respectively (see appendix B for the entire derivation):

$$\lambda^1_{m,p,n} = [z^{\mu}] X(z) = \sum_{k=n}^{\delta} \sum_{l=0}^{k} (-1)^l \binom{n}{l} \binom{k - mpl - 1}{n - 1} \tag{11}$$

and

$$\lambda^2_{m,p,n} = [z^{\mu+\Delta}] X(z) = \sum_{k=1+\delta}^{nmp} \sum_{l=0}^{k} (-1)^l \binom{n}{l} \binom{k - mpl - 1}{n - 1} \tag{12}$$

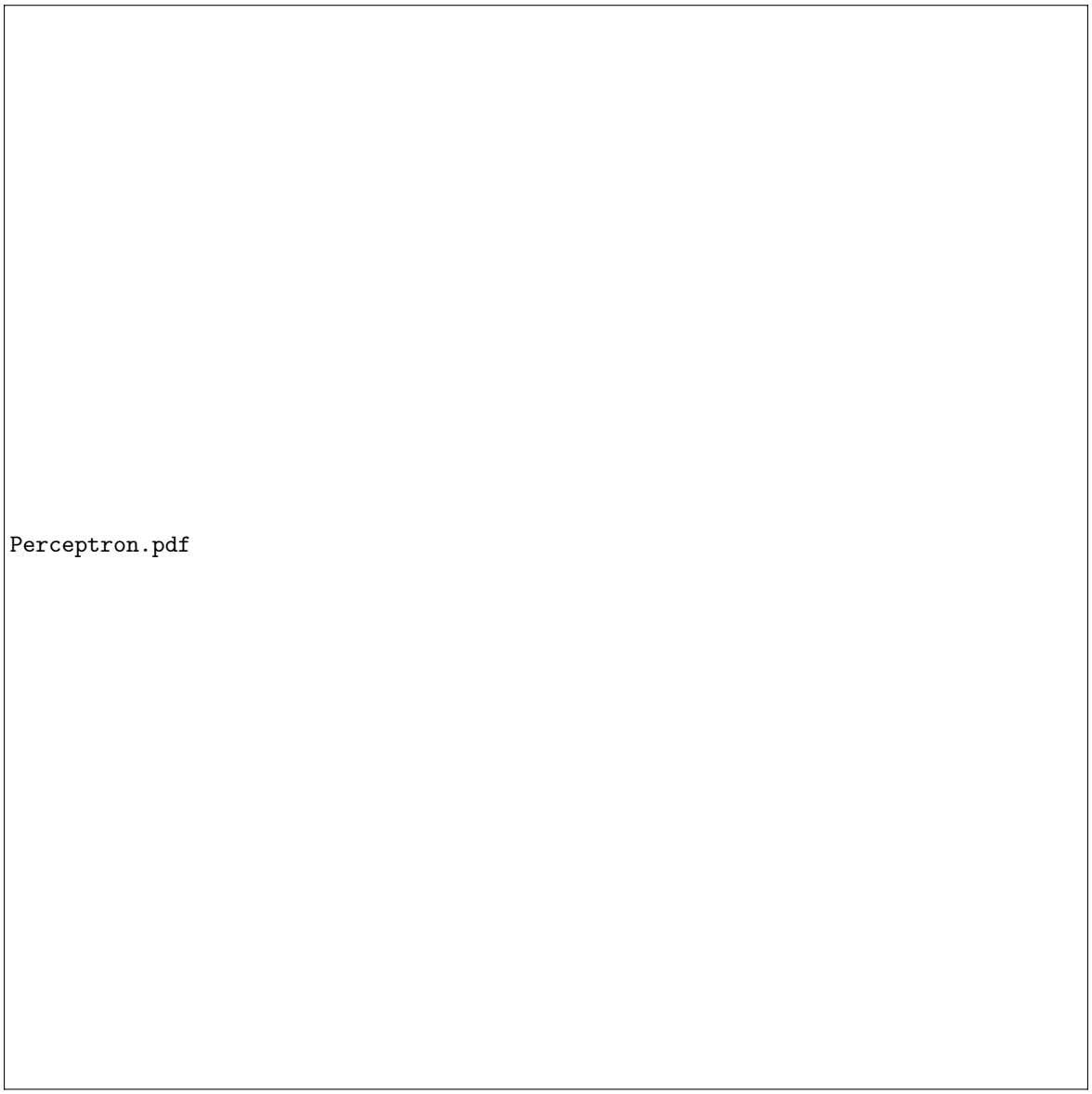

Figure 1. Graphical illustration of the combinatorial formulation of the perceptron (left side of the figure). The inputs, and the adaptive weights of the perceptron are modeled in terms of random multi-state variables. Thus, the weighted sum operation performed by the perceptron gets converted into a integer composition (interpreted as an energy value) where the summands (or parts) of the composition are fixed to the number of inputs of the perceptron and, at the same time, they are only allowed to be taken from the integer set $1, 2, 3, \cdots, mp$. The elements of this set are in direct correspondence with the number of states that result from the products $u_i = X_i W_i$, where $1 \leq i \leq n$ (the bottom right part of the figure). This simplification captures the computational scheme of the perceptron but abstracting, at the same time, the particularities of the training set and/or the learning algorithm employed.

Integer Composition	Combinatorial Object	Energy	Energy Reassignment
u_1 u_2 $\{1,2,3\}$ $\{1,2,3\}$ $u_1+u_2=l \qquad l\in[2..6]$	1+1 1+2 2+1 1+3 3+1 2+2 2+3 3+2 3+3	2 3 3 4 4 4 5 5 6	1 1 1 3 3 3 3 3 3

Integer Composition	Combinatorial Object	Energy	Energy Reassignment
u_1 u_2 u_3 u_4 $\{1,2\}$ $\{1,2\}$ $\{1,2\}$ $\{1,2\}$ $u_1+u_2+u_3+u_4=l \qquad l\in[4..8]$	1+1+1+1 1+1+1+2 1+1+2+1 1+2+1+1 2+1+1+1 2+2+1+1 1+2+2+1 1+1+2+2 2+1+1+2 1+2+1+2 2+1+2+1 2+2+2+1 1+2+2+2 2+1+2+2 2+2+1+2 2+2+2+1 2+2+2+2	4 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8	1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3

Figure 2. The figure contains two tables with examples of how the effect of the transfer function of the perceptron is modeled. This modelization accounts with the typical sigmoidal or hard-limiting functions typically used in deep learning paradigms. The transfer function is modeled as an energetic barrier that reassigns the energy values of the combinatorial objects of the composition onto two arbitrary energy levels μ and $\mu+\Delta$ that model the effect of the energetic barriers as well as the corresponding energetic gap (i.e., Δ). Specifically, all the combinatorial objects whose energy value (i.e., the size of the composition) is bigger or equal to a predetermined threshold δ (i.e., $\delta=4$ for the table on top, and $\delta=6$ for the other) are assigned the highest energetic value of the barrier (i.e., $\mu+\Delta$), otherwise the lowest value is used (i.e., μ). Each table is composed of four columns that show graphically (from left to right) the integer composition that models the weighted sum operation performed by the perceptron, the combinatorial objects associated to the integer composition, the energy value assigned to each combinatorial object which is equal to the size of the composition l, and the values resulting from the energy reassignment that models the effect of the transfer function. The values of the energetic barriers in this example are +1 and +3 respectively. These values are obtained when $\mu=1$ and $\Delta=2$.

3.2. The Partition Function

The first step to calculate an expression of the thermodynamic potentials of a deep network is to calculate the partition function of the system. From equation (**??**) it can be deduced that the generating function of the counts $M(n)$ can be expressed as the product of the generating functions of the combinatorial classes describing the energy values at each layer of the network L_k where $1 \leq k \leq L$.

$$M(z) = \prod_{k=1}^{L} L_k(z) \tag{13}$$

Similarly, the generating function of the combinatorial class L_k representing layer k of the network is expressed in terms of the product of the generating functions of the combinatorial classes X_j that reads:

$$L_k(z) = \prod_{j=1}^{g_k} X_j(z) \tag{14}$$

Thus, the partition function of the network can be finally written in a generic form as:

$$Z = \sum_{n} M(n)e^{-\beta n} = M(z)|_{z=e^{-\beta}} = \prod_{k=1}^{L}\prod_{j=1}^{g_k} X_j(e^{-\beta}) \tag{15}$$

The previous expression particularized for a fully connected feed-forward network of L layers gives rise to (see appendix B):

$$Z = e^{-\mu g_1 \beta}\left[\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right]^{g_1} \prod_{i=2}^{L} e^{-\mu g_i \beta}\left[\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right]^{g_i} \tag{16}$$

It is important to note that in a fully connected feed-forward architecture a neuron in any layer of the network is connected to all the nodes/neurons in the previous layer. Signal flow through the network progresses in a forward direction, from left to right and on a layer by layer basis. The coefficients g_i represents the number of neurons associated to layer i ($0 \leq i \leq L$), where the coefficient g_0 is representing the number of inputs to the network (remember that the bias inputs are not considered for simplification purposes). The coefficients λ that appear in the previous equation are those derived in section 3.1.1 but assuming now the form:

$$\lambda^1_{m,p,g_0} = \sum_{n=g_0}^{\delta_1}\sum_{l=0}^{n}(-1)^l\binom{g_0}{l}\binom{n-mpl-1}{g_0-1} \tag{17}$$

$$\lambda^2_{m,p,g_0} = \sum_{n=\delta_1+1}^{g_0 mp}\sum_{l=0}^{n}(-1)^l\binom{g_0}{l}\binom{n-mpl-1}{g_0-1} \tag{18}$$

$$\lambda^1_{m,g_{i-1}} = \sum_{n=g_{i-1}}^{\delta_i}\sum_{l=0}^{n}(-1)^l\binom{g_{i-1}}{l}\binom{n-2ml-1}{g_{i-1}-1} \tag{19}$$

$$\lambda^2_{m,g_{i-1}} = \sum_{n=\delta_i+1}^{2mg_{i-1}}\sum_{l=0}^{n}(-1)^l\binom{g_{i-1}}{l}\binom{n-2ml-1}{g_{i-1}-1} \tag{20}$$

The coefficients δ_1 and δ_i (where $2 \leq i \leq L$) that appear in the equations of the coefficients λ correspond to the thresholds that model the squashing effect of the activation function that read respectively:

$$\delta_1 = g_0 + \frac{g_0(mp-1) - (g_0(mp-1)) \bmod 2}{2} \tag{21}$$

$$\delta_i = g_{i-1} + \frac{g_i(2m-1) - (g_{i-1}(2m-1)) \bmod 2}{2} \tag{22}$$

Moreover, equation (16) can also be written in a recursive form (see appendix B) as:

$$Z_k = Z_{k-1} e^{-\mu g_k \beta} \left(\lambda^1_{m,g_{k-1}} + e^{-\Delta\beta} \lambda^2_{m,g_{k-1}} \right)^{g_k} \tag{23}$$

Where the initial step for the recursion Z_1 reads:

$$Z_1 = e^{-\mu g_1 \beta} \left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta} \lambda^2_{m,p,g_0} \right)^{g_1} \tag{24}$$

Similarly, using the equations (16) and (23) the expression of the partition function of a feedforward fully connected deep neural network of $L-1$ layers with sigmoidal units and an output layer L composed of linear units can be derived (see appendix B) giving rise to:

$$\begin{aligned} Z &= e^{-\mu g_1 \beta} \left[\lambda^1_{m,p,g_0} + e^{-\Delta\beta} \lambda^2_{m,p,g_0} \right]^{g_1} \\ &\times \prod_{i=2}^{L-1} e^{-\mu g_i \beta} \left[\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta} \lambda^2_{m,g_{i-1}} e^{-(\mu+\Delta)\beta} \right]^{g_i} e^{-\beta g_L g_{L-1}(m+\frac{1}{2})} \left(\frac{sinh(\beta m)}{sinh(\frac{\beta}{2})} \right)^{g_L g_{L-1}} \end{aligned} \tag{25}$$

Finally, it is important to note that according to the modelization, and specially, within the context of feedforward fully-connected networks the number of degrees of freedom of any artificial neuron is described by the number of inputs to the neuron, the total number of adaptive weights, and its state (i.e., the output of the artificial neuron). The number of inputs to a neuron strictly depends on the total number of neurons of the preceding layer where the neuron is situated (or to the number of inputs to the network if the neuron is situated in the first hidden layer). Thus, the degrees of freedom of a neuron are linked to its position within the architecture leading to the fact that the topology of the network (i.e., the number of layers of the network, and the number of neurons per layer in this case) has a strong influence in the resulting degrees of freedom of the entire network.

3.3. The Thermodynamic Potentials

The equation (16) corresponding to the partition function of a deep network of L layers composed of sigmoidal units is used hereafter to calculate the free energy, the entropy, and the internal energy of a deep learning feed-forward network. It is important to remember the principal interest in the characterization from a thermodynamics point of view of these machine learning models. The appendix B also provides the expressions derived from equation (25) of the thermodynamic potentials of a fully connected deep neural network of $L-1$ layers with sigmoidal units and one output layer composed of linear units.

3.3.1. Free Energy

The free energy for a deep learning feedforward network composed entirely of sigmoidal units reads:

$$F = \mu \sum_{i=1}^{L} g_i - \frac{g_1}{\beta} \log \left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta} \lambda^2_{m,p,g_0} \right) - \frac{1}{\beta} \sum_{i=2}^{L} g_i \log \left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta} \lambda^2_{m,g_{i-1}} \right) \tag{26}$$

The previous equation can also be rewritten in a recursive form leading to:

$$F_k = F_{k-1} + \mu g_k - \frac{1}{\beta} g_k \log\left(\lambda^1_{m,g_{k-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{k-1}}\right) \tag{27}$$

Where the initial step for the recursion F_1 reads:

$$F_1 = \mu g_1 - \frac{1}{\beta} g_1 \log\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right) \tag{28}$$

3.3.2. Entropy

The entropy for a fully connected deep learning feedforward network of L layers composed of sigmoidal units reads:

$$\begin{aligned} S &= g_1 \log\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right) + \sum_{i=2}^{L} g_i \log\left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right) + \\ &+ \beta\Delta e^{-\Delta\beta}\frac{g_1\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} + \beta\Delta e^{-\Delta\beta}\sum_{i=2}^{L}\frac{g_i\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}} \end{aligned} \tag{29}$$

The entropy equation rewritten in a recursive form reads:

$$S_k = S_{k-1} + g_k \log\left(\lambda^1_{m,g_{k-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{k-1}}\right) + \beta\Delta e^{-\Delta\beta}\frac{g_k\lambda^2_{m,g_{k-1}}}{\lambda^1_{m,g_{k-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{k-1}}} \tag{30}$$

Where the initial step for the recursion S_1 reads:

$$S_1 = g_1 \log\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right) + \beta\Delta e^{-\Delta\beta}\frac{g_1\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} \tag{31}$$

3.3.3. Internal Energy and its fluctuations

The internal energy for a fully connected deep learning feedforward network of L layers composed of sigmoidal units reads:

$$U = \mu\sum_{i=1}^{L} g_i + \Delta e^{-\Delta\beta}\frac{g_1\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} + \Delta e^{-\Delta\beta}\sum_{i=2}^{L}\frac{g_i\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}} \tag{32}$$

Similarly, the internal energy can also be written in a recursive form of the equation of the internal energy reads:

$$U_k = U_{k-1} + \mu g_k + \Delta e^{-\Delta\beta}\frac{g_k\lambda^2_{m,g_{k-1}}}{\lambda^1_{m,g_{k-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{k-1}}} \tag{33}$$

and the initial step for the recursion U_1 reads:

$$U_1 = g_1\frac{\lambda^1_{m,p,g_0} + \Delta\lambda^2_{m,p,g_0}e^{-\Delta\beta}}{\lambda^1_{m,p,g_0} + \lambda^2_{m,p,g_0}e^{-\Delta\beta}} \tag{34}$$

Finally, the fluctuations of the internal energy reads:

$$
\begin{aligned}
(\Delta U)^2 &= \Delta^2 e^{-\Delta\beta} g_1 \frac{\lambda^1_{m,p,g_0}\lambda^2_{m,p,g_0}}{\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right)^2} + \\
&+ \Delta^2 e^{-\Delta\beta} \sum_{i=2}^{L} g_i \frac{\lambda^1_{m,g_{i-1}}\lambda^2_{m,g_{i-1}}}{\left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right)^2}
\end{aligned} \tag{35}
$$

3.4. The Specific Heat

The quantity of heat which, when added to a body raises its temperature by one unit of temperature (e.g., one degree) is called in thermodynamics the specific heat (denoted as C_v hereafter) [34]. This parameter is calculated here because it provides conditions for equilibrium and stability of the states of a thermally isolated physical system as a result of the law of increase of entropy (i.e., the second law of thermodynamics), and it expresses the idea how much heat a body can absorb.

This parameter will play an important role in understanding the thermodynamic efficiency with which information can be processed by these models. It is important to remember that during the learning phase of deep learning machines their adaptive weights are gradually fixed according to the learning task at hand. Thus, according to the theoretical model the training phase of these models can be interpreted as changing the probability of occurrence of the microstates of the system. From a thermodynamics point of view this fact can be seen as energy that is brought into the system while no work is delivered , that is, the external parameters of the hamiltonian are unchanged (i.e., the topology does not change),and the energy levels of the system remain identical. The average energy of the system changes due to the changes in the probabilities of ocurrence of the microstates and the variation of internal energy is identified with the inifinitesimal heat absorbed [31].

The specific heat can be obtained from the derivative of the Entropy (see equation (29)) with respect to the inverse temperature β as follows:

$$
C_v = -\beta \frac{\delta S}{\delta \beta} = \sum_{i=1}^{4} C_{v_i} \tag{36}
$$

whereas the terms C_{v_i} $(1 \leq i \leq 4)$ read respectively:

$$
C_{v_1} = \Delta\beta e^{-\Delta\beta} g_1 \frac{\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} \tag{37}
$$

$$
C_{v_2} = \Delta\beta e^{-\Delta\beta} \sum_{i=2}^{L} g_i \frac{\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}} \tag{38}
$$

$$
\begin{aligned}
C_{v_3} &= -\Delta\beta(1-\Delta\beta)e^{-\Delta\beta} g_1 \left(\frac{\lambda^1_{m,p,g_0}\lambda^2_{m,p,g_0}}{\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right)^2} + \left(\frac{\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} \right)^2 \right) - \\
&- g_1 \left(\frac{\Delta\beta e^{-\Delta\beta}\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} \right)^2
\end{aligned} \tag{39}
$$

$$
\begin{aligned}
C_{v_4} &= \Delta\beta(1-\Delta\beta)e^{-\Delta\beta} \sum_{i=2}^{L} g_i \left(\frac{\lambda^1_{m,g_i}\lambda^2_{m,g_i}}{\left(\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}\right)^2} + \left(\frac{\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}} \right)^2 \right) - \\
&- \sum_{i=2}^{L} g_i \left(\frac{\Delta\beta e^{-\Delta\beta}\lambda^2_{m,g_i}}{\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}} \right)^2
\end{aligned} \tag{40}
$$

3.5. Quenched Averages

The expression for the free energy (26), the entropy (29), and the internal energy (32) are randm variables as a result of the disorder introduced by the structural parameters m and p. In order to obtain, the typical behavior of the physical system, the expectation of these random vaiables (i.e., quenched averages) must be calculated:

$$\begin{aligned}\langle \Theta(\beta, m, p) \rangle &= \sum_m \sum_p P(m,p)\Theta(\beta, m, p, g_0, g_1, g_2, \ldots, g_L) \\ &= \sum_m \sum_k P(m)P(p)\Theta(\beta, m, p, g_0, g_1, g_2, \ldots, g_L) \end{aligned} \tag{41}$$

In expression (41) Θ represents any of the thermodynamic potentials S,F, an U whereas $P(p)$ and $P(m)$ are the probability distributions of the random variables describing the inputs to the networks and their adaptive weights respectively. It is assumed that the random variables p and m are independent. Furthermore, limit Gaussian distributions are assumed for $P(p)$ and $P(m)$. These assumptions permit to capture the essence of the processes being modeled allowing at the same time a considerable simplification of the calculation. Specifically, Dirac delta functions (see expression (42)) are used simulating the concentration of the probability mass around the mean of the Gaussian distribution:

$$\begin{aligned} P(m,p) &= \\ &= P(m)P(p) \\ &= \frac{1}{\sqrt{2\pi}\sigma_m} e^{-\frac{(m-\mu_m)^2}{2\sigma_m^2}} \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{(p-\mu_p)^2}{2\sigma_p^2}} \\ &\cong \delta(m-\mu_m)\delta(p-\mu_p) \end{aligned} \tag{42}$$

Hence, the expectations of the Entropy, Free Enregy and Internal Energy read exactly as in expressions (26),(29),and (32) simply substituting the variables m, and p by μ_m, and μ_p respectively.

4. Theoretical Results

The first goal of this section is to verify that the thermodynamic formalism derived in section 3 is able to capture the intrinsic differences that may exists in these models as a result of the topology and/or the structural parameters that characterize them (e.g., total number of nodes, total number of adaptive weights, the input space dimension, or the number of hidden layers). Accordingly, a preliminary set of network architectures is used to assess the validity of the theoretical model.

The second goal is to analyze the influence of the neural network topology,that is, the influence of their structural parameters such as the total number of adaptive weights, the total number of nodes, the depth of the networks , as well as their inter-relationships in terms of the thermodynamic formalism derived in section 3.To this end, three scenarios imposing different restrictions on the structural parameters of these models are used. Furthermore, they quantitatively sample the whole range of combinations that may arise when considering the topology and the structural parameters of these models. It is important to note that different network topologies may correspond to an identical value of one of the aforementioned structural parameters.

The final goal is to explore from a theoretical perspective the differences between shallow networks (i.e., networks with one hidden layer) and deep networks (i.e., networks with more than one hidden layer) aimed at understanding the relationship of those differences with their learning and generalization capabilities.

Finally, it is important to note that for simplification purposes it is also assummed hereafter that the discrete random variables m and p that model the number of energy levels associated to the adaptive weights and the inputs to the networks are identical, that is, $\mu_p = \mu_m$.

4.1. The Thermodynamic Formalism

To assess the validity of the theoretical model, a preliminary set of nine architectures were conveniently selected to examine their macroscopic thermodynamic behavior. Specifically, shallow networks, and deep networks with two and three hidden layers are considered in the following for comparison purposes. Networks with more than three hidden layers are only considered in the next sections for assessing the influence of the depth of the networks. Furthermore, It is also assumed that all the units in these networks are sigmoidal excepting the output layer that is composed of a single neuron with a linear activation function. The qualifier "multiplex" is used hereafter to denote multilayer artificial neural networks since the concept of multilayer network is broader than the traditional use that it has been done in the machine learning field [13]. Furthermore, the notation used for describing the parameters of the multiplex networks is similar to that used in [13]. Specifically, the notation g_k, where $0 \leq k \leq L$ is used to denote the number of neurons associated to layer k, where the value for $k = 0$ indicates the number of inputs to the network, and the value $k = L$ the number of output units. Furthermore, shallow networks are denoted as M_1 (multiplex networks with one hidden layer), deep networks of two hidden layers as M_2 , M_3 for deep networks of 3 hidden layers, and so forth (M_L for multiplex networks with L hidden layers).

Having said this, it is important to note that from a thermodynamics point of view a measure of the disorder of a physical system is the number of accesible states (or the size of the accesible volume of the phase space). The number of accesible states is related to both the number of particles (artificial neurons in this case) and to the number of degrees of freedom of each particle, although in the general case, the particles may have other degrees of freedom since they may be subjected to a potential energy [31,34]. Furthermore, it also important to remember the two distinct but related interpretations of the concept of entropy, the first based on degree of disorder, and the second on information content [11], especially taking into consideration that entropy is a macroscopic quantity defined from the number of accesible states of the physical system under consideration. Thus, both interpretations of the entropy coexists in this particular case since according to the theoretical model the accesible states of the physical system under study (i.e, fully connected feedforward neural networks) encode information. In other words, as higher is the entropy of a network as higher will be its resulting degree of disorder but at the same time its information encoding capacity of the network (i.e., the storage capacity of information of the network).

Moreover, two structural parameters are commonly used to compare the complexity of a network, namely, the total number of nodes, and the total number of adaptive weights of the network. Unfortunately, fixing any of these parameters (or even both) may result in a whole variety of network architectures that display radical differences in terms of their topology (e.g., the total number of layers, and/or the number of units per layer). According to the theoretical model of section 3, the degrees of freedom of an artificial neuron (i.e., the particles in the physical system under study) are described by its total number of adaptive weights, by the set of values of its inputs, and by the cardinality of the set of energy states used to describe both the adaptive weights and the inputs to the neuron, all of them defining its state (active or not active) which is also a degree of freedom when the artificial neuron is part of an artificial neural network. Thus, the degree of disorder of a deep learning fully connected feedforward network is a function of the total number of neurons and of the number of degrees of freedom of the artficial neurons comprising its structure. In other words, the number of accesible states of a Deep Learning fully-connected feedforward machine is intrinsically linked to the topology of the network but also to their total number of neurons N.

The panels of figures (3) and (4) show the evolution of the thermodynamic potentials (i.e., entropy, free energy, and internal energy), and also the specific heat for the set of network architectures

selected as a function of the average number of energy levels (or states) used to represent the adaptive weights of the networks (i.e., the parameter μ_m).The whole set of networks possess an identical number of adaptive weights (i.e., an identical effective complexity), and some of them (such as the pairs of architectures 12x300x1 and 301x11x1) even have an identical number of units and adaptive weights (this is also true for the pair of architectures 25x149x1 and 150x24x1). The qualifier structural complexity of networks is used hereafter to denote artificial neural networks with and identical number of units N and adaptive weights W. The expectations of the Entropy, Free Energy, Internal Energy, and the specific heat of a shallow network of 1308x1 units (M_1) are plotted against four deep networks (M_2) with architectures 12x300x1, 25x149x1,150x24x1, and 301x11x1 in panel (3). Similarly, the thermodynamic potentials, and the specific heat of the shallow network 1308x1 are plotted in panel (4) together with those of four deep networks of three hidden layers (M_3) with architectures 162x21x9x1, 24x59x41x1,144x11x171x1, and 30x32x88x1 respectively for comparison purposes. It is important to remember that it is assummed that the input space to the networks is a two-dimensional space (i.e., $g_0 = 2$), and also that the activation function of the output unit of the networks is linear.

The selection criterion for deep network architectures was principally based on the different possibilities (or logical schemes) that arise when considering the total number of units associated to each layer, namely, $g_1 > g_2$ or $g_1 < g_2$ for deep networks of two hidden layers (i.e., multiplex networks M_2), and the six possibilities for multiplex networks of three hidden layers (M_3). For deep networks of three hidden layers (i.e., multiplex networks M_3) only those logical schemes typically used in the machine learning literature were considered. In particular, the logical scheme $l_1 \equiv g_1 > g_2 < g_3$ has found applications in data encryption (compressive autoencoder architecture), whereas the logical schemes $l_3 \equiv g_1 < g_2 > g_3$ (expansive autoencoder architecture) and $l_4 \equiv g_1 > g_2 > g_3$ correspond to the arquetype of deep learning applications. Finally, the logical scheme $l_2 \equiv g_1 < g_2 < g_3$ (expansive architecture scheme) was also included simply to investigate its behavior compared to the rest of selected schemes.

In both figures, the most important characteristic of the behavior of the entropy is that it grows (as expected) with the parameter μ_m. It is important to note that the number of microstates of the artificial neural network model increases exponentially with the average number of energy levels (or states) associated to the adaptive weights of the networks μ_m. Independently of the value of the parameter μ_m considered the entropy of the shallow network is higher with respect to any of the deep network architectures. This fact is of particular interest taking into consideration that the effective complexity of both shallow and deep networks is identical, thereby pointing to the existence of intrinsic differences between both types of artificial neural network models. Particularly, these results appear to suggest that under the conditions defined (i.e., identical effective complexity) shallow networks possess a higher storage capacity of information compared to deep networks, but at the same time they appear to present a higher degree of disorder compared to deep networks. In contrast, the free energy decreases to the extent the parameter μ_m increases. Although as occured with the entropy the values reached by the free energy of the shallow network are always higher compared to those values reached by deep networks.

Finally, the internal energy is practically constant and thus it appears to be independent of the statistics associated to the random variable that model the energy states associated to the adaptive weights of the networks (i.e., the random variable m). It is important to note that the effect of the training algorithm used during the learning phase is simply to alter the probability of occurrence of the energy states of the network according to the particularities imposed by the data set to be learnt (i.e., its unknown probability distribution). In other words, the influence of the training set and the learning algorithm is implicitly embedded within the statistics of the random variables m and p of the theoretical model. Thus, taking into consideration the fact that the internal energy is constant with μ_m it is a plausible to state that this thermodynamic potential is principally linked to the topology of the networks (besides to the total number of units comprising the networks). Again it highest value is reached again for the shallow network. Surprisingly, when comparing the set of deep

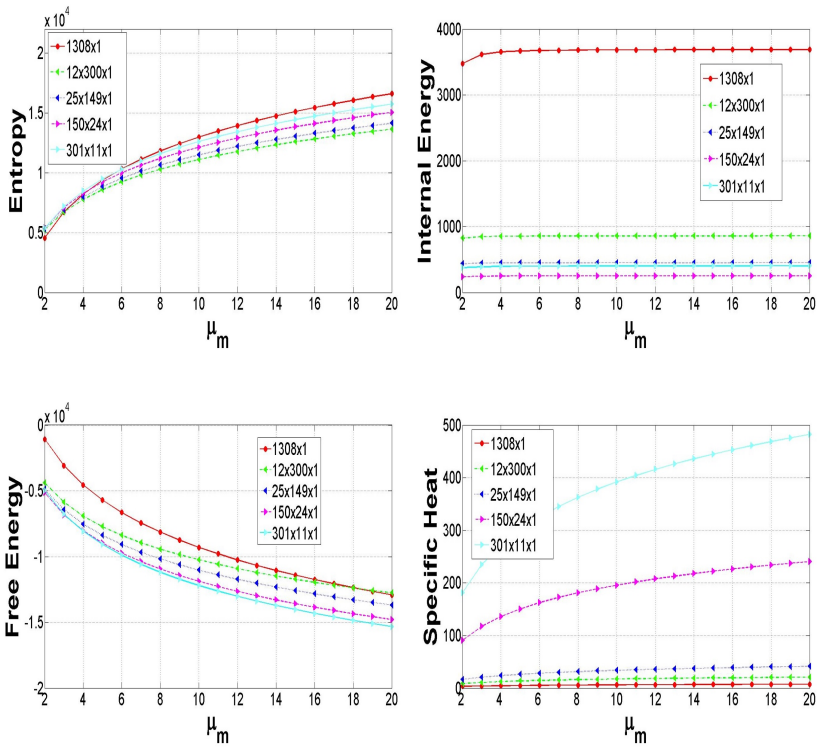

Figure 3. Graphical illustration of the evolution of the thermodynamic potentials and the specific heat in terms of the average number of energy levels used to represent the adaptive weights and the inputs to the networks. The set of architectures represented posses and identical effective complexity, and is composed of a shallow network of 1309 units, and two groups of deep networks with two hidden layers with an identical structural complexity (175 units the first group and 313 units the second). Networks with a hierarchical structure are those exhibiting higher entropy values, and thus larger storage capacities. Particularly, the network attaining the highest entropy values is the shallow network but at the same time is the network possessing the highest number of units. Hierarchical networks (excepting the shallow network) are those exhibiting the lowest internal and free energy values but at the same time the larger specific heat values. Of particular interest is the fact that the shallow network is the network attaining the lowest values for the specific heat in spite of the fact of being the network with the largest neuron numbers. Expansive architecture networks such as 12x300x1 (N=313) and 25x149x1 (N=175) are those exhibiting the worst storage capacities and thus the lowest entropy values. Surprisingly, the former attains the lowest entropy values in spite of the fact of possessing a larger number of neurons, thereby evidencing the profound influence of the topology in the thermodynamic properties of the networks.Each network appears to have its own thermodynamic signature even in those cases where the total number of adaptive weights and/or the total number of neurons is identical.

network architectures selected, networks possessing lower internal energy values are those obeying a hierarchical structure, i.e, the logical scheme $g_1 > g_2$ for multiplex networks M_2, or the logical scheme $g_1 > g_2 > g_3$ for multiplex networks M_3. This fact is of particular interest because this kind of hierarchical structures are the most popular in deep learning applications due to their learning and generalization performance.

However, the most important conclusion that can be gleaned from the graphs is that each architecture appears to possess its own thermodynamic signature. In other words, the numerical values obtained when evaluating the expressions of the thermodynamic potentials (i.e., entropy, free energy, and internal energy), and also those concerning the specific heat appear to be specific to each of the architectures considered. This fact is of particular interest taking into consideration that the theoretical formalism derived is even sensitive to capture existing topological differences in networks with an identical structural complexity (i.e., networks with an identical number of units and adaptive weights) such as the pairs of networks 12x300x1 and 301x11x1, or the networks 25x149x1 and 150x24x1 respectively, thereby suggesting that the observed differences in the thermodynamic potentials are strongly linked to the topology of the networks.

Moreover, it is important to remember that the problem of how well an artificial neural network can infer an unknown classification rule from input/output examples defines the concept of generalization error in the Machine Learning field [11,30], and at the same time, learning input-output relations also defines the problem of storage capacity [28]. The Vapnik-Chevornenkis (VC hereafter) dimension is the quantity that relate both problems, since it provides a measure of the capacity and complexity of the space of functions that can be learned by any statistical classification algorithm.Taking these considerations into account, the behavior of the entropy appears to suggest that those networks leading to larger entropy values have the potential to provide better learning and generalization performance since as higher is the entropy of those networks as higher will be the complexity of the space of functions that they are able to represent, and thus its VC dimension. Indeed, complex learning functions need (in average) higher amounts of information for being described, that is, networks with higher entropy values. Nevertheless, it is important to emphasize that according to the model (remember that both interpretations of entropy coexists) networks with higher entropy values present a higher degree of disorder, but at the same time a higher information storing capacity, and a priori it is not possible to assess without experimentation how the generalization error is affected by the degree of disorder of the networks. These preliminary results were tested with a whole variety of network architectures playing with the topology and the structural parameters of the networks obtaining identical results, that is, each network possess its own thermodynamic signature according to the theoretical model presented in section 3. Thus, changes in the structural parameters of these models lead to changes in their thermodynamic signature, however, the whole thing is that the observed changes are indicating the existence of variations in the number of accesible states and therefore in the structure of the phase space of the networks that are interesting to analyze and understand from a theoretical perspective. More specifically, understanding the observed differences in entropy between shallow and deep networks is of capital importance to provide insights in the current debate about their learning and generalization capabilities.

Finally, It is important to remember that one of the goals of the present study is to better understand the process of learning and generalization of deep-learning machines something that is strongly linked to the role (and structure) of the phase space of the networks as well as to the Vapnik-Chervonenkis dimension [11,30,68], especially taking into account that the phase space and the VC dimension may play independent roles in the process of generalization [47]. However, it is important to emphasizes that the theoretical model derived is an equilibrium model, that is, the analysis is focused on the typical behavior of the system or in the presence of small fluctuations around its typical behavior, in other words, the emphasis is put on the analysis of the typical learning and generalization capabilities that may be expected from these Machine learning models.

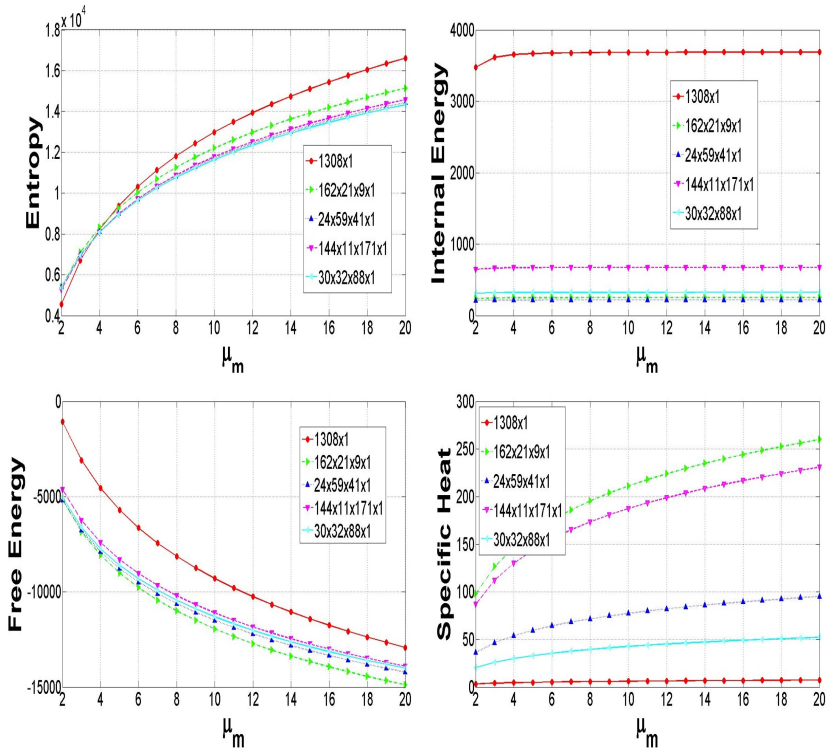

Figure 4. Graphical illustration of the evolution of the thermodynamic potentials and the specific heat in terms of the average number of energy levels used to represent the adaptive weights and the inputs to the networks for an input space dimension $g_0 = 2$. The set of network architectures represented posses and identical effective complexity but different topological schemes, and is composed of a shallow network of 1309 units, and four deep networks of three hidden layers each. Hierarchical networks are again those networks exhibiting higher entropy values, and thus larger storage capacities. Particularly, the network with the highest entropy is the shallow network but at the same time is the network possessing the highest number of units. Excluding the shallow network, the hierarchical network 162x21x9x1 exhibits the lowest internal and free energy values, and the larger specific heat values compared to the rest of architectures. After hierarchical networks the autoencoder architectures 24x59x41x1 (expansive scheme) and 144x11x171x1 (compressive scheme) are those attaining higher entropy values, and thus higher storage capacities. The compressive scheme also presents the highest internal energy value after the shallow network although both networks possess a larger number of neurons. The network with an expansive topology 30x32x88x1 is the network with the worst storage capacity, attaining the lowest entropy and specific heat values. In summary, the number of neurons is a necessary but not sufficient condition for networks to attain higher entropy values, and thus higher storage capacities since the thermodynamic signature of networks is strongly linked to its topology.

In the following, to shed light on this issue a more detailed analysis is presented using three illustrative scenarios that impose different restrictions on the structural parameters of the networks leading to important differences in the resulting phase space of the networks according to the theoretical model presented.

4.2. Controling the Structural Complexity: Case Study I

This section is aimed at studying the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers), and the effect of the input space dimension to the networks when controling the structural complexity of the networks. Due to the strong restrictions imposed to the structural parameter of the networks this scenario is particularly suited to analyze the phase space of both shallow and deep networks. Specially, taking into consideration that under these circunstances the theoretical capacity for encoding information of the networks is exactly the same.

The effective complexity of a neural network is expressed by the total number of adaptive weights.Thus, using the superscript d for deep network and s for shallow networks, denoting g_k^d as the number of units associated to layer k such as $0 \leq k \leq L$, whereas g_0^d is representing the number of inputs to the network, and assuming that the networks possess only one neuron in the output layer (i.e., $g_L^d = 1$), comparing a deep network of L layers with a shallow network with g_1^s hidden layer units under identical effective complexity is equivalent to solve the following diophantine equation:

$$g_{L-1}^d + \sum_{k=0}^{L-2} g_k^d g_{k+1}^d = (1 + g_0^s) g_1^s \tag{43}$$

From the set of solutions of equation (43) to ensure that the resulting networks possess and identical structural complexity only those solutions that satisfy the equation:

$$\sum_{i=1}^{L-1} g_i^d = g_1^s \tag{44}$$

can be used. The equation (43) with the restriction imposed by (44) was numerically solved for different range of values in the space of parameters, for later picking up some of the resulting values appropriately for the analysis and the illustration of concepts presented hereafter. The final goal is to understand from a theoretical level the intrinsic properties (i.e., differences and commonalities) that characterize deep networks when compared to shallow networks.

4.2.1. Network Architectures: A Preliminary Analysis

A total set of twelve network architectures were selected for the study ranging from one up to three hidden layers and following different topological schemes.

Table 1 shows the set of network architectures selected together with their corresponding thermodynamic signature. The values of the entropy S, free energy F, internal energy U and its fluctuations ΔU are shown together with the specific heat C_v, the total number of adaptive weights W, and the total number of units N assuming a discretization level of the adaptive weights equal to 20 (i.e., $\mu_m = 20$), and an eight dimensional input space to the networks (i.e., $g_0 = 8$), excepting for the two hidden layer network architectures (i.e, multiplex networks M_{2a}, and M_{2b}) where the set of values shown correspond with a two dimensional input space (i.e., $g_0 = 2$).

There are four groups of network each of them with an identical structural complexity. More specifically, the first group is composed by multiplex networks M_{1a}, and M_{3a}, that is, a shallow network 21x1 and two three hidden layer networks 12x6x3x1 and 11x7x3x1. The second group (mutiplex networks M_{2a}) with architectures 12x300x1 and 301x11x1 following opposed topological schemes (i.e, hierarchical vs. expansive). The third group is composed of two hidden layer networks (architectures 150x24x1 and 25x149x1 also following opposed topological schemes but with a lower number of units ($N = 175$) compared to the second group. Finally the fourth group (multiplex

networks M_{3b}) composed of four networks following the possible topological schemes for three hidden layer networks with architectures 108x50x108x1,16x54x196x1, 16x211x39x1 and 180x47x39x1.

The shallow network 1308x1 (M_{1b}) does not belong to any of the aforementioned groups and it was included for comparison purposes in the following sections because it has an identical effective complexity with respect to the group of networks M_{3b}.

Table 1. The set of network architectures selected for the study of the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers), and the effect of the input space dimension to the networks when controling the structural complexity of the networks, that is, the total number of units N and the total number of adaptive weights W. The table shows the thermodynamic potentials associated to each of the networks considered (i.e., entropy S, free energy F, internal energy and its fluctuations $U \pm \Delta U$) , the specific heat C_v, the total number of adaptive weights W for the input space dimension considered ($g_0 = 8 for the groups M_{1a}, M_{3a}$, and M_{3b}, and $g_0 = 2$ for the rest of groups), and the total number of units N.

Multiplex M_{1a}	**S**	**F**	**U** $\pm\Delta U$	C_v	**W**	**N**
21x1	1021.54	-962.33	59.21 ± 9.72	1.64×10^{21}	189	22
Multiplex M_{1b}						
1308x1	63627.6	-59939.2	3688.4 ± 23.44	1.02×10^{23}	11772	1309
Multiplex M_{2a}						
12x300x1	13634.7	-12775	859.7 ± 2.44	2.44×10^{21}	3924	313
301x11x1	15730.2	-15236.7	403.5 ± 11.26	2.18×10^{482}	3924	313
Multiplex M_{2b}						
150x24x1	15045.5	-14792.2	253.3 ± 7.98	5.82×10^{240}	3924	175
25x149x1	14138.7	-13687.5	451.2 ± 3.37	1.99×10^{41}	3924	175
Multiplex M_{3a}						
12x6x3x1	903.46	-872.75	30.71 ± 9.67	0.95×10^{21}	189	22
11x7x3x1	885	-854.26	30.75 ± 9.73	0.86×10^{21}	189	22
Multiplex M_{3b}						
108x50x108x1	45042	-44542.1	499.9 ± 8.19	6.26×10^{173}	11772	267
16x54x196x1	43114.4	-42475.5	638.9 ± 5.48	6.55×10^{87}	11772	267
16x211x39x1	43530.4	-43140	390.4 ± 9.75	5.03×10^{338}	11772	267
180x47x39x1	46550	-46159.9	390.1 ± 9.8	1.36×10^{289}	11772	267

Having said this, it is important to note that according to the model presented in section 3, each of the artificial network architectures considered in this section (i.e., those of table 1) correspond to the equilibrium states of a phyical system. For example, the network architectures of the first group (i.e., M_{1a} and M_{3a}) would represent a subset of the equilibrium states of a physical system composed of 22 artificial neurons with 8 inputs and with a total number of 189 adaptive weights. Similarly, the two networks of the second group (i.e., M_{2a}) would represent a subset of the equilibrium states of a physical system composed of 313 units with two inputs, and with a total number of 3924 adaptive weights, and so forth.

The most important observation that can be gleaned from the inspection of the table is the substantial differences that exists on the entropy values attained at the equilibrium points of the four theoretical systems considered (i.e., physical systems whose equilbrium states are artificial neural networks with an identical structural complexity). This fact is of particular interest taking into consideration that from a theoretical point of view networks subject to these strong restrictions (i.e., identical structural complexity) should have an identical capacity for storing information. Thus, the observed differences in the entropy values of the equilibrium states represented by the networks can only be explained in terms of the existing differences in their topology (the external field) which is affecting not only the structure of the phase space of the system, but also its thermal properties. Independently of this fact, it is also important to remember that larger entropy values indicate a higher

degree of disorder (more probable macroscopic states) but at the same time networks with a higher capacity for storing information.

Moreover, comparing the values of the entropy attained by the different set of network architectures it can be deduced that hierarchical networks (including shallow networks) are those attaining the highest entropy values, thus indicating that this topological scheme leads to networks with the highest information-storage capacities but also to larger heat capacities (excluding shallow networks). It is important to note that in this scenario the qualifiers heat capacity or specific heat can be used indistinctly since the total number of units is identical. In contrast, the expansive topological scheme (e.g., network architectures 12x300x1, 25x149x1, and 16x54x196x1) leads to networks with the lowest information-storage capacities. Of particular interest is the fact that networks that store the largest amount of energy belong to this topological scheme together with shallow networks, and at the same time are those topological schemes presenting the smallest heat capacities.

Moreover, networks presenting a higher degree of disorder are those presenting larger energy fluctuations. In contrast, network architectures belonging to the expansive topological scheme are those presenting the smallest energy fluctuations.

Bearing in mind the intrinsic link between the entropy and the number of accesible states, that is, with the inherent structure of the phase space, in the next section the emphasis is put in understanding the phase space of these networks under the strong restrictions imposed by this scenario. To this end the number of accesible states associated to the equilibrium states represented by these networks are calculated for comparison purposes.

4.2.2. The Generating Function of Energies

Equation (45) correspond to the generating function of energies of a generic fully connected feed-forward network of L layers under the assumption that all the units of the network are sigmoidal (or hyperbolic tangent). This function permits to calculate the number of microscopic states accesible to any given fully connected feed-forward neural network of L layers according to the model presented in section 3.

$$M(z) = \left(\lambda^1_{m,p,g_0} z^{\mu} + \lambda^2_{m,p,g_0} z^{\mu+\Delta}\right) \prod_{i=2}^{L} \left(\lambda^1_{m,g_{i-1}} z^{\mu} + \lambda^2_{m,g_{i-1}} z^{\mu+\Delta}\right)^{g_i} \tag{45}$$

The coefficients of this polynomial function express the total number of microstates (or microscopic configurations) leading to the energy values expressed by the power of the complex variable z, thus the total number of accessible microstates for a concrete feed-forward fully connected neural network is obtained evaluating the aforementioned equation for $z = 1$ particularized for the topological charateristics of the network such as the input-space dimension g_0, the number of layers L, and the number of units associated to each layer, that is, g_i, where $1 \leq i \leq L$.

The data presented in table 1 was generated assuming networks composed of nonlinear units excepting the output layer, however, without a loss of generality the expressions of the generating functions that are shown hereafter were obtained using equation (45) so as to get finite polynomial expansions.

$$\begin{aligned} M_{21x1}(z) = {} & 1.49 \times 10^{464} z^{22} + 3.29 \times 10^{465} z^{24} + \cdots \\ & \cdots + 9.48 \times 10^{469} z^{42} + 1.03 \times 10^{470} z^{44} + 9.43 \times 10^{469} z^{46} + \cdots \\ & \cdots + 3.15 \times 10^{465} z^{64} + 1.43 \times 10^{464} z^{66} \end{aligned} \tag{46}$$

Equations (46), (47), and (48) show respectively the generating functions of energy of one of the groups of architectures of table 1 under identical structural complexity for $g_0 = 8$ (i.e., the input space

dimension to the networks). Specifically, one shallow network with architecture 21x1, and two deep networks of three hidden layers with architectures 12x6x3x1 and 11x7x3x1 respectively.

$$\begin{aligned} M_{12x6x3x1}(z) = {} & 1.51 \times 10^{392} z^{22} + 3.32 \times 10^{393} z^{24} + \cdots \\ & \cdots + 9.62 \times 10^{397} z^{42} + 1.05 \times 10^{398} z^{44} + 9.59 \times 10^{397} z^{46} + \cdots \\ & + 3.21 \times 10^{393} z^{64} + 1.46 \times 10^{392} z^{66} \end{aligned} \tag{47}$$

The shallow network is the multiplex network that reaches the highest entropy value, with a total number of $M_{21\times1}(1) = 6.13 \times 10^{470}$ accesible microscopic states for the $N = 22$ artificial neurons of this network, a value that is exponentially higher compared to the number of accesible microstates for the deep network architectures $M_{12\times6\times3\times1}(1) = 6.23 \times 10^{398}$, and $M_{11\times7\times3\times1}(1) = 6.22 \times 10^{390}$ respectively. Summarizing, the network with the highest entropy value is the network possessing the highest number of accesible microstates, and the following inequality holds: $M_{21\times1}(1) > M_{12\times6\times3\times1}(1) > M_{11\times7\times3\times1}(1)$.

$$\begin{aligned} M_{11\times7\times3\times1}(z) = {} & 1.48 \times 10^{384} z^{22} + 3.26 \times 10^{385} z^{24} + \cdots \\ & \cdots + 9.59 \times 10^{389} z^{42} + 1.05 \times 10^{390} z^{44} + 9.59 \times 10^{389} z^{46} + \cdots \\ & + 3.26 \times 10^{385} z^{64} + 1.48 \times 10^{384} z^{66} \end{aligned} \tag{48}$$

Following an identical procedure, the number of accesible states of the networks corresponding to the second group of table 1 was also calculated, that is, four networks with different topology but with identical structural complexity was calculated obtaining the values $M(1)_{108\times50\times108\times1}(1) = 2.44 \times 10^{19703}$, $M_{16\times54\times196\times1}(1) = 2.82 \times 10^{18987}$, $M_{16\times211\times39\times1}(1) = 2.82 \times 10^{18987}$, and $M_{180\times47\times39\times1}(1) = 1.03 \times 10^{20266}$.

Having said this, it is important to remember again that the number of accesible states of a physical system composed of N particles is related to both the number of particles and to the number of degrees of freedom of each particle (artificial neurons in this case). Taking into consideration that in this scenario the structural complexity of the networks is identical, the observed differences in the number of accesible states are the result of the differences in the degrees of freedom of the individual neurons comprising the networks as a result of differences in the topology of the networks. Furthermore, it is important to remember the interpretation (or view) of the topology of a network as an especial kind of external field that generates a potential energy that affects the artificial neurons differently depending on their position within the network (i.e., the layer to which the neuron belongs). In other words, as a result of the influence of the topology the degrees of freedom of a neuron vary depending on its position within the network architecture (i.e., the layer to which the neuron belongs), thereby affecting the degrees of freedom of the entire network and thus its total number of accesible states, that is, the intrinsic structure of the phase space of the network. Specifically, the degrees of freedom of a neuron situated in layer L_i are a function of the number of neurons of the precedent layer (i.e., the layer L_{i-1}), or a function of the number of inputs to the network when $L_i = 1$.

Moreover, the total number of units (and their adaptive weights) of the first hidden layer of the network play an important role in the resulting degrees of freedom, since as opposed to any other hidden layer the average number of energy levels used to represent the inputs to the neurons of the first hidden layer are in general bigger compared to those used to represent the inputs to the rest of neurons of the network (i.e., neurons belonging to layer L_i with $i > 1$), that is, two levels in average in case of artificial neurons with sigmoidal transfer functions. In other words, as higher is the number of neurons in the first hidden layer of the network as higher will be its contribution to the number of degrees of freedom of the entire network. Indeed, the importance of early-layer weights (and thus of

the number of neurons of early-layers) for reliable computation has been also recently highlighted in [36].

Independently of those facts, the whole thing is that under identical structural complexity the shallow network M_{21x1} presents a higher degree of disorder compared to the deep networks $M_{12\times6\times3\times1}$ and $M_{11\times7\times3\times1}$ (remember that the larger the number of accesible states, the larger the uncertainity or disorder of the physical system) but at the same time it presents a higher storage capacity compared to deep networks (remember the two distinct but related interpretations of the entropy coexist in this case).

In a similar way, for the second group of networks, the hierarchical network $M_{180\times47\times39\times1}$, and the network with a topology scheme used for data encryption $M_{108\times50\times108\times1}$ are those presenting a higher degree of disorder when compared to the rest but also a higher storage capacity for encoding information (i.e., higher entropy values). In contrast, the networks $M_{16\times54\times196\times1}$, and $M_{16\times211\times39\times1}$ are those presenting a lower level of thermodynamic disorder but at the same time a lower capacity for encoding information. Of particular interest is the fact that in this case both networks present an identical number of accesible states in spite of the fact of having different entropy values ponting to differences in the probabilities of occurrence of the microstates. Specifically, the higher value of the entropy for the network architecture 16x211x39x1 is indicating the existence of a larger number of more probable microstates compared to those of network 16x54x196x1.

Clearly, the difference found in the entropy of both netwoks is evidencing changes in the structure of the phase space of the networks, but surprinsingly, they are simply the result of the influence of the topology. Furthermore, as a result of this fact a subset of miscroscopic configurations becomes more probable in spite of the fact that the cardinality of the set of microscopic configurations is almost identical for both networks.

Taking these considerations into account, it is plausible to state that this fact might be responsible of changes in the resulting energy loss landscape [52,69] that face a machine learning algorithm during the learning phase making the learning task easier or more difficult depending on its characterstics (i.e., intuitively, harder for the network 16x54x196x1 in this case). Thus, one might expect better learning and generalization performance for the network 16x211x39x1 compared to the network 16x54x196x1. In other words, under identical structural complexity, these results appear to suggest the strong influence of the topology in the structure of the phase space of the networks, and thus in their resulting learning and generalization performance independently of the particularities of the functions to be learned. It is important to remember that one of the important questions about Deep Learning machines which to some extent still remains unsolved at the present time is about their learning capabilities (e.g., whether good minima are easier to find in deep rather than in shallow networks).

4.2.3. Infuence of the Structural Parameters

Given the restrictions imposed by this scenario only two structural parameters are considered in this case. Namely, the influence of the input space dimensionality, and the depth of the networks.

With regards the influence of the depth in the thermodynamic potentials figure (8) shows the evolution of the entropy and free energy, whereas figure (9) shows the evolution of the internal energy and the specific heat both for hierarchical deep networks from three up to seven hidden layers for a number of neurons comprised within the integer interval [125..350] with an effective complexity $W = 3924$,and assuming a two dimensional input space to the networks ($g_0 = 2$).

This graph also contains the solutions for networks with an identical structural complexity (i.e, the set of solutions obtained after imposing the restrictions of equation (44)). Specifically, solutions with an identical structural complexity correspond in the graphs to those values of N that has associated two or more points in the vertical axis. However, due to the strong restrictions imposed by this scenario fixing the total number of neurons N it is difficult to find solutions for the entire range of network depth sampled. Independently of those facts, the most important conclusion that can be extracted is that the entropy of the networks appears to decrease (although not monotonically) to the extent the

number of hidden layers increases, indicating changes in their degree of disorder as well as in their capacity. Particularly, the information-storage capacity of networks does not increase monotonically with neuron numbers.

Similarly, both the energy stored by the networks and their heat capacity appear to decrease to the extent the number of hidden layers of the networks increases. There are strong fluctuations of the heat capacity of the networks that appears to increase with both neuron numbers N and the depth of the networks, thus indicating strong variations in the thermal properties of the networks, particularly, its capacity for absorving heat from the point of view of thermodynamics, which in turn is linked to the characteristics of the energetic landscape that is seen by any learning protocol from the point of view of machine learning. In other words, the process of finding good minima during a learning task may be easier or harder depending on that. In contrast, the energy stored by the networks presents certain fluctuations that appear to decrease with both neuron numbers N and the depth of the networks.

The graph of figure (5) represent the evolution of the entropy (top part of the graph) and the internal energy (bottom part of the graph) as a function of the input space dimensionality for the group of networks with a total number of units $N = 22$, that is, two deep networks of three hidden layers with architectures 12x6x3x1 and 11x7x3x1, and one shallow network with architecture 21x1. The range of values used from 2 dimensions up to 100 is wide enough to check the evolution of both the entropy and the internal energy of the networks when the number of inputs (i.e., g_0) is higher enough to surpass the total number of units in the first hidden layer of any of the networks considered. This set of networks possess and identical structural complexity when the input space dimension g_0 is equal to eight.

Similarly, the graph of figure (6) represents the evolution of the logarithm of the heat capacity of the aforementioned networks with respect to the dimension of the input space. In this graph only 10 dimensions are shown (from 2 up to 11) because from an input space dimension $g_0 \geq 8$ the behavior of the logarithm of the heat capacity simply grows linearly with g_0 for the set of networks sampled and the observed differences between those values are minimal.

The most relevant aspect of this graph is that the value attained by the internal energy is not affected by the number of inputs to the network. In other words, the internal energy of the network does not depend on the dimensionality of the input space to the network. The oscillatory behavior of the internal energy (although practically not appreciated in this graph) can be explained in terms of the procedure used to model the squashing effect of the activation function of the artificial neuron (see expression (7)). Particularly, the fact of using a discrete threshold that employs the modular function is responsible of the observed small oscillatory behavior. The aforementioned behavior of the internal energy was also checked for different multiplex network models that appear in table 1 obtaining similar results. Thus, it can be concluded that according to the theoretical model the internal energy of the networks is not affected by the dimensionality of the input space to those networks.

Moreover, it is important to remember that from the perspective of the theoretical model the effect of the training algorithm used during the learning phase is simply to alter the probability of occurrence of the energy states of the network according to the particularities imposed by the data set to be learnt (i.e., its unknown probability distribution). In other words, the influence of the training set and the learning algorithm is implicitly embedded within the statistics of the random variables m and p of the theoretical model. Thus, taking into consideration the fact that the internal energy is constant with μ_m as well as with the dimension of the input space to the networks it is a plausible to state that this thermodynamic potential is principally linked to the topology of the networks (besides to the total number of units comprising the networks). In other words, the energy stored by deep learning machines is an intrinsic parameter that is linked to its topology and neuron numbers. Furthermore, this parameter is constant with the dimensionality of the input space independently whether the networks considered are shallow or deep. However, according to the model the amount of energy stored by shallow networks is always larger compared to deep networks, thereby evidencing important thermodynamic differences between both models.

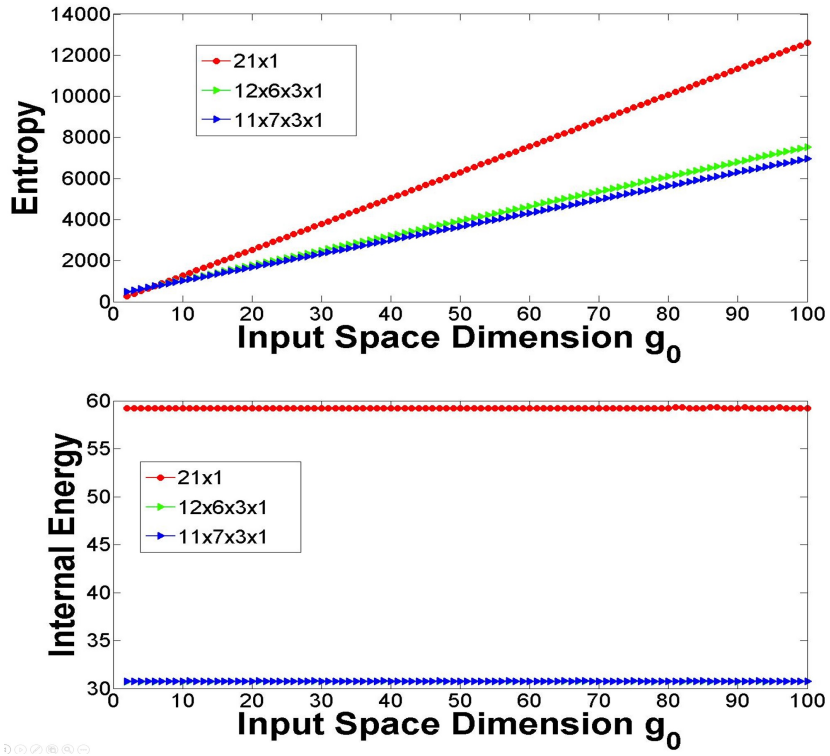

Figure 5. Graphical illustration of the evolution of the entropy and the internal energy as a function of the input space dimension to the networks for two deep networks of three hidden layers and one shallow network that present an identical structural complexity for dimension $g_0 = 8$.The entropy of the networks grows with the dimensionality of the input space, and thus their storage capacity, but at the same time their degree of disorder. The entropy of the shallow network is lower compared to the deep networks when the input space dimension g_0 is lower than 6 being higher otherwise. Comparing the growth of the entropy for the three networks it becames clear the strong influence of the number of units in the first hidden layer of the networks when increasing the dimensionality of the input space. In contrast, the internal energy does not change with the dimensionality of the input space for any of the networks considered. This fact is of particular interest since appears to suggest that the internal energy is an intrinsic property of the networks only linked to its topology and number of units. The internal energy of the shallow network is again higher compared to values attained by the deep networks.

Concerning the behavior of the entropy, as expected it grows with the dimensionality of the input space. The entropy of the shallow network is lower compared to the deep networks when the input space dimension g_0 is lower than 6 being higher otherwise. Comparing the growth of the entropy for the three networks it becames clear the strong influence of the number of units in the first hidden layer of the networks when increasing the dimensionality of the input space [36]. For example, the entropy values of the deep network of three hidden layers 12x6x3x1 with respect to the network 11x7x3x1 become apparent approximately when $g_0 \geq 50$ and it is only due to the fact that the former has an extra neuron in the first hidden layer compared to the second independently of the fact that both networks possess an identical number of neurons.

With regards the behavior of the heat capacity with respect to the dimension of the input space to the networks the most important aspect of this graph (see figure (6)) is the existence of a criticical dimension $g_0 = g_c = 8$ from which the values of the logarithm of the heat capacity (remember the qualifiers heat capacity and specific heat can be used indistinctly in this case study) grows linearly. More specifically, for dimensions $g_0 < g_c$ there exists substantial differences (more than 20 orders of magnitude for 2 or 3 dimensional input spaces) between shallow and deep networks evidencing strong differences in the thermal properties of those networks. In contrast, for $g_0 \geq g_c$ the heat capacity (i.e, the logarithm of the heat capacity) of the networks grows linearly with g_0 and with an identical slope, thereby keeping their relative differences between each other constant with g_0. However, it is also important to note that the condition of identical structural complexity for these networks only holds for $g_0 = 8 = g_c$ that is, the critical dimension g_c appears to be linked to the condition of identical structural complexity.

These facts are of particular interest taking into consideration that the thermodynamic concept of heat capacity expresses the idea how much heat a body it can absorb and it was shown through the theoretical model the link between the heat capacity of a network and its expected learning and generalization performance. Specifically, for the set of networks considered the heat capacity of the shallow network (see table 1) is practically one order of magnitude larger compared to the deep networks, thus, quantitatively the learning and generalization performance of the shallow network 21x1 will be typically better compared with the three hidden layer networks 12x6x3x1 and 11x7x3x1 for any eight dimensional learning dataset.

Having said this, from a learning and generalization perspective increasing the dimensionality of the space rapidly leads to the point where the data is very sparse, in which case it provides a very poor representation of the mapping (i.e., the classification or regression task performed by the neural network), especially, taking into consideration that in practice one is forced to work with a limited amount of data. Indeed, irrespective of the approximation technique employed, learning functions in high dimensions is very difficult (i.e., the curse of dimensionality problem) [7,12]. Surprisingly, from an empirical point of view deep networks behave well even in learning tasks that involve high dimensional data sets [50,51]. The only possibility to achieve a rate of convergence independent of the input dimensionality g_0 is to provide increasing smoothness of the unknown underlying function with increasing g_0, since the number of parameters needed for the approximating function to attain a prescribed degree of accuracy increases exponentially with the input dimensionality g_0 [46].

Thus, taking into consideration firstly, that the energy stored by deep learning machines (shallow or deep) is independent of its number of inputs(i.e., the dimension of the input space). Secondly, that increasing the depth of the networks leads to a progressive reduction of the energy stored by hierarchical networks accompanied of a reduction of the level of disorder, and finally the unavoidable influence of the internal energy in the geometrical complexity of the loss landscape, if deep networks behave well even in learning tasks that involve high dimensional data sets (something that is only possible to achieve providing increasing smoothness of the unknown underlying function with increasing g_0) it is plausible to hypothesize that this fact might be linked to the amount of energy stored by the networks since is the only parameter that quantitatively decreases when increasing the number of hidden layers together with the level of disorder.

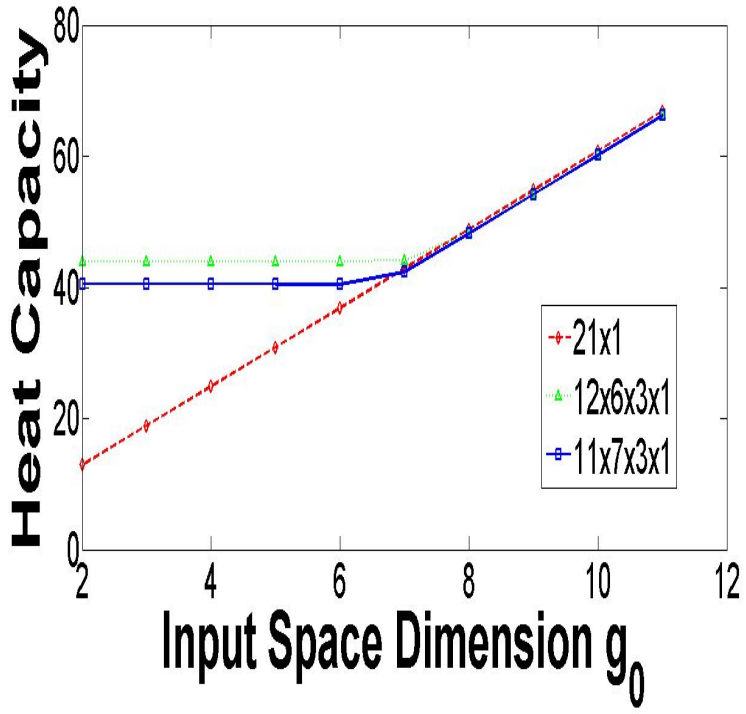

Figure 6. Graphical illustration the behavior of the logarithm of the heat capacity for three of the network architectures shown in figure (5) as a function of the input space dimension to the networks. These set of networks present an identical structural complexity for dimension $g_0 = 8$ and this value marks a transition on the behavior of the heat capacity of the networks. The most important characteristic of this graph is the behavior of the heat capacity before and after the critical dimension $g_0 = g_c = 8$. For dimensions $g_0 < g_c$ there exists substantial differences (more than 20 orders of magnitude for 2 or 3 dimensional input spaces) between shallow and deep networks evidencing strong differences in the thermal properties of those networks. In contrast, for $g_0 \geq g_c$ the heat capacity (i.e, the logarithm of the heat capacity) of the networks grows linearly with g_0 and with an identical slope, thereby keeping their relative differences between each other constant with g_0.

Summarizing, the input space dimension does not affect to the amount of energy stored by the networks, however increasing the depth of the networks leads to a quantitatively reduction of the energy stored by the networks as well as its capacity to absorb heat accompanied with a reduction of the level of disorder of the equilibrium states represented by the networks.

4.2.4. Thermodynamic Interpretation

In thermodynamics the change in energy of a body from one state to another can be divided into the quantity of heat gained (or lost) by the body, and the work done on it (or by it on other bodies). However, the internal energy of a body cannot be split into heat and mechanical energy. In other words, one can speak of the internal energy of the body in a given state but one cannot, for instance, talk of the quantity of heat possessed by a body in a given state [34]. This division is only possible if one refers to a change of energy.

According to the model presented in section 3, each of the group of networks considered in this section (i.e., groups $M_{1a} \cup M_{3a}$, M_{2a},M_{2b}, and M_{3b}) correspond to the equilibrium states of a physical system, that is, feedforward fully-connected artificial neural networks with an identical structural complexity in this particular case. Thus, this physical situation makes the canonical ensemble the most adapted statistical ensemble to the analysis of the macroscopic properties of this system. Moreover, these equilibrium states are characterized by an energy value (internal energy) that is shown in Table 1. Furthermore, the values of the entropy, free energy, and the specific heat are also available. It is important to note that the condition of thermal equilibrium is guaranteed because of the fact that the specific heat is positive for the whole set of equilibrium states represented by the network architectures.

To better understand the macroscopic thermodynamic behavior of the model it is assumed, in the following, the existence of an hypothetical process (or external force) that doing work on the system perturbates the equilibrium states represented by the group of network architectures of table 1 (i.e., those that follow an identical structural complexity scheme) by making changes in their topology so as to pass from one multiplex network architecture to any other of those defined within its group. This is equivalent to the general formulation of the law of increase of entropy where the system steadily passes from states of lower entropy to those of higher entropy, that is, the successive states through which the system passes correspond to energy distributions of successively greater probabilities. It is important to remember that if a closed system is at some instant of time in a macroscopic state not in equilibrium, then the most probable consequence will be that the entropy of the system will increase monotonically in succesive instants of time. Indeed, the probability of a change to a state of higher entropy is so overwhelmingly large compared with the probability of any appreciable decrease that the latter cannot in fact be observed in nature [34]. Thus, the kind of theoretical transformations (from one equilibrum state to another) studied hereafter are those that follow the law of increase of the entropy (i.e., the second law of thermodynamics). The main goal is to understand the influence of the topology of the networks in the structure and role of the phase space of the system, the stability of the equilibrium states, and its relationship with the resulting level of disorder of the networks, as well as with its information storage capacity. It is important to remember that the final goal is to relate the macroscopic thermodynamic behavior of the model with the observables of these machine learning models in real-world applications, that is, their learning and generalization capabilities. Intuitively, those networks attaining higher entropy values have the potential to provide better learning and generalization performance since as higher is the entropy as higher will be the complexity of the space of functions they are able to represent, and thus its VC dimension. Thus, the theoretical transformation of state studied hereafter reflect the intuitive plausibility that generalization tends to improve to the extent the entropy increases.

With a total number of 22 units the network architectures 21x1, 12x6x3x1, and 11x7x3x1 form a group of networks with an identical structural complexity. Specifically, this group of architectures form a subset of the whole set of possible equilibrium states of an hypothetical physical system whose thermal equilibrium states are defined by the set of network architectures composed of $N = 22$

units,with an input space dimension to the networks $g_0 = 8$, and with an effective complexity $W = 189$. The architecture 21x1 correspond to one of the possible solutions of the set of equilibrium states.The presence of these set of equilibrium states opens the possibility of the existence of both: stable and metastable equilibrium states because the aforementioned conditions for equilibrium are not sufficient for the equilibrium to be completely stable.

Assuming that the physical system is found in the equilibrium state represented by the deep network of three hidden layers 12x6x3x1 (this state is denoted herafter as state 2). The system is then perturbated by rearranging neurons and connections (i.e., by doing work on the system) passing to another equilibrium state represented by the multiplex networks M_{1a}, corresponding to the shallow network 21x1. Let us denote this state as state 3. Clearly, the result of this transformation is an increase of the internal energy of the body (i.e., the neural network) since $\Delta U = U_3 - U_2 = 59.21 - 30.71 = 28.5 > 0$, the entropy of the system increases $\Delta S = S_3 - S_2 = 1021.54 - 903.457 = 118.083 > 0$, but at the same time, the free energy of the system decreases $\Delta F = F_3 - F_2 = -89.575 < 0$.

Moreover, it is important to remember that in conservative mechanical systems (e.g., a mass raised in a gravitational field), work can be stored in the form of potential energy and subsequently retrieved. This is also true for thermodynamic systems. For example, it is possible to store energy in a thermodynamic system by doing work on it through a reversible process, and that energy can eventually be retrieved in the form of work. The energy which is stored and retrievable in the form of work is called the free energy. Indeed, there are as many forms of free energy in a thermodynamic system as there are combinations of constraints [67].

Under the assumption that the deep network 12x6x3x1 (i.e., the body) is thermally isolated the observed change in the internal energy is entirely due to the work done on it, that is, the work done on the artificial network is work done against the external field represented by the topology through the displacement of neurons from the second and third layers to the first and rearranging their connectivity, so as to end up in the equilibrium state reflected by the shallow network 21x1. Thus, as a result of this process, the number of degrees of freedom of these set of neurons (i.e, six neurons from the second layer, and three neurons from the third layer) has changed , being this process equivalent to displace the artificial neurons (i.e., the particles) in the presence of an external field represented by the topology, that is, the external field generates a "force" that affects the neurons of the network and that varies depending on the position where the neuron is situated (i.e., the layer where the neurons are situated). The result of this process is that part of this work is transformed into potential energy stored in these internal degrees of freedom contributing to the energy content of the network (i.e., the internal energy) but not to its temperature.

Moreover, according to the law of increase of entropy the equilibrium state represented by the shallow network (with a higher entropy) is more probable compared to any of those represented by the deep networks within this group. Of particular interest is also the fact that the degree of disorder in the equilibrium state represented by the shallow network is higher compared to the degree of disorder of the deep networks architectures considered. Furthermore, the higher degree of disorder of the shallow network is also acompanied with a larger value of the specific heat when compared to the deep networks.

In a similar way, using now the architectures of the group M_{3b}, that is , another physical system (i.e., a theoretical body) whose thermal equilibrium states are defined by the set of all network architectures composed of $N = 267$ units, with an input space dimension to the networks $g_0 = 8$, and with an effective complexity $W = 11772$. Let as sort the states represented by these four architectures according to the value attained by the entropy.

Assuming that this physical system is found in the equilibrium state represented by the deep network of three hidden layers 16x54x196x1 (this state is denoted as before state 1), the system is perturbated by displacing neurons from the third hidden layer, to the second and re-arranging connections (i.e., by doing work on the system) so as to end up in the equilibrium state represented by the expansive autoencoder architecture 16x211x39x1 (state 2), the result of this transformation is

also a loss in the internal energy of the body $\Delta U = U_2 - U_1 = 390.4 - 638.9 = -248.5 \pm 11.18 < 0$, the free energy decreases $\Delta S = F_2 - F_1 = -664.5$, but at the same time a scarcely gain of entropy $\Delta S = S_2 - S_1 = 43530.4 - 43111.4 = 416 > 0$. The small gain in the entropy can be explained as the result of a small expansion of the accesible volume of the phase space of the system during this process. Specifically, in this transformation the accesible volume of the phase space remains practically constant ($M(1)_{16\times54\times196\times1} \cong M(1)_{16\times211\times39\times1}$) in spite of the fact of passing to a state of higher entropy. In other words, the constraints imposed by the topology (i.e., the external field) are slightly relaxed and the accesible volume of the phase space for the set of 267 neurons slightly increases as a result of the displacement of neurons from the third layer to the second. However, there is a substantial change in the heat capacity of the body (i.e., passing from 6.55×10^{87} to 5.03×10^{338}), thus suggesting substantial changes in the thermal properties of the body (i.e., the neural network). It is important to note that the qualifiers heat capacity and specific heat can be used indistinctly for networks with an identical structural complexity.

Similarly, assuming now that the system is found in the equilibrium state corresponding to the network 16x211x39x1 (state 2) and it is perturbated so as to end up in the equilibrium state represented by the compressive autoencoder architecture 108x50x108x1 (state 3), there is a also a loss in the internal energy of the body $\Delta U = U_4 - U_3 = 390.1 - 499.9 = -109.8 \pm 12.73 < 0$, but lower compared to the previous transformation. In contrast, there is larger gain of entropy $\Delta S = S_4 - S_3 = 45042 - 43530.4 = 1511.6 > 0$ indicating an expansion of the accesible volume of the phase space. Specifically, the accesible volume of the phase space passes from $M(1)_{16\times211\times39\times1} = 2.82 \times 10^{18987}$ accesible states to a total number of $M(1)_{108\times50\times108\times1} = 2.44 \times 10^{19703}$ accesible states, that is, a difference of 716 orders of magnitude. However, as opposed to the transformation $T_{1\to2}$ this transformation is characterized by a reduction of the heat capacity of the body passing from a value of 5.03×10^{338} to 6.26×10^{173}.

Finally, if one now assumes that from the equilibrium state 3 (the compressive autoencoder architecture), the system is then perturbated by displacing neurons from the third layer to the second, and first hidden layers of the network re-arranging connections (i.e., by doing work on the system) so as to pass to state 4 (i.e. the equilibrium state corresponding to the hierarchical network 180x47x39x1). The result of this transformation is again a decrease of the internal energy of the body since $\Delta U = U_4 - U_3 = 390.1 - 499.9 = -109.8 \pm 12.77 < 0$, the entropy of the system increases $\Delta S = S_4 - S_3 = 46550 - 45042 = 1508 > 0$ in a similar way compared to the previous transformations, but the free energy of the system substantially decreases $\Delta F = F_4 - F_3 = -1617.8 < 0$ compared to the previous transformations $\Delta F = F_3 - F_2 = -1402$. Of particular interest is the fact that in this transformation the accesible volume of the phase space passes from $M(1)_{108\times50\times108\times1} = 2.44 \times 10^{19703}$ to $M(1)_{180\times47\times39\times1} = 1.03 \times 10^{20266}$), that is, an expansion of 563 orders of magnitude that is 153 orders of magnitude lower compared to the previous transformation $T_{2\to3}$ in spite of the fact of obtaining a similar entropy gain. Furthermore, the heat capacity of the body decreases in this transformation of state passing from 6.26×10^{173} to 1.36×10^{289}., thus pointing again to changes in the thermal properties of the body.

Having said this, it is important to remember that entropy can be seen as an indicator of the probability of a state, thus, from the four equilibrium states considered the network 180x47x30x1 is the most probable state. Indeed, the number of accesible states for this network is exponentially larger compared to those of the networks 16x54x196x1, 16x211x39x1, and 108x50x108x1 suggesting a higher degree of disorder for this equilibrium state compared to the others. It is important to remember that a measure of the disorder of a physical system is the number of accesible states, therefore, the networks 16x211x39x1, 108x50x108x1, and 180x47x39x1 present a higher degree of disorder, and thus a higher information storage capacity compared to the expansive network architecture 16x54x196x1. Furthermore, the compressive autoencoder architecture 108x50x108x1 together with the expansive network architecture 16x54x196x1 constitute the equilibrium states that store larger amounts of energy, and at the same time are those presenting smaller heat capacities compared to the rest of network architectures.

Those facts are of particular interest if one takes into account that the concept of heat capacity provides information about the thermal properties of a body. The training phase of these models involves the modification of the values of the adaptive weights of the network. This process can be interpreted as changing the probability of occurrence of the microstates of the system according to the theoretical model. Furthermore, from a thermodynamics point of view the aforementioned process can be seen as energy that is brought into the system while no work is delivered [31], more specifically, the external parameters of the hamiltonian are unchanged (i.e., the topology of the network does not change), the energy levels of the system remain identical but the average energy of the system changes throughout this process due to the changes in the probabilities of ocurrence of the microstates and the variation of internal energy is identified with the inifinitesimal heat absorbed by the deep learning machine.Taking into consideration that the heat capacity expresses the idea how much heat a body it can absorb it is plausible to state that as lower (or higher) is the heat capacity of a network as lower (or higher) will be its capacity to absorb heat and therefore more difficult (or less difficult) will be to change the probability of occurrence of the microstates.

Thus, owing to the fact that changing the probability of occurrence of the microstates is equivalent in machine learning terms to the process of finding good minimizers during the learning phase this process will be harder (or easier) depending on the heat capacity of the network as a result of its influence in the geometrical structure of the loss landscape [52,69]. In other words, the topology not only exert influence upon the values attained by the entropy, and the internal energy of the networks but especially in their thermal properties, which in turn influence the geometrical complexity of the loss landscape being this fact independent of the learning algorithm and/or the traininig set used (remember that the learning and recall phases of deep learning machines, the particularities of the training set, and the learning algorithm used are abstracted in the theoretical model).

Accordingly, one would expect that is specially more difficult to learn any input-output mapping for the network 16x54x196x1 compared to the rest. Furthermore, the compressive autoencoder architecture 108x50x108x1 should also present some difficulties in spite of the fact of attaining a higher entropy value. Due to the strong restrictions imposed by the topology (the external field) the accesible volume of the phase space of the system is mimimal for the equilibrium state represented by the network 16x54x196x1. The capacity of the physical system to store information shrinks for this equilibrium state, and its heat capacity is severely reduced, thus leading to a harder energetic landscape compared to those induced by the rest of architectures. As a matter of fact the architecture 16x54x196x1 represents the network storing the largest amount of energy (followed by the compressive autoencoder architecture) compared to the rest. Thus, the observed behavior of the internal energy (under the restrictions imposed by this case study) is evidencing the intuitive plausibility that as higher is the energy stored by a network as lower will be its capacity for heat absortion.

Having said this, it is important to remember that changes in the external conditions in which a body is situated form a special class of the external influences to which a body is subject [34]. For example, in thermodynamics the external conditions most frequently take the form of a predescribed volume for the body, for example, a gas contained in a chamber. When the predescribed volume for the body increases (like in the Joule-Guy Lussac free expansion process [31]), the existing constraints on the positions of its molecules change, thus, the accesible volume for the molecules of the gas increases. In other words the number of "cells" of the one molecule phase space have increased with the accesible volume in this process, that is, the number of accesible states of the system increases. The physical system under consideration is a theoretical one, however, in the transformations of state (those according to the law of increase of entropy), the accesible volume of the phase space changes (expands or collapse) depending on the intrinsic characteristics of the equilibrium states involved.

Thus, changes in the number of accesible states in the theoretical transformations studied can be interpreted as changes in the predescribed "volume" in which the neurons (i.e, the particles) are confined, and accordingly the concept of pressure may also be defined from one of the most important

thermodynamic identities [34]: $\Delta U = T\Delta S - P\Delta V$ that relates energy, entropy, pressure, temperature, and volume. Accordingly, pressure finally reads as:

$$P = \frac{T\Delta S - \Delta U}{\Delta V} \tag{49}$$

where the variations in volume ΔV (see equation (49)) are calculated in terms of the difference on the total number of accesible states when passing from one equilibrium state to another.

It is important to note that statistical mechanics [15,67] defines the temperature of any system to be the derivative of energy with respect to entropy. The entropy of the model is justified simply from information-theoretic considerations. The system under study contains information which is represented in terms of the adaptive weights and the states of the neurons. Similarly, the physical model described also has an energy function. Therefore, the idea of volume, and pressure can be applied in this context even though the model has no resemblance to the physical systems (e.g., gas particles) where these thermodynamic concepts began with. Furthermore, the whole thing is that at any equilibrium state the "pressure" of the body (i.e., the neural network) must be positive. Equilibrium states with negative pressure correspond to metastable states, that is, states that are stable within certain limits (e.g., if a body is in a metastable state, then after a sufficiently large deviation from it the body cannot return to its initial state [34]).

Taking these considerations into account, from the set of transformations considered before (i.e., those for networks with $N = 267$), it can be verified that both, the variation of the accesible volume of the phase space is positive for all of the transformations considered, and the gain in entropy obtained ΔS in the set of possible transformations of state is always big enough to ensure that the pressure (49) is positive. For example, the transformation $T_{1\to 2}$ (i.e., from the equilibrium state represented by the network 16x54x196 to the network 16x211x39x1), that is, $\frac{(S_2-S_1)-(U_2-U_1)}{M(1)_{16\times 211\times 39\times 1}-M(1)_{16\times 54\times 196\times 1}} > 0$, or from state 2 to state 3. Similarly, when considering the transformation $T_{2\to 3}$, that is, from the network 16x211x39x1 (equilibrium state 2) to the network 108x50x108x1 (equilibrium state 3). The "pressure" is also positive $\frac{(S_3-S_2)-(U_3-U_2)}{M(1)_{108\times 50\times 108\times 1}-M(1)_{16\times 54\times 196\times 1}} > 0$. Thus, it can be concluded that the equilibrium state represented by the network 108x50x108x1 is also a stable state.

However, as oppossed to the previous transformation (see table I) this transformation is characterized by a reduction of the heat capacity of the body passing from a value of 5.03×10^{338} to 6.26×10^{173}. It is important to remember that from the point of view of thermodynamics the observed changes in the heat capacity of the body as a result of the theoretical evolution of the system throughout its different equilibrium states are indicating changes in the thermal properties of the body. In summary, it can be asserted that these set of equilibrium states are stable.

Moreover, for the first group of architectures, it can be also verified that the shallow network 21x1, and the deep networks 12x6x3x1, and 11x7x3x1 also correspond to stable equilibrium states. Specifically, the theoretical transformation of state according the second law of thermodynamics from the equilibrium state represented by the network 12x6x3x1 to the shallow network 21x1 leads to a positive variation of "pressure" indicating also stability, or from the state 11x7x3x1 to 12x6x3x1. Of particular interest is the fact that the transformation from 12x6x3x1 to 21x1 is accompanied of an increase of the heat capacity of the body whereas in the transformation from 16x211x39x1 to 108x50x108x1 do not, specially taking into consideration the influence of the heat capacity in the geometrical complexity of the loss landscape.

Finally, with regards equilibrium states that are stable wihin certain limits the three hidden layer network 1122x2x184x1 (state A) for $g_0 = 8$ possess an identical structural complexity with respect to the equilibrium state represented by shallow network 1308x1 (state B) of table 1 (i.e., both networks with $W = 11772$, and $N = 1309$), attaining an entropy, free energy, and internal energy that are respectively $S = 63179.5$, $F = -61271$,and $U = 1908.5$. Similarly, the number of accesible microstates for these networks is $M(1)_{1122\times 2\times 184\times 1} = 2.82 \times 10^{27835}$, and $M(1)_{1308\times 1} = 2.8 \times 10^{29323}$. Using equation (49) to analyze the stability of the transformation $T_{A\to B}$ it can be verified that the

equilibrium state represented by the shallow network is a metastable state. Of particular interest is the fact that this transformation leads to a positive increase of the free energy ($\Delta F > 0$).

Summarizing, the most important conclusions of this section can be summarized as follows:

1. The theoretical analysis performed appears to corroborate the strong influence of the topology of the networks in the structure of its phase space which in turn affects their thermal properties, their resulting degree of disorder, the stability of the equilibrium states represented by the networks, and its information storage capacity. More specifically, when conserving the structural complexity of networks the degrees of freedom responsible of the accesible volume of the phase space of the system (i.e., the total number of adaptive weights W and the total number of units N of the networks) do not change when the system passes from one equilibrium state to another. However, the accesible volume of the phase space changes (i.e., the number of accesible microstates) as a result of the strong influence of the topology (remember the interpretation of the topology as an external field). In other words, the topology constrains the number of accesible microstates of the physical system, thereby exerting a strong influence in the thermodynamic properties of the networks.
2. The topology affects the thermal properties of the networks, which in turn influence the geometrical complexity of the loss landscape being this fact independent of the learning algorithm, traininig set used, and/or the existence (or not) of overparametrization conditions (i.e., networks with more adaptive weights than data points). More specifically, the process of finding good minima during a learning task for any machine learning algorithm might be easier or harder depending on the available accesible volume of the phase space, the amount of energy stored by the network and its heat capacity.
3. Quantitatively stable equilibrium states (networks) leading to better generalization performance are expected to be those with larger information-storage capacities, however, a larger capacity (i.e., a higher entropy value) might not be enough to provide better learning and generalization performance as a result of the influence of the specific heat and the energy stored by the networks in the geometrical complexity of the loss landscape. Metastable states appear to be linked to those transformations of state involving a positive increase of the free energy.

In order to assess the theoretical predictions of the model, that is, how the differences of the thermodynamic signature of the networks and their learning and generalization capabilities of these models correspond some experimental results are shown in the next section.

4.2.5. Experimental Results

The graphs of figure (7) represent the generalization performance for the set of multiplexes architectures studied in section 4.2 (see table I) when using as training algorithm standard gradient descent (i.e., Backpropagation).

The dataset used for training the networks is an artificial database *Gauss8d.dat* used in [12] to study the effect of the inpùt dimension on the classification error rates. The distribution is gaussian for the two classes, and there is a fully overlapping between these classes, the center of gravity is the same for the two classes. As the dimension increases (from 2 to 8), the number of available classes remains the same, 2500 patterns for each class. The theoretical Bayes errors for dimension eight is 9%. It is important to note that for dimension eight the database does not contain an enough number of samples to reliably estimate the underlying probability distribution that generated the data [12]. The set of 3 hidden layer architectures are overparameterized (i.e., the total number of adaptive weights is bigger than the total number of patterns contained in the dataset. The qualifier speed of learning is used hereafter to define the minimal number of epochs needed to reach a predefined learning and/or generalization error.

A ten fold crossvalidation procedure [25] models was used to assess the generalization performance of the networks using ten averages per fold so as to minimize the presence of artifacts

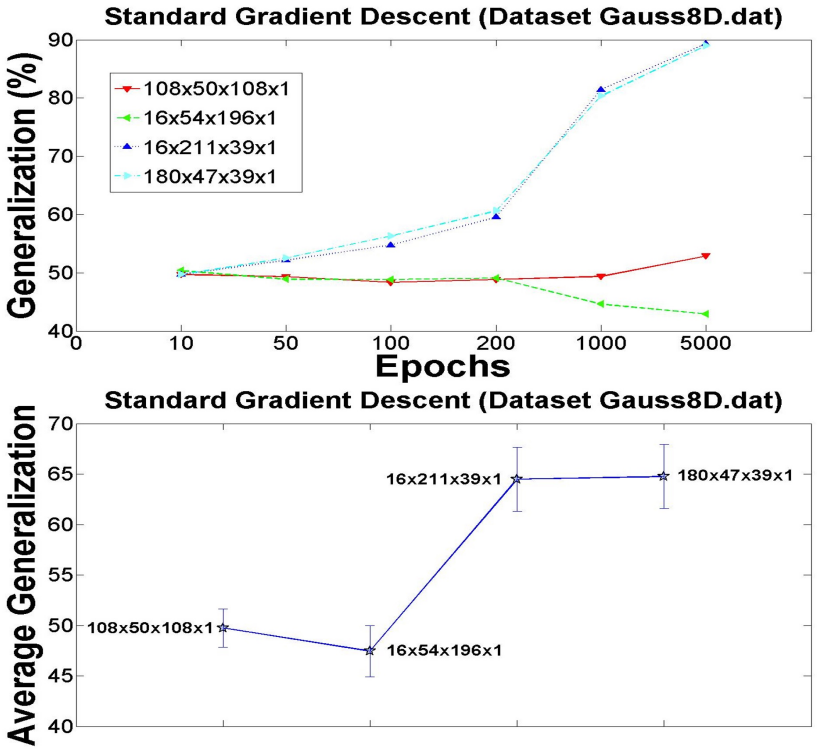

Figure 7. Graphical illustration of the generalization performance obtained for the group of networks M_{3b} of table 1 presenting an identical structural complexity using as learning algorithm backpropagation with the dataset *Gauss8d.dat* [12]. The top part represents the evolution of the generalization performance of the networks sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The botton part represents the average generalization (and its standard deviation is represented as an error bar) obtained when averaging the generalization at the sampling points described before. The goal is to assess the typical behavior of the network architectures in terms of generalization performance, that is, which is the typical expected behavior of the generalization independently of the number of epochs used to train the networks. Networks storing larger amounts of energy (i.e., 108x50x108x1 and 16x54x196x1) and attaining smallest heat capacities are those presenting the worst learning and generalization performance for standard gradient descent. The generalization performance improves with the value attained by the entropy (see table I) excepting for the compressive autoencoder architecture 108x50x108x1, thus evidencing the existence of a more complex geometrical structure of the loss landscape [69].

that may appear due to the implicit randomness associated to the training phase of these models, thereby increasing the accuracy of the results.

Each graph is composed of two parts. The top part represents the evolution of the generalization performance of the network sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The idea is to assess the generalization capabiliy of the networks against underfitting and overfitting conditions. It is important to remember that the toy datasets used are composed of 5000 patterns thus it is likely that overfitting effects appear for a number of epochs close to the total number of patterns of the datasets, especially for those architectures that are overparametrized (i.e., the total number of adaptive weights is bigger than the total number of patterns of the dataset). The botton part represents the average generalization (and its standard deviation is represented as an error bar) obtained when averaging the generalization at the sampling points described before. It is important to remember that the goal is to assess the typical behavior of the network architectures in terms of generalization performance, that is, which is the typical expected behavior of the generalization independently of the number of epochs used to train the networks.

The networks of figure (7) are overparameterized for the dataset *Gauss8D.dat*, and were selected precisely to study the influence of the topology in their learning and generalization performance for networks with an identical depth (i.e., an identical number of hidden layers) under the presence of the curse of dimensionality phenomenon.

When comparing the generalization performance obtained with standard gradient descent techniques, the worst results are obtained for the network architecture 16x54x196x1, that is the network attaining the lowest value of entropy thus appearing to corroborate the predictions of the theoretical model. Specifically, due to the strong restrictions imposed by the topology the accesible volume of the phase space of the system is mimimal for the equilibrium state represented by this expansive network architecture. Compared to the rest of architectures, the capacity to store information shrinks for this equilibrium state, and its heat capacity is severely reduced leading to a more complex geometrical structure of the loss landscape (i.e., the landscape of empirical risk) [52,69] both facts affecting the resulting learning and generalization performance of the network.

Of particular interest is the fact that this network together with the compressive autoencoder architecture 108x50x108x1 constitute the networks storing largest amounts of energy and attaining smallest heat capacities compared to the rest of architectures, and precisely, these networks are those attaining the worst learning and generalization performance for standard gradient descent.

Moreover, the generalization performance appears to improve with the value of the entropy attained by the networks (see table 1) excepting for the network 108x50x108x1. Particularly, the hierarchical network 180x47x39x1 is the topological scheme attaining the best learning and generalization performance. Surprisingly, the compresive autoencoder 108x50x108x1 is the network after the hierarchical network 180x47x39x1 attaining the largest entropy value. Thus, the reduced heat capacity of this network compared to the expansive autoencoder 16x211x39x1, and the hierarchical network 180x47x39x1 appear to indicate a more complex geometrical structure of the loss landscape in spite of having a larger number of accesible microstates (i.e., a larger accesible volume of the phase space), and thus a larger storage capacity compared to the networks 16x54x196x1 and 16x211x39x1 respectively.

To test this hypothesis the generalization capability of this network was sampled again but at 10000,20000,25000, and 50000 epochs obtaining the values 53.03%, 60.58%,63.15%, and 63.98% thereby appearing to corroborate the aforementioned hypothesis, that is, the speed of learning is being clearly limited by the difficulties imposed by the complexity of the geometrical structure of the loss landscape as a result of the topology of the network (i.e., compressive autoencoder architecture) but specially by its reduced heat capacity.

Moreover, to further investigate this issue, the generalization performance of these networks was also assessed using the scaled conjugate gradients algorithm [11,30] (see appendix B), that is, a second order optimization method with accelerated convergence properties compared to backpropagation, corroborating the predictions of the theoretical model, that is, using a learning protocol with better convergence properties leads to the fact that the learning and generalization performance obtained are now in direct correspondence with the entropy values attained by the networks. Having said this, it is important to remember that according to the theoretical model each network architecture represents an equilibrium state of a physical system, and the evolution from one equilibrium state to another is drawn according to the law of the increase of the entropy. Taking into consideration that entropy is linked to the generalization performance of the networks, the second law of thermodynamics can be used to compare the generalization performance of different network architectures.

Quantitatively, for stable states the generalization performance improves with the value attained by the entropy, excepting for those transformation of state accompanied of a reduction of the heat capacity where the improvement on generalization performance is limited by the convergence properties of the learning algorithm used. For example, when using as training algorithm the scaled conjugate gradients, for the state represented by network 108x50x108x1 the generalization improves with respect to the theoretically preceding state represented by the expansive autoencoder network

architecture 16x211x39x1 (transformation $T_{2\to3}$), or when moving according to the law of increase of the entropy from the state 108x50x108x1 towards the stable equilibrium state 180x47x39x1 the generalization improves as expected (transformation $T_{3\to4}$, or in any other transformation respecting the accesibility of states such as $T_{1\to3}$, $T_{1\to4}$, or $T_{2\to4}$), thus corroborating the theoretical predictions of the model.

4.2.6. Summary of Results

When conserving the structural complexity of the networks the principal conclusions that can be extracted when combining both the theoretical and experimental results can be synthesized as follows:

1. Quantitatively networks leading to better generalization performance are those with larger information-storage capacities. From the whole set of network topologies, hierarchical networks are those leading to higher entropy values, and thus to better generalization performance. However, a larger capacity (i.e., a higher entropy value) might be not enough to provide better learning and generalization performance depending on the convergence properties of the learning algorithm used because of the influence of the specific heat and the energy stored by the networks in the geometrical complexity of the loss landscape, evidencing the strong influence of the topology in their thermodynamic signature.
2. Network architectures represent the equilibrium states of a physical system. Entropy is linked to the generalization performance of the networks. The second law of thermodynamics permit us to calculate entropy variations, and thus to compare the generalization performance of different network architectures. Specifically, if A and B are two network architectures with identical structural complexity if $S_A > S_B$ the state A is adiabatically accesible from state B, thus the typical generalization performance of Network A will be better compared to Network B if both states are stable excepting for those transformations accompanied with a reduction of the heat capacity indicating the existence of a more complex geometrical structure of the loss landscape. In that case, the improvement on generalization performance is limited by the convergence properties of the learning algorithm used. Metastable states are characterized by storing larger amounts of energy and they are linked to those transformations involving a positive increase of the free energy.
3. The input space dimension does not affect the amount of energy stored by the networks (both shallow and deep networks), however increasing the depth of the networks leads to a quantitatively reduction of the energy stored by the networks as well as its capacity to absorb heat accompanied with a reduction of the level of disorder of the equilibrium states represented by those networks. In other words, deep networks behave well even in learning tasks that involve high dimensional data sets because they are able to provide increasing smoothness of the unknown underlying function with increasing input dimensionality as a result of a progressive reduction of the energy stored by the networks when increasing its depth. Shallow networks store larger amounts of energy compared to deep networks leading to the generation of more complex geometrical structures of the loss landscape being this fact independent of the dataset or the learning algorithm used as a result of the profound influence of the topology in the thermodynamic signature of the networks.

4.3. Controling the Effective Complexity: Case Study II

This section is aimed at studying the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers), and the total number of units N when controling the total number of adaptive weights of the networks using the thermodynamic formalism derived in section 3. The influence of the input space dimension is not presented since the results of this study are identical to those obtained for case study I.

In this case the total number of units of the networks N plays the role of a variable parameter. Depending on the topology of the network fixing the total number of adaptive weights may lead to networks with substantial differences on the total number of neurons.

4.3.1. Network Architectures: A Preliminary Analysis

A set of seven network architectures with an identical effective complexity were selected for the study under the assumption of an input space dimensionality $g_0 = 2$.

One multiplex network M_1 with architecture 1308x1 (i.e., a shallow network), and six multiplex networks M_3 (deep networks with three hidden layers) to study the influence of the topology and the total number of nodes following the possible topological schemes commonly used in the machine learning literature for this depth (i.e., hierarchical $g_1 > g_2 > g_3$, expansive autoencoder $g_1 < g_2 > g_3$, and compressive autoencoder $g_1 > g_2 < g_3$ and $g_1 < g_2$) but also including the expansive topological scheme $g_1 < g_2 < g_3$ described in section 4.1. More specifically, it is composed of the three hierarchical networks with architectures 74x49x3x1 ($N = 127$), 162x21x9x1 ($N = 193$), and 300x11x2x1 ($N = 314$) with an increasing number of neurons to study the influence of the number of neurons for networks with an identical topology. Two autoencoder architectures 144x11x171x1 and 24x59x41x1 with a compressive and expansive scheme respectively, and finally one network 30x32x88x1with an expansive topological scheme for comparison purposes.

Table 2. The set of network architectures selected for the study of the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers), and the effect of the input space dimension to the networks when controling the total number of adaptive weights of the neworks W. The table shows the thermodynamic potentials associated to each of the networks considered (i.e., , entropy S, free energy F, internal energy and its fluctuations $U \pm \Delta U$) , the specific heat C_v, the total number of adaptive weights W for an input space dimension $g_0 = 2$, and the total number of units N.

Multiplex M_1	**S**	**F**	**U** $\pm\Delta U$	C_v	W	N
1308x1	16607.2	-12919.5	3687.7 ± 23.41	2.49×10^7	3924	1309
Multiplex M_3						
24x59x41x1	14436.6	-14218.4	218.2 ± 7.26	1.62×10^{95}	3924	125
74x49x3x1	14766.8	-14606.2	160.6 ± 7.32	2.08×10^{119}	3924	127
30x32x88x1	14331.4	-14006.7	324.7 ± 7.96	1.92×10^{52}	3924	151
162x21x9x1	15134.6	-14882.8	251.8 ± 9.01	8.54×10^{259}	3924	193
300x11x2x1	15749.1	-15358.6	390.5 ± 11.45	5.93×10^{480}	3924	314
144x11x171x1	14578.7	-13904.7	674 ± 11.73	6.5×10^{230}	3924	327

Table 2 shows the particular values of the thermodynamic potentials (entropy S, free energy F, the internal energy U and its fluctuactions ΔU), the specific heat C_v, and the total number of nodes N for the aforementioned set of networks for $\beta = 1$, $\mu_m = 20$, and assuming a bi-dimensional input space ($g_0 = 2$) to the networks.

In this case each of the multiplex networks networks considered in this section (i.e., those of table 2) would correspond to the equilibrium states of a physical system defined by the set of all feedforward fully-connected artificial neural networks with two inputs (i.e., $g_0 = 2$), and an effective complexity $W = 3924$. It is important to emphasize that there may exists a whole number of network architectures with an identical effective complexity but with substantial variations in the total number of units N and/or the topology. Thus, networks with a larger number of neurons are expected to attain (in general) larger entropy values and/or a larger number of accesible states because each additional neuron contributes with an extra degree of freedom that comes from its own state (active or not active) independently of the fact that the total number of adaptive weights are conserved.

Comparing the values of the entropy attained by the different set of network architectures it can be deduced that hierarchical networks (including shallow networks) are those attaining the highest entropy values, thus indicating that this topological scheme leads to networks with the highest

information-storage capacities, and at the same time those equilibrium states tending to present a higher degree of disorder. The influence of the topology is again so strong that for example the hierarchical network 74x49x3x1 ($N = 127$) attains a larger entropy value than the compressive autoencoder 144x11x171x1 ($N = 327$) in spite of possesing a smaller number of units (almost three times less). In contrast, the network that attains the lowest entropy value, and thus the poorest information-storage capacity correspond again to the expansive network topology 30x32x88x1 ($N = 151$), which evidences at the same time the equilibrium state with the lowest degree of disorder.

The energy stored by the networks increases with neuron numbers as well as their fluctuations, however as occurred before hierarchical networks represent the topological scheme (excepting the shallow network) that stores the lowest amount of energy, for example the expansive network architecture 30x32x88x1 with $N = 151$ stores a higher amount of energy compared to the hierarchical network 162x21x9x1with $N = 193$ in spite of the fact of having a lower number of units. The specific heat is a state variable that is extensive (i.e., change in value when the size of the system is changed), and thus, it changes with neuron numbers. Specifically, it appears that its value increases with neuron numbers although not monotonically because of the influence of the topology. For example, for hierarchical networks such as 74x49x3x1 ($N = 127$), 162x21x9x1 ($N = 193$), and 300x11x2x1 ($N = 314$), the specific heat increases with neuron numbers. Indeed, this topological scheme attains the largest values excepting for the shallow network. In contrast, the shallow network together with the expansive network 30x21x88x1 are those topological schemes attaining the lowest values.

Of particular interest is the fact that networks with lower specific heat values are those that store larger amounts of energy as occurred in case study I, that is, when conserving the structural complexity. Particularly, the expansive network architecture 30x32x88x1 ($N = 151$), the compressive autoencoder architecture 144x11x171x1 ($N = 327$), and the shallow network 1308x1 ($N = 1309$) are the networks that attain the lowest values of specific heat, and at the same time are those that store larger amounts of energy. Furthermore, the shallow network correspond to the topological scheme that attains the lowest value of specific heat in spite of the fact of being the network with the largest amount of neurons as well as the network that stores the largest amount of energy. In other words, the number of units N together with the topology of the networks strongly affects the thermal properties of the networks, that is, their capacity for absorbing heat from a thermodynamic point of view being the shallow network the topology scheme leading to the poorest capacity for heat absortion.

4.3.2. The Generating Function of Energies

Five network architectures of three hidden layers were selected from table 2 following an increasing order on the total number of units, and using the topological schemes described before (i.e., logical schemes l_1,l_2,l_3,and l_4). Namely, the networks 24x59x41x1 ($N = 125$), 74x49x3x1 ($N = 127$), 30x32x88x1 ($N = 151$), 300x12x2x1 ($N = 315$), and 144x11x171x1 ($N = 327$).

It is important to remember that in this scenario, the expected differences on the number of accesible states are the result of both the total number of artificial neurons, and the differences in the degrees of freedom of neurons comprising the networks as a result of the topology. Furthermore, taking into account that the total number of units N may vary (remember that fixing the effective complexity of a network it may exists a whole number of network architectures with an identical effective complexity but with substantial variations in the topology and the total number of units N) networks with a larger number of neurons are expected to show (in general) a larger number of accesible states because each additional neuron contributes with an extra degree of freedom that comes from its own state (active or not active) independently of the fact that the total number of adaptive weights of the networks is conserved.

$$M_{24\times59\times41\times1}(z) = 1.24\times10^{6297}z^{125} + 1.54\times10^{6299}z^{127} + \cdots + 9.62\times10^{6332}z^{235} + 1.19\times10^{6333}z^{44} + 1.43\times10^{6333}z^{46} + \cdots + 5.39\times10^{6298}z^{373} + 4.28\times10^{6296}z^{375} \quad (50)$$

$$M_{74\times49\times3\times1}(z) = 2.62\times10^{6396}z^{127} + 3.3\times10^{6398}z^{129} + \cdots + 9.32\times10^{6432}z^{239} + 1.15\times10^{6433}z^{241} + 1.38\times10^{6433}z^{243} + \cdots + 1.56\times10^{6398}z^{379} + 1.22\times10^{6396}z^{381} \quad (51)$$

$$M_{30\times32\times88\times1}(z) = 2.38\times10^{6301}z^{151} + 3.56\times10^{6303}z^{153} + \cdots + 1.21\times10^{6345}z^{287} + 1.45\times10^{6345}z^{289} + \cdots + 1.07\times10^{6303}z^{451} + 7.03\times10^{6300}z^{453} \quad (52)$$

$$M_{300\times12\times2\times1}(z) = 6.86\times10^{7275}z^{315} + 2.15\times10^{7278}z^{317} + \cdots + 9.34\times10^{7367}z^{591} + 1.18\times10^{7368}z^{593} + 1.48\times10^{7368}z^{595} + \cdots + 4.61\times10^{7277}z^{943} + 1.46\times10^{7275}z^{945} \quad (53)$$

$$M_{144\times11\times171\times1}(z) = 1.65\times10^{6476}z^{327} + 5.37\times10^{6478}z^{329} + \cdots + 9.69\times10^{6572}z^{639} + 1.05\times10^{6573}z^{641} + 1.13\times10^{6573}z^{643} + \cdots + 2.46\times10^{6478}z^{979} + 7.52\times10^{6475}z^{981} \quad (54)$$

The number of accesible states associated to these networks is obtained evaluating the equations for $z = 1$ giving rise to the values $M_{24\times59\times41\times1}(1) = 3.1\times10^{6334}$, $M_{74\times49\times3\times1}(1) = 3.04\times10^{6434}$, $M_{30\times32\times88\times1}(1) = 3.02\times10^{6346}$, $M_{300\times12\times2\times1}(1) = 2.11\times10^{6886}$, and $M_{144\times11\times171\times1}(1) = 3.04\times10^{6374}$. For the rest of networks shown in table 2, that is, the shallow network 1308x1 and the hierarchical network 162x21x9x1 the the number of accesible microstates are respectively $M_{1308\times1}(1) = 3.04\times10^{8902}$, and $M_{162\times21\times9\times1}(1) = 3.04\times10^{6610}$.

Observing the polynomial expressions of the generating functions it can be deduced that the number of neurons N is affecting (as expected) the energy states. Specifically, as higher is N as higher are the number of energy states in which the neural network can be found. In addition, there is also a shift (or displacement) of the energy states towards higher energy values to the extent N increases. The shallow network is the multiplex network that reaches the highest entropy value, with a total number of $M_{1308\times1}(1) = 3.04\times10^{8902}$ accesible states for $N = 1309$ artificial neurons, a value that is exponentially larger compared to the number of accesible microstates for rest of the deep network architectures considered in table 2.

Moreover, in general the number of accesible states increases to the extent the total number of units of the networks increases. In other words, the degree of disorder of the networks increases with the number of neurons. Similarly, networks with higher entropy values are in general those with a larger number of accesible states. However, as observed in the previous case study, there are several exceptions that are mainly motivated by the topological differences between the selected networks. More specifically, networks with higher entropy values are not necessarily those having

a larger number of units and/or a larger number of accesible states. For example, the hierarchical network 74x49x3x1 has 127 units, that is, a number that is less than the half of the total number of neurons of network 144x11x171x1 ($N = 327$) but possess a higher entropy value, and a larger number of accesible states simply as a result of the strong influence of the topology.

Similarly, the network 30x32x88x1 has a lower entropy value compared to the network 24x59x41x1 in spite of having a larger number of accesible states as well as a larger number of neurons. According to the number of accesible states the expected degree of disorder for the network 30x32x88x1 appears to be higher compared to the network 24x59x41, however, the entropy indicates the opposite, that is, the higher entropy value of the network 24x59x41x1 is indicating a higher degree of disorder compared to the network 30x32x88x1, and at the same time a higher storage capacity of information. The explanation to this aparent paradoxical behavior is that the probabilities of occurrence of the microstates associated to the phase space of network 24x59x41x1 are higher compared to those of network 30x32x88x1, however, in this case the larger number of accesible states of the network 30x32x88x1 is simply due to its larger number of units. Furthermore, the expansive network architecture 30x32x88x1 presents a reduced specific heat value 1.92×10^{52}, that is, 43 orders of magnitud lower compared to the expansive autoencoder architecture 24x59x41x1 in spite of the fact of posessing a larger number of neurons.

Taking into consideration that from the perspective of the model that the process of finding good minimizers during the learning phase of deep learning machines is equivalent to change the probability of occurrence of their accesible microstates. The reduced capacity for heat absortion of the network 30x32x88x1 compared to the hierarchical network 24x59x41x1 is evidencing more difficulties to change the probability of occurrence of its microstates, and thus harder learning phases in general, being this fact independent of both the learning protocol and the learning task. In other words, the typical geometrical complexity of the energy loss landscapes generated by the network 30x32x88x1 during any learning task will be higher compared to those generted by the network 24x59x41x1.

Of particular interest is also the fact that from all the network architectures that are shown in table 2, the shallow network represents the network that attains the lowest value of specific heat in spite of the fact of being the network with the largest number of neurons. According to the theoretical model its reduced capacity for heat absortion is evidencing again the fact that changing the probability of occurrence of the microstates of the phase space of this network is even more difficult compared to the network 30x32x88x1, thereby any learning protocol operating on this network will find more difficulties during the training phase compared to those presented by any other network of table 2.

Summarizing, the resulting degree of disorder of networks under identical effective complexity is the consequence of the interplay between the total number of neurons and the topology of the network. In other words, as occurred in the previous scenario the topology of the networks is a not only a determinant factor in the structure of the phase space of the networks but also it appears to be reponsible of the geometrical complexity of the loss landscape being this fact independent of the dataset and/or the learning algorithm used.

4.3.3. Infuence of the Structural Parameters

Given the restrictions imposed by this scenario only two structural parameters are considered in this case. Namely, the total number of units N, and the depth of the networks. The input space dimension is not considered since the results are identical to those obtained in scenario I. It is important to remember that depending on the topology of the network fixing the total number of adaptive weights (i.e., the effective complexity) may lead to networks with substantial differences on the total number of neurons.

The graphs of figure (8) show the evolution of the entropy and the free energy for deep network architectures from three up to seven hidden layers with a hierarchical topology with an effective complexity $W = 3924$, and for a number of neurons comprised within the integer interval [125..350]. Specifically, the diophantine equation (43) was numerically solved in the aforementioned integer interval assuming an input space dimension $g_0 = 2$, and an effective complexity $W = 3924$ searching

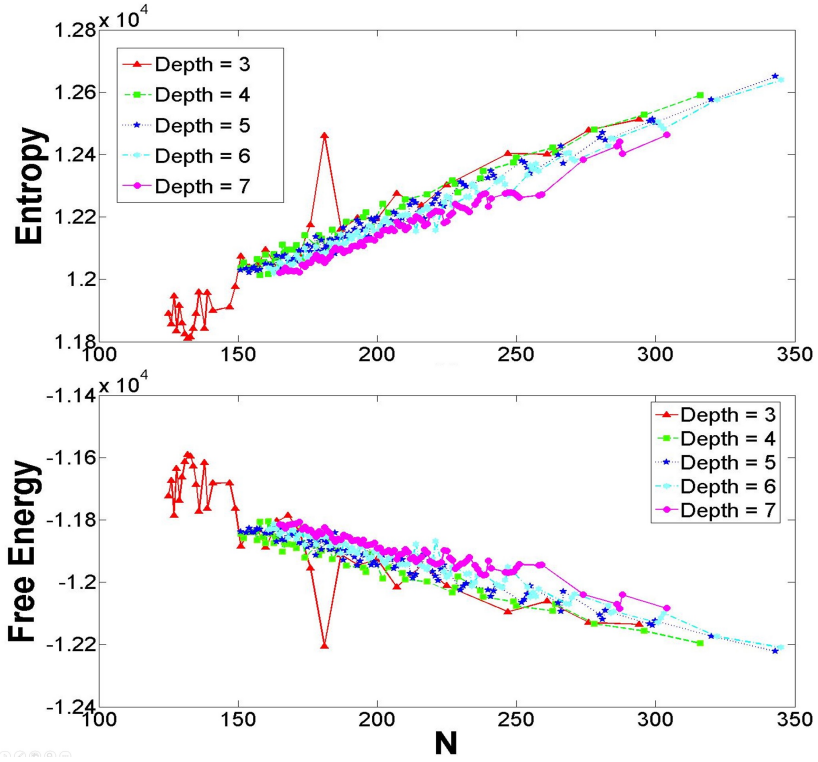

Figure 8. Graphical illustration of the evolution of the entropy and the free energy for deep networks from three hidden layers up to seven for a number of neurons comprised within the integer interval [125..350]. Specifically, the diophantine equation (43) was numerically solved for networks with a total number of units defined within the aforementioned interval and for networks with a depth ranging from 3 up to seven hidden layers and only selecting those with a hierarchical structure. The most important characteristic of this graph is that the entropy of the networks decreases to the extent the number of hidden layers increases, whereas the free energy of the networks increases with the depth. Independently of the depth, the entropy values increases to the extent the number of units of the networks increases whereas the free energy decreases, however, such an increase or decrease in case of the free energy is not monotonic due to the strong influence of the topology. Therefore, when conserving the efective complexity of networks their degree of disorder, and also their storage capacity decreases to the extent their number of hidden layers increases.

for solutions with a depth ranging from three up to seven hidden layers having a hierarchical structure. Similarly, the graphs of figure (9) show the evolution of the internal energy together with the logarithm of the specific heat under identical conditions to those defined before for the entropy and the free energy graphs.

With regards figure (8) the most important characteristic of these graphs is that the entropy of the networks decreases to the extent the number of hidden layers increases, whereas the free energy of the networks increases with the depth. Independently of the depth, the entropy values increases to the extent the number of units of the networks increases whereas the free energy decreases, however, such an increase or decrease in case of the free energy is not monotonic due to the strong influence of the topology. Therefore, when conserving the effective complexity of networks their degree of disorder, and also their storage capacity decreases to the extent their number of hidden layers increases.

Of particular interest are the oscillations observed in figure (9) for both the internal energy and the specific heat when increasing neuron numbers. Specifically, the oscillations of the internal energy appear to decrease when the depth of the networks increases. The internal energy increases with neuron numbers although not monotonically principally due to the strong influence of the topology, especially due to the effect of the number of units in the first hidden layer of the network in hierarchical topologies. In average the values of the internal energy are higher for networks with a lower number of layers even when their structural complexity is identical (identical number of units and adaptive weights). The specific heat present oscillations that are higher to the extent the depth of the networks increases. Fixing the depth of the networks the amplitude of the oscillations decreases when increasing neuron numbers, although when considering neuron numbers, the period of oscillations is larger to the extent the number of hidden layers increases. From a thermodynamics point of view these oscillations are related to changes in the thermal properties of the equilibrium states represented by the networks, and thus evidencing changes with regards to their expected learning and generalization capabilities.

Having said this, to further investigate the growth of the entropy with neuron numbers using the data of figure (8), a regression fit was performed for the density of entropy in logarithmic scale (see figure (10)) with respect to the total number of neurons for hierarchical networks of 3 hidden layers with two inputs and identical effective complexity ($W = 3924$) for a total number of neurons ranging from 124 up to 566.

The regression fit evidence the existence of a power law. This fact is not surprising since the appearance of the power law distributions is common in complex networks and evidences the tendency of social, technological, and especially biological networks toward "ordering". The whole thing is that this tendency is at work regardless of the mechanism that is driving their evolution [10]. With regards to deep learning machines this fact is equivalent to state that increasing the depth in hierarchical networks with identical effective complexity leads to higher ordered thermodynamic states being this fact independently of the learning protocol and/or the dataset used for training these models.

Moreover, the entropy density grows as $S \cong S_0 N^{-x}$, with $x = 0.916 \pm 0.004$, and $S_0 = 9700 \pm 200$. This graph shows that the entropy density decreases practically as $1/N$, and that is the reason why entropy increases smoothly with neuron numbers (the entropy grows with a rate slightly smaller than $N^{0.1}$). The explanation of this behavior is simple if one takes into account that networks possess an identical effective complexity thus, when increasing neuron numbers, the observed differences on the number of degrees of freedom that appear between different network architectures are the result of the difference on neuron numbers, and each neuron only contributes with a single degree of freedom,that is, its state (active or not active).

Summarizing, when conserving their effective complexity of networks increasing neuron numbers leads to a an increase of the entropy and to the energy stored by the networks although such a growth is not monotonically due to the strong infuence of the topology. Furthermore, for hierarchical networks, increasing the number of hidden layers leads to a progressive reduction of the entropy that is implicitly evidencing a tendence towards ordering of the equilibrium states represented by the networks, and thus to a reduction of their information-storage capacity. Furthermore, this behavior is also accompanied by

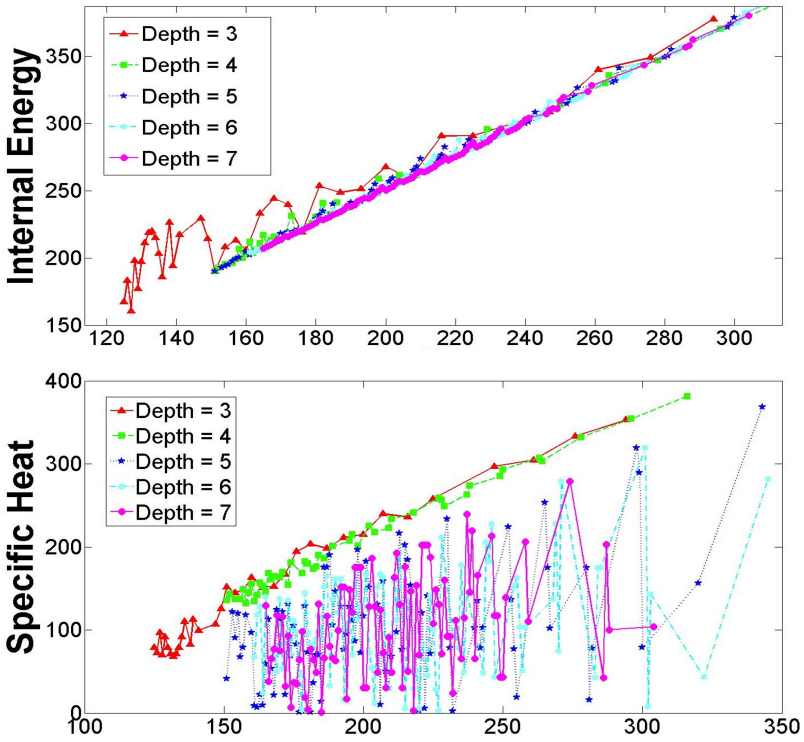

Figure 9. Graphical illustration of the evolution of the internal energy and the specific heat for deep networks from three hidden layers up to seven for a number of neurons comprised within the integer interval [125..350]. Specifically, the diophantine equation (43) was numerically solved for networks with a total number of units defined within the aforementioned interval and for networks with a depth ranging from 3 up to seven hidden layers. Afterwards, only networks with a hierarchical structure were selected. The most important characteristic of this graph are the oscillations that are observed for both the internal energy and the specific heat when increasing neuron numbers. In particular, the oscillations of the internal energy appear to decrease when the depth of the networks increases. The internal energy increases with neuron numbers although not monotonically principally due to the strong influence of the topology. In average the values of the internal energy are higher for networks with a lower number of layers even when their structural complexity is identical (identical number of units and adaptive weights). The specific heat present oscillations that are higher to the extent the depth of the networks increases. Fixing the depth of the networks the amplitude of the oscillations decreases when increasing neuron numbers, although when considering neuron numbers, the period of oscillations is larger to the extent the number of hidden layers increases.

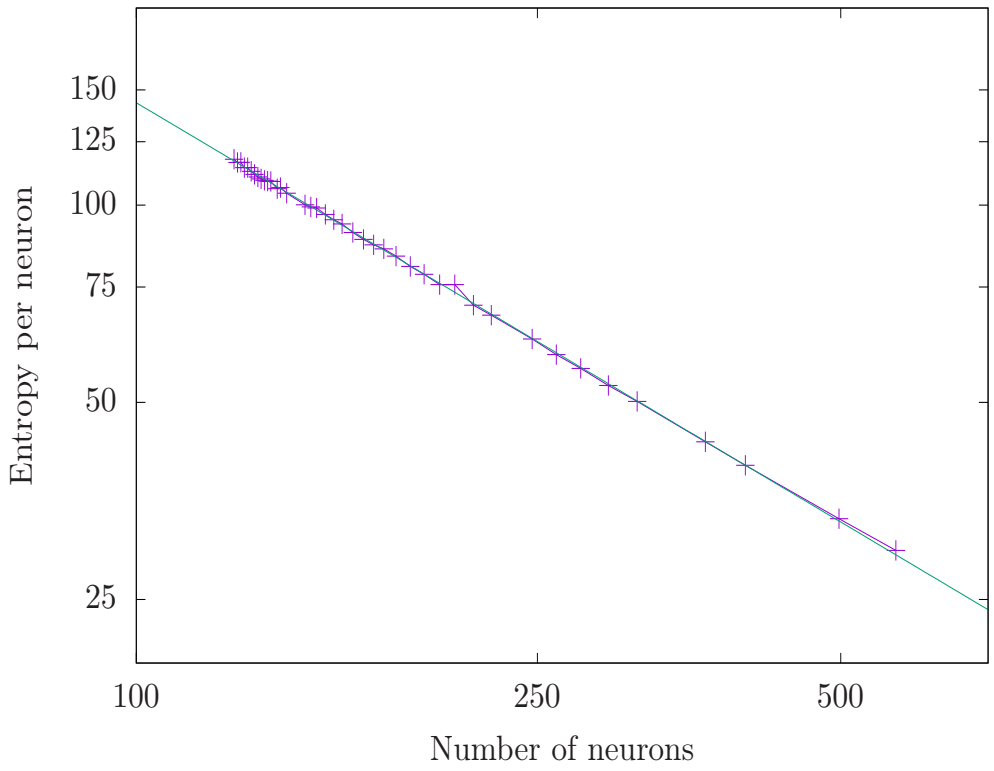

Figure 10. Regression fit of the density of entropy (i.e., entropy per neuron) in logarithmic scale with respect to the total number of neurons for hierarchical networks of 3 hidden layers with identical effective complexity, and a total number of neurons ranging from 124 up to 566. The regression fit shows a power law. Specifically, entropy density grows as $S \cong S_0 N^{-x}$, with $x = 0.916 \pm 0.004$, and $S_0 = 9700 \pm 200$. This graph shows that the entropy density decreases practically as $1/N$, and that is the reason why entropy increases smoothly with neuron numbers (the entropy growth with a rate slightly smaller than $N^{0.1}$). The explanation is simple if one takes into account that networks possess an identical effective complexity thus, when increasing neuron numbers, the extra degrees of freedom between networks comes only from the difference on neuron numbers, and each neuron only contributes with a single degree of freedom,that is, its state (active or not active).

a successive decrease of both the heat capacity and the energy stored by the networks, thus evidencing changes with regards to their expected learning and generalization capabilities.

4.3.4. Thermodynamic Interpretation

Aimed at understanding the macroscopic thermodynamic behavior of the model an identical theoretical procedure of that described in section 4.2.4 is assumed hereafter but with the particularity that now the chemical potential, that is, the total number of neurons (the particles) is permitted to vary while conserving the effective complexity of the networks. Specifically, the existence of an hypothetical process (or external force) that perturbates the equilibrium states represented by the the network architectures of table 2 by making changes in their topology so as to pass from one multiplex network architecture to any other is assumed. It is important to remember that the whole set of architectures that appear in table 2 possess an identical effective complexity. This theoretical scenario is adapted to the grand canonical ensemble which corresponds to a situation where the physical system is in contact with a heat reservoir and a particle reservoir, that is, two coupled systems separated by a porous partition through which energy and particles (artificial neurons in this case) can be exchanged, the combined system being isolated from its surroundings. An thermodynamic example of such a situation is water molecules under two phases, liquid and gas, contained in a fixed volume, the water molecules belonging to either phase [31]. Accordingly, the concept of pressure may also be defined in this context but taking into consideration that the number of neurons (i.e.,particles) varies from one equilibrium state to another. Thus, the thermodynamic identity used in section 4.2.3 reads now [34]: $\Delta U = T\Delta S - P\Delta V + \Delta N$ relating now energy, entropy, pressure, temperature, particles, and volume. Then, equation (49) reads now as:

$$P = -\left(\frac{T\Delta S - \Delta U + \Delta N}{\Delta V}\right) \tag{55}$$

Independently of this fact, the main goal is to understand the role of the topology in the structure of the phase space of the networks (i.e., the degree of disorder, the stability and thermal properties of the equilibrium state represented by the networks, and its information storage capacity) but specially the influence of the total number of units N in the resulting level of disorder of the networks when its effective complexity is conserved.

In this case each of the multiplex networks networks considered in this section (i.e., those of table 2) correspond to the equilibrium states of a physical system defined by the set of all feedforward fully-connected artificial neural networks with two inputs (i.e., $g_0 = 2$), and an effective complexity $W = 3924$. The equilibrium states represented by the networks of table 2 are denoted following the cardinality given by the value attained by the entropy: 30x32x88x1 (State 1 with $N = 151$), 24x59x41x1 (State 2 with $N = 125$), 144x11x171x1 (State 3 with $N = 327$), 74x49x3x1 (State 4 with $N = 127$), 162x21x9x1 (State 5 with $N = 193$), 300x12x2x1 (State 6 with $N = 315$), and 1308x1 (State 7 with $N = 1309$) are used to understand (under the restrictions imposed by this scenario) the thermodynamic behavior of the model. Furthermore, given the theoretical nature of the transformations, the evolution of the system from one state to another is assumed again that occurs according to the law of increase of the entropy.

Having said this, it is important to note that when conserving the total number of adaptive weights W, the degrees of freedom of the system vary at the different equilibrium states that the system may visit since the topology of the networks and/or the neuron numbers are changing when the system passes from one equilibrium state to another. The restrictions imposed by this scenario together with the fact that the transformations of state are drawn according to the law of increase of entropy lead to the fact that the accesible volume of the phase space gets expanded or compressed depending on the particularities of the topology together with the difference on neuron numbers associated to the equilibrium states visited by the system. However, it is important to remember that when the effective complexity of the networks is conserved, and ignoring the infuence of the topology, the difference

on the number of degrees of freedom when passing from one equilibrium state to another comes only from the difference on neuron numbers, and each neuron only contributes with a single degree of freedom,that is, its state (active or not active), that is, the kind of degrees of freedom that are not directly configurable by the learning protocol.

Assuming that the physical system is found in the equilibrium state represented by the multiplex network 30x32x88x1 (this state is denoted herafter as state 1). The system is then perturbated by removing 26 neurons and rearranging the resulting neurons and connections (i.e., by doing work on the system) passing to the equilibrium state represented the multiplex network 24x59x41x1. Let us denote this state as state 2. Clearly, the result of this transformation is a reduction of the internal energy of the body (i.e., the neural network) since $\Delta U = U_2 - U_1 = -106.5$, the entropy of the system increases $\Delta S = S_2 - S_1 = 105.2$, but at the same time, the free energy of the system decreases $\Delta F = F_2 - F_1 = -211.7$. Indeed, there is a reduction of 12 orders of magnitude on the number of accesible states of the equilibrium state represented by the network 24x59x41x1 (remember the calculation performed in the previous section), that is, the system experiences a compression of the accesible volume of the phase space.

Therefore, the loss on the number of accesible states is in part motivated by the loss on the total number of units, that is, a difference of 26 units (i.e., $\Delta N = N_2 - N_1 = -26$) from one state to the other when passing from state 1 to the state 2. It is important to remember that in this scenario (i.e., the effective complexity of the networks is conserved) there is a correlation between the total number of units N of a network, the number of accesible states, and the resulting entropy value, that is, as higher is N as higher will be the entropy because each additional neuron is contributing with an extra degree of freedom that comes from its own state. Surprisingly, the entropy increases in spite of the fact of the loss of 26 neurons accompanied with the contraction of the accesible volume of the phase space. The increase of the entropy is a clear indicator that the system is passing to a state with a higher degree of disorder but also of a higher capacity.

Of particular interest is also the fact that the specific heat passes from $C_v = 1.92 \times 10^{52}$ to $C_v = 1.62 \times 10^{95}$ in spite of a reduction on neuron numbers, thereby indicating a radical change in the thermal properties of the body since the specific heat is an extensive magnitude, that is, its grows with neuron numbers (remember that the specific heat is calculated as the derivative of the entropy with respect to the temperature, being the entropy an extensive magnitude) but surprisingly, as stated before, in this transformation of state there is a reduction on neuron numbers. Particularly, the larger value of the specific heat of network 24x59x41x1 compared to the network 30x32x88x1 is evidencing its higher capacity for heat absorption, and this fact is hypothesized to be responsible of better learning and generalization performance. With regards the stability of this transformation $T_{1\to 2}$, using equation (55) it can be verified that the equilibrium state represented by the expansive autoencoder architecture 24x59x41x1 is a metastable state since $\frac{(S_2-S_1)-(U_2-U_1)+(N_2-N_1)}{M(1)_{24\times 59\times 41\times 1}-M(1)_{30\times 32\times 88\times 1}} < 0$. Summarizing, this analysis appears to corroborate the idea suggested before concerning that the probabilities of occurrence of the microstates. Specifically, the probabilities of occurrence of the microstates associated to the phase space of network 24x59x41x1 are higher compared to those of network 30x32x88x1.

Let us assume now that the system is found in the equilibrium state represented by the network 24x59x41x1 (state 2). Afterwards the system is pertubated again by adding 202 neurons and rearranging the resulting neurons and connections so as to pass to the equilibrium state represented by the compressive autoencoder network 144x11x171x1 (state 3).

As expected in this case, the accesible volume of the phase space gets expanded principally because of the increase on neuron numbers (i.e., $\Delta N = 202$) Specifically, the number of accesible states increases from $M_{24\times 59\times 41\times 1}(1) = 3.1 \times 10^{6334}$ to $M_{144\times 11\times 171\times 1}(1) = 3.04 \times 10^{6434}$, that is, a gain of 40 orders of magnitude. This fact is of particular interest since it is evidencing once again the strong influence of the topology with regards the structure of the phase space of the system (i.e., a gain of 202 neurons leads to an expansion of the accesible volume of the phase space of 40 orders of magnitude whereas a loss of 26 neuron leads to a compression of only 12 orders of magnitude).

Similarly, the result of this transformation is an increase of the internal energy of the body (i.e., the neural network) since $\Delta U = U_3 - U_2 = 456 \pm 13.79$ which that is mainly motivated by the interplay of the change in the topology accompanied by the increase on neuron numbers. The gain in entropy obtained is $\Delta S = S_3 - S_2 = 142.1$, and the free energy increases as $\Delta F = F_3 - F_2 = 313.7$ leading to a positive value. A positive variation of the free energy $\Delta F > 0$ is indicating that work is positive, that is, the change of state from the equilibrium state represented by the expansive autoencoder network 24x59x41x1 to the compressive autoencoder network is only possible (as opposed to the previous transformation) if an external force performs work on the system. More specifically, the first law of thermodynamics states that there is a store of energy in the system, called the internal energy which can be changed by causing the system to do work, or by adding heat to the system. Under the hypothetical transformations studied here no heat is added to the system, thus the work is due to changes in any of the relevant extensive variable, that is, the "volume" (i.e., the number of accesible microstates) and the number of neurons in this case.

To further explore this issue let us assume now that the system is found in the equilibrium state represented by the network 144x11x171x1 (state 3). Afterwards the system is pertubated again by rearranging neurons and connections so as to pass to the equilibrium state represented by the multiplex network 74x49x3x1 (state 4).

Surprinsingly, in this case the accesible volume of the phase space gets expanded, that is the accesible number of microstates increases exponentially by 60 orders of magnitude, passing from a number of accesible states $M(1)_{144\times11\times171\times1} = 3.04 \times 10^{6374}$ to $M(1)_{74\times49\times3\times1} = 3.04 \times 10^{6434}$, in spite of the fact of a larger decrease on the number of neurons (i.e., $\Delta N = -200$). The result of this transformation is a reduction of the internal energy of the body (i.e., the neural network) since $\Delta U = U_4 - U_3 = -513.4 \pm 13.83$ which is mainly motivated by the interplay of the change in the topology accompanied by the reduction on the neuron numbers, the gain in entropy obtained is $\Delta S = S_4 - S_3 = 188.1$, but at the same time, the free energy of the system decreases $\Delta F = F_4 - F_3 = -701.5$. The specific heat of the body increases passing from a value $C_v = 6.5 \times 10^{230}$ to $C_v = 2.08 \times 10^{119}$, that is, an exponential loss of 111 orders of magnitude, indicating a change in the thermal properties of the body.

However, in this case it is not possible to precisely assess the impact of this change in the learning phase of both networks since the change in the thermal properties of the system are masked by the fact that both the specific heat, and the neuron numbers decrease in this transformation. Specifically, the smaller value of the specific heat for network 74x49x3x1 compared to the network 144x11x171x1 is evidencing its lower capacity for heat absortion but at the same time is the result of having a larger number of neurons (remember that the specific heat is an extensive magnitude). Furthermore, using equation (55) it can be verified that the equilibrium state represented by the hierarchical network 74x49x3x1 ($N = 127$) is a stable state. Finally, in order to analyze the behavior of the system when the transformation conserves the topological scheme, the equilibrium state represented by the multiplex network 162x21x9x1 (state 5) was chosen together with those of the hierarchical network 300x12x2x1 (state 5) and the shallow network 1308x1 (state 6).

Thus, if now it is assumed that the system is found in the equilibrium state represented by the network 162x21x9x1 (State 5) and passes to the equilibrium state represented by the network 300x11x2x1 (State 6).The result of this transformation is an increase of the internal energy of the body (i.e., the neural network) since $\Delta U = U_6 - U_5 = 138.7$, the entropy of the system increases $\Delta S = S_6 - S_5 = 614.5$, but at the same time, the free energy of the system decreases $\Delta F = F_6 - F_5 = -475.8$, that is, the work is performed by the system. The specific heat of the body increases passing from a value $C_v = 8.54 \times 10^{259}$ to $C_v = 5.93 \times 10^{480}$, that is, an exponential gain of 221 orders of magnitude, indicating a change in the thermal properties of the body. However, as occurred in the transformation $T_{3\to4}$ it is not possible to precisely assess the impact of this change in the learning phase of both networks since the change in the thermal properties of the system are masked by the fact that both the specific heat, and the neuron numbers increase in this transformation. Specifically, the larger value

of the specific heat for network 300x11x2x1 compared to the network 162x21x9x1 is evidencing its higher capacity for heat absortion but at the same time is the result of having a larger number of neurons (remember that the specific heat is an extensive magnitude). The accesible volume of the phase space expands as a result of the increase in neuron numbers (there is a gain of $\Delta N = 121$ neurons in this transformation) passing from a number of accesible states $M(1)_{162\times21\times9\times1} = 3.04 \times 10^{6610}$ to $M(1)_{300\times11\times2\times1} = 3.044 \times 10^{6886}$, that is, a number that is 276 orders of magnitude larger. Furthermore, using equation (55) it can be verified that the equilibrium state represented by the hierarchical network 300x11x2x1 ($N = 314$) is a stable state.

Finally, the transformation $T_{6\rightarrow7}$ is analyzed hereafter because the thermodynamic conditions that take place are opposed to those found in transformation $T_{1\rightarrow2}$. Particularly, the heat capacity of the system decreases in spite of the fact there is a substantial increase on neuron numbers (i.e., $\Delta N = 995$). Thus, if it is assumed that the system is found in the equilibrium state represented by the network 300x11x2x1 (State 6) and passes to the equilibrium state represented by the shallow network 1308x1 (State 7). The result of this transformation is a substantial increase of the internal energy since $\Delta U = U_7 - U_6 = 3297.2 \pm 26.06$, the entropy of the system increases $\Delta S = S_7 - S_6 = 858.1$, but at the same time, the free energy of the system increases $\Delta F = F_7 - F_6 = 2439.1$ to a positive value as occurred in the transformation $T_{2\rightarrow3}$, and thus with an identical interpretation. In this transformation the specific heat decreases exponentially in spite of the substantial increase on neuron numbers passing from a value $C_v = 5.93 \times 10^{480}$ to $C_v = 2.49 \times 10^{7}$, that is, an exponential loss of 221 orders of magnitude, indicating a radical change in the thermal properties of the body. Specifically, a substantial reduction on the capacity for heat absortion of the shallow network compared to the hierarchical network 300x11x2x1 . Thus, according to the reasonings presented before those facts are hypothesized to be responsible in average of a worst learning and generalization performance of the former independently of the learning protocol and/or the learning task. Moreover, the accesible volume of the phase space expands as a result of the increase in neuron numbers (there is a gain of $\Delta N = 995$ neurons in this transformation) passing from a number of accesible states $M(1)_{300\times11\times2\times1} = 3.044 \times 10^{6886}$ to $M(1)_{1308\times1} = 3 \times 10^{8902}$, that is, a number that is 2016 orders of magnitude larger. Furthermore, using equation (55) it can be verified that the equilibrium state represented by the shallow network 1308x1 ($N = 1309$) is also a metastable state.

Having said this, it is important to remember that as opposed to the first case study, when conserving the effective complexity of networks the degrees of freedom of the system change when the system passes from one equilibrium state to another principally as a result of the difference on neuron numbers between both equilibrium states. More specifically, when the system passes from one equilibrium state to another the accesible volume of the phase space of the system gets contracted or expanded not only as a result of a change of the topology of the network but also as a result of the difference on the number of neurons between those states. The most important conclusions of this section can be summarized as follows:

1. The observed differences in the resulting level of disorder of the system when passing from one equilibrium state to another, and thus the differences in its information-storage capacity are explained as a result of the interplay between total number of units and the restrictions imposed by the topology of the networks (the external field), being these two factors also responsible of the structure of the phase space of the system together with the resulting thermal properties of those equilibrium states.
2. The observed changes in the thermodynamic properties of the system during the transformations of state according to the law of increase of the entropy permit to compare the expected learning and generalization properties of the networks representing those equilibrium states. However, the impact of these changes in their learning and generalization performance cannot be precisely assessed for those trasformations leading to an increase (decrease) of the specific heat also accompanied of an increase (decrease) on neuron numbers since it is masked by the fact that the specific heat is an extensive magnitude.

3. Metastable states appear to be linked to those transformations involving a positive increase of the free energy ($\Delta F > 0$), or a compression of the accesible volume of the phase space being this fact independent of the change on neuron numbers from one equilibrium state to another.

4.3.5. Experimental Results

The dataset used for training *Clouds.dat* is an artificial database [12] containing 5000 sample data points distributed in a two dimensional space. The samples distribution for this artificial database is Gaussian. There are 5000 patterns, 2500 in each class. Class 0 is the sum of three different Gaussian distribution, Class 1 is a single Gaussian distribution. The aim of this database is the study of the machine learning classifier behavior for heavy intersection of the class distributions and for high degree of nonlinearity of the class boundaries. There is an important overlapping between the two classes. The theoretical error is 9.66 % (Bayes error). The set of networks used in this benchmark are under-parameterized (i.e., $W < 5000$) for this dataset. Furthermore, the procedure followed to validate the models is exactly the same used in section 4.3, that is, a ten fold crossvalidation procedure using ten averages per fold [25].

Each figure is composed as before of two parts. The top part represents the evolution of the generalization performance of the network that is typically sampled at 10, 50,100,200, 1000 and 5000 epochs arriving in some cases up to 20000 epochs to show overfitting effects. The botton part represents again the average generalization (and its standard deviation that is represented as an error bar) obtained when averaging the generalization at the sampling points described before. The qualifier speed of learning is used as in the previous case study to define the minimal number of epochs needed to attain a predefined learning and/or generalization error.

When comparing the generalization performance obtained with Backpropagation, and with the scaled conjugate gradients algorithm (figures (11) and (12)) the worst results are obtained for the compressive autoencoder network architecture 144x11x171x1, and the shallow network 1308x1. Furthermore, the shallow network 1308x1 represents the equilibrium state attaining the largest entropy value, but at the same time the worst generalization performance compared to the rest of networks. Surprisingly, the equilibrium states represented by these networks are metastable (remember the results obtained in section 4.3.4), and at the same time they constitutes the networks storing the largest amount of energy.

The effective complexity of the networks is identical, thus, the larger amount of energy stored by these networks compared to the rest of architectures in this group is the result of the interplay between the topology and the number of units N. Particularly, they are both the networks with the largest amount of neurons N in this case. It is important to remember that the entropy (and also the internal energy) grows with neuron numbers, however, the values attained are strongly linked to the topology of the networks. For example, the entropy attained by the hierarchical network 162x21x9x1 surpass the value attained by the compressive autoencoder architecture 144x11x171x1 in spite of possessing almost twice the neurons of the former. Similarly, with fewer neurons the entropy attained by expansive autoencoder architecture 24x59x41x1 is larger compared to the network 30x32x88x1.

Of particular interest is the fact that for standard gradient descent algorithm the compressive autoencoder network 144x11x171x1 within the sampling range considered (from 10 up to 5000 epochs) start to overfit after 200 epochs approximately, whereas the rest of networks do not show any overfitting effect. The learning and generalization improves with the scaled conjugate gradients algorithm as occurred in section 4.2 with the topological-equivalent architecture 108x50x108x1. However, in this case the generalization performance of this network is unable to outperform that of network 24x59x41x1 in spite of attaining a larger entropy value. This behavior is evidencing the strong influence of the internal energy in the complexity of the energetic landscape that faces backpropagation and the scaled conjugate gradients algorithms (i.e., the learning protocols) independently of the characteristics of the learning task but especially the limited stability of this equilibrium state. The hierarchical network 162x21x9x1 together with the expansive autoencoder architecture 24x59x41x1 are those networks presenting the

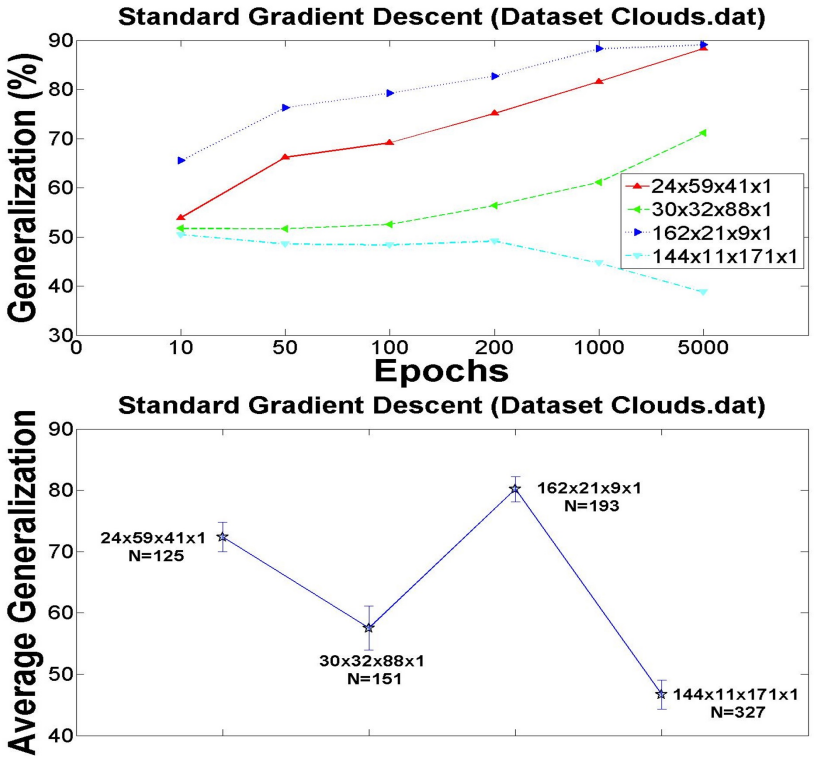

Figure 11. Graphical illustration of the generalization performance obtained for the group of networksM_3 with an identical effective complexity (see table 2) using as learning algorithm backpropagation with the dataset *Clouds.dat* [12]. The top part represents the evolution of the generalization performance of the networks sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The botton part represents the average generalization (and its standard deviation represented as an error bar) obtained when averaging the generalization at the sampling points described before using a ten fold cross-validation procedure. The goal is to assess the typical behavior of the generalization independently of the number of epochs used to train the networks. Networks storing larger amounts of energy (i.e., 30x32x88x1x1 and 144x11x171x1) are those presenting the worst learning and generalization performance for standard gradient descent. The generalization performance improves with the value of the entropy attained by the networks (see table 2) excepting for the compressive autoencoder architecture 144x11x171x1, perhaps evidencing the existence of a more complex geometrical structure of the loss landscape [69]. However, according to the model the equilibrium state represented by the aforementioned network correspond to a metastable state.

best generalization performance and also a better speed of convergence independently of the learning algorithm considered. Surprisingly, the energy stored by these networks is smaller compared to the rest, but in spite of this fact the equilibrium state represented by the network 24x59x41x1 is also a metastable state. Thereby, evidencing again the strong influence of the internal energy in the complexity of the generated energy loss landscapes from a learning point of view.

Having said this, it is important to remember that those networks attaining higher entropy values have the potential to provide better learning and generalization performance since as higher is the entropy as higher will be the complexity of the space of functions they are able to represent, and thus its VC dimension. Furthermore, when passing from one equilibrium state to another the entropy increases because the transformations of state studied in section 4.3.4 are drawn according to the law of increase of the entropy. Thus, the theoretical transformation of state evidenced the intuitive plausibility that generalization tends to improve to the extent the entropy increases.

However, the increase of the entropy when passing from one equilibrium state to another is always linked to a change in the structure of the phase space of the system motivated by the change in the topology and/or the structural parameters such as the number of hidden layers, the dimension of the input space to the networks, and the number of units N associated to the particularities of the equilibrium states involved in the transformation, all of these facts affecting the thermodynamic properties and stability of the equilibrium states of the system.

4.3.6. Summary of Results

When conserving the effective complexity of the networks the principal conclusions that can be extracted when combining both the theoretical and experimental results can be synthesized as follows:

1. Quantitatively networks leading to better generalization performance are those with larger information-storage capacities, however, a larger capacity (i.e., a higher entropy value) is a neccessary but not sufficient condition for a network to provide better generalization performance because of the strong influence of the internal energy in the generalization performance when conserving the effective complexity of the networks. From the whole set of network topologies, deep hierarchical networks are those leading to both: higher entropy values, and also to the lowest internal energy values, therefore leading to better generalization performance.
2. As larger is the energy stored by a network as complex will be the geometrical complexity of the energy loss landscape (i.e., the landscape of the empirical risk) generated when facing a learning task being this fact independent of the convergence properties of the learning algorihm used and/or the particularities of the learning task. Furthermore, the amount of energy stored by a network limits the speed of learning, and is quantitatively correlated with the expected deviations of the generalization performance from its typical value. Shallow networks constitute the network topology that stores the largest amount of energy and possess the largest amount of neurons because of the restrictions imposed when conserving the effective complexity of networks, thus typically leading to the generation of more complex geometrical structures of the loss landscape, and thus harder from a learning point of view compared to any other network topology.
3. Compared to shallow networks, when approximating a function in high dimensions deep networks are able to achieve a rate of convergence independent of the input dimensionality g_0 since they are able provide increasing smoothness of the unknown underlying function with increasing g_0 because the energy stored by shallow or deep learning machines is independent of their number of inputs (i.e., the dimension of the input space) but increasing its depth leads to a progressive reduction of the energy stored by the networks accompanied with a reduction of the level of disorder.
4. Keeping the effective complexity of networks while increasing neuron numbers leads to networks with larger entropy and internal energy values. This fact may lead to equilibrium states with limited stability and/or with more complex geometrical structures of the loss landscapes from a

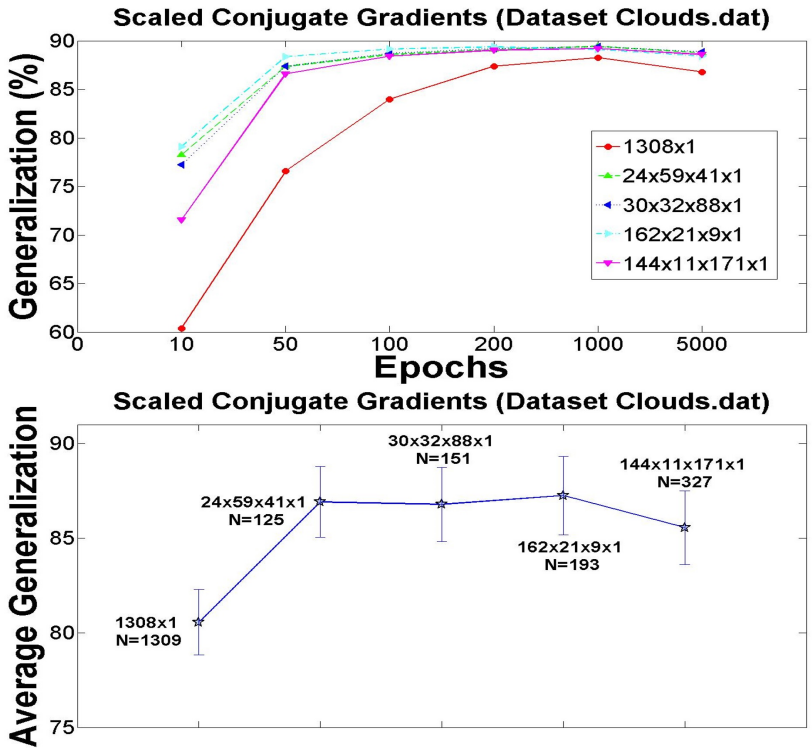

Figure 12. Graphical illustration of the generalization performance obtained using as learning protocol the scaled conjugate gradients algorithm [11,30] with the dataset *Clouds.dat* [12] for the group of networks used in figure (12) but including now the shallow network 1308x1. The entire set of networks present an identical effective complexity. The top part represents the evolution of the generalization performance of the networks. The botton part represents the average generalization (and its standard deviation represented as an error bar) following in both cases an identical procedure as described in figure (12). Quantitatively networks presenting better generalization performance are those storing smaller amounts of energy (i.e., 24x59x41x1 and 162x21x9x1), and attaining larger entropy and specific heat values. Conversely, the worst generalization performance results correspond to those networks storing larger amounts of energy and attaining lower specific heat values, also corresponding in this case to equlibrium states that are stable within certain limits (metastable states). In other words, the geometrical complexity of the loss landscape associated to the compressive autoencoder network 144x11x171x1 and the shallow network 1308x1 combined with the limited stability of these equilibrium states impairs their generalization performance in spite of the fact of being states attaining relatively larger entropy values and/or the fact of using a learning protocol with better convergence properties.

learning point of view as a result of the increased amount of energy stored by these networks. Metastable states are equilibrium states that are characterized by storing larger amounts of energy compared to stable equilibrium states. Furthermore, learning and generalization performance tends to degrade in metastable states.

4.4. Controling the Number of Artificial Neurons: Case Study III

This section is aimed at studying the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers) when controling the total number of neurons comprising their structure using the thermodynamic formalism derived in section 3. As occurred in the previous case study, the influence of the input space dimension is not presented since the results are identical to those obtained in case study I (i.e., the internal energy does not depend on the number of inputs to the networks).

In this case the effective complexity of the networks plays the role of a variable parameter. Depending on the topology of the network fixing the total number of neurons may lead to networks with substantial differences on the total number of adaptive weights.

4.4.1. Network Architectures: A Preliminary Analysis

A measure of the disorder of a physical system is the number of accesible states or the size of the accesible volume of the phase space [31]. Furthermore, the number of accesible states is related to both the number of particles and to the number of degrees of freedom of each particle, that is, artificial neurons in our case.

In this case study the total number of neurons (i.e., the number of particles) is fixed, particularly to $N = 101$ for the networks of table III. The equilibrium states of the physical sytem under consideration are those network architectures with a total number of neurons equal to $N = 101$, with a single neuron in the output layer, an input space dimension $g_0 = 8$, and any arbitrary topology respecting the aforementioned restrictions. Thus, differences on the total number of accessible states between network architectures (i,e, the equilibrium states of the system) are the result of differences in their total number of adaptive weights, that is, the kind of degrees of freedom that are directly configurable by the learning algorithm.

A total set of eleven network architectures ranging from one up to six hidden layers were selected for the study. Specifically, one multiplex network M_1 with architecture 100x1 (i.e., a shallow network), three multiplex networks M_2 (deep networks with two hidden layers) following the possible topologies for this depth ($g_1 > g_2$ and $g_1 < g_2$) of the networks with architectures 10x90x1, 90x10x1, and 55x45x1 respectively. In this particular case, to gain further insights concerning the properties of hierarchical networks the total number of nodes in the hidden layers ($g_1 + g_2$) was parameterized $g_1 = \alpha g$ and $g_2 = (1 - \alpha)g$, where $0 < \alpha < 1$ so as to obtain the hierarchical network with maximum entropy resulting in the network architecture 55x45x1.

Four more multiplex networks M_3 with architectures 47x6x47x1, 10x20x70x1, 20x60x20x1, and 70x20x10x1 corresponding to the topologies (i.e., logical schemes l_1, l_2, l_3, and l_4) described in the previous section for networks with three hidden layers, and finally, three multiplex networks M_4, M_5, and M_6, with architectures 40x30x20x10x1, 35x30x20x10x5x1, and 35x25x20x10x8x2x1 respectively following a hierarchical structure to study the influence of the depth of the networks for the most commonly used topology in deep learning applications.

Table 3 shows the particular values of the thermodynamic potentials (entropy S, free energy F, the internal energy U, and its fluctuactions ΔU), the specific heat C_v as well as the total number of adaptive weights associated to each network (denoted as W in the table) for the aforementioned set of networks for $\beta = 1$, $\mu_m = 20$, and assuming an eight dimensional input space ($g_0 = 8$).

Table 3. The set of network architectures selected for the study of the effect of the topology of the networks, its depth (measured in terms of the number of hidden layers), and the effect of the input space dimension to the networks when controling the total number of neurons comprising their structure. The table shows the thermodynamic potentials associated to each of the networks considered (i.e., , entropy S, free energy F, internal energy and its fluctuations $U \pm \Delta U$) , the specific heat C_v, the total number of adaptive weights W for an input space dimension $g_0 = 8$, and the total number of units N

Multiplex M_1	**S**	**F**	**U** $\pm\Delta U$	C_v	W	N
100x1	4864.5	-4582.51	281.99 ± 6.55	7.8×10^{21}	900	101
Multiplex M_2						
10x90x1	3860.52	-3594.71	265.81 ± 2.26	7.8×10^{20}	1070	101
90x10x1	7611.52	-7471.92	139.6 ± 6.21	1.82×10^{144}	1630	101
55x45x1	11780.3	-11585.3	195 ± 4.897	6.96×10^{88}	2960	101
Multiplex M_3						
10x20x70x1	6421.85	-6187.59	234.26 ± 3.66	9.12×10^{32}	1750	101
20x60x20x1	9800.17	-9644.93	155.24 ± 5.85	3.16×10^{96}	2580	101
47x6x47x1	4349.69	-4151.79	197.9 ± 4.81	1.42×10^{75}	987	101
70x20x10x1	9235.14	-9095.58	139.56 ± 6.21	3.31×10^{112}	2170	101
Multiplex $M4$						
40x30x20x10x1	9272.79	-9133.31	139.48 ± 6.11	4.31×10^{64}	2330	101
Multiplex M_5						
35x30x20x10x5x1	8659	-8527.38	131.62 ± 6.37	4.22×10^{56}	2185	101
Multiplex M_6						
35x25x20x10x8x2x1	7811.11	-7684.2	126.91 ± 6.47	3.51×10^{56}	1953	101

Comparing the values of the entropy it can be deduced as expected that there exists a correlation of this thermodynamic potential with the total number of adaptive weights of the networks (right most part of the table), as higher is the total number of adaptive weights as higher is the resulting value of the entropy. It's important to remember that the total number of units for the set of network architectures considered is identical ($N = 101$), thus the resulting effective complexity of the networks (defined in terms of the total number of adaptive weights) varies as a result of the particularities imposed by the topology of these networks.

However, there are several exceptions to this rule evidencing the strong influence of the topology, particularly the effect of the total number of neurons in the first layer of the networks, for example the shallow network 100x1 is the network with the smallest effective complexity but attains an entropy value larger than the networks 10x90x1 (expansive topological scheme), and the compressive autoencoder architecture 47x6x47x1 respectively. Similarly, the compressive autoencoder 47x6x47x1 posess a lower number of adaptive weights compared to the the network 10x90x1 but attains a higher entropy value.

Moreover, the networks that store higher amounts of energy are the shallow network 100x1 together with the networks following an expansive topology scheme suchs as 10x90x1 and 10x20x70x1 as occured in the previous case studies. Similarly, they constitute the networks attaining the lowest values of specific heat. In contrast, the topological scheme that stores the lowest amount of energy are hierarchical networks (excepting the shallow network). Of particular interest is the fact that the amount of energy stored by these networks appears to decrease (athough not monotonically) with the depth of the networks, that is with the number of hidden layers. Furthermore, the largest values of the specific heat are attained by networks belonging to this topological scheme, thereby evidencing once again the strong influence of the topology in the resulting thermodynamic properties of the networks.

Finally, from the point of view of the fluctuations of energy (ΔU), hierarchical networks (including the shallow network) appear to have the largest energy fluctuations whereas networks that are more resilient to fluctuations are those following an expansive topological scheme (e.g., 10x90x1).

4.4.2. The Generating Function of Energies

Four network architectures with a variable number of hidden layers ranging from one up to six were selected from table 2 following an increasing order on the total number of adaptive weights, all of them with a hierarchical topological scheme. Namely, the networks 100x1 ($W = 900$), 90x10x1 ($W = 1630$), 35x25x20x10x8x2x1 ($W = 1953$),and 70x20x10x1 ($W = 2170$).

It is important to remember that in this scenario, the expected differences on the number of accesible states are principally due to the difference on the total number of adaptive weights between the network architectures considered. More specifically, taking into consideration that the total number of units N is fixed there may exists a whole variety of network architectures with an identical number of neurons but with substantial variations in the total number of adaptive weights W). However, networks with a larger number of adaptive weights are expected to show (in general) a larger number of accesible states because each additional adaptive weight contributes with a number of degrees of freedom in correspondence with the average number of energy levels used for representing the adaptive weights.

$$M_{70\times20\times10\times1}(z) = 2.241\times10^{2913}z^{101} + 2.248\times10^{2915}z^{103} + 1.116\times10^{2917}z^{105} + \cdots \\ \cdots + 1.143\times10^{2915}z^{301} + 1.124\times10^{2913}z^{303} \tag{56}$$

$$M_{35\times25\times20\times10\times8\times2\times1}(z) = 1.306\times10^{2832}z^{101} + 1.311\times10^{2834}z^{103} + 6.518\times10^{2835}z^{105} + \cdots \\ \cdots + 7.293\times10^{2833}z^{301} + 7.178\times10^{2831}z^{303} \tag{57}$$

$$M_{100\times1}(z) = 1.94\times10^{651}z^{101} + 1.907\times10^{653}z^{103} + 9.283\times10^{654}z^{105} + \\ \cdots + 1.318\times10^{652}z^{301} + 1.269\times10^{650}z^{303} \tag{58}$$

$$M_{90\times10\times1}(z) = 4.014\times10^{4293}z^{101} + 3.995\times10^{4295}z^{103} + 1.968\times10^{4297}z^{105} + \\ \cdots + 9.22\times10^{4294}z^{301} + 8.994\times10^{4292}z^{303} \tag{59}$$

Observing the polynomial expressions of the generating functions of energy associated to the equilibrium states represented by each network architecture, it can be deduced that the total number of energy states (i.e. the powers of the complex variable z in the polynomial expressions) is fixed and its cardinality is in direct correspondence with N because of the fact that the total number of units is identical (remember that $N = 101$).

Moreover, as expected the number of accesible states, and thus the accesible volume of the phase space increases to the extent the total number of adaptive weights of the networks increases. In other words, the degree of disorder of the networks increases with the number of adaptive weights, and also its storage capacity.

The number of accesible microstates for the aforementioned networks follows the cardinality dictated by the adaptive weights W, that is, $M(1)_{100\times1} = 7.14\times10^{2241}$ microstates for the shallow network 100x1, $M(1)_{90\times10\times1} = 2.28\times10^{3331}$ microstates for the two hidden layers hierarchical network 90x10x1, $M(1)_{35\times25\times20\times10\times8\times2\times1} = 6.66\times10^{3408}$ microstates for the six hidden layer hierarchical network 35x25x20x10x8x2x1, and finally $M(1)_{70\times20\times10\times1} = 2.95\times10^{4036}$ microstates for the three hidden layer network 70x20x10x1.

This is also true for the expansive network architecture 10x90x1 ($W = 1070$, and $M(1)_{10\times90\times1} = 1.6 \times 10^{1794}$) and the compressive autoencoder architecture 47x6x47x1 ($W = 987$, and $M(1)_{47\times6\times47\times1} = 1.71 \times 10^{1957}$). These two networks together with the shallow network 100x1 constitute the networks with the lowest number of adaptive weights W, and the cardinality of the number of accessible microstates is in direct correspondence with the cardinality of its adaptive weights.

However, it is important to remember that these networks are those attaining entropy values that are not in direct correspondence with the number of adaptive weights W. For example, the shallow network attains an entropy value that is bigger than the values attained by the entropy in the networks 47x6x47x1 and 10x90x1 in spite of the fact of possesing both a lower number of adaptive weights and a lower number of accesible microstates, or the compressive autoencoder network 47x6x47x1 compared to the network 10x90x1. This fact is of particular interest since it is evidencing again not only the strong influence of the topology of the networks but also the influence of the input space dimension and the number of units in the first hidden layer of the networks in the values attained by the entropy [36].

4.4.3. Infuence of the Structural Parameters

Given the restrictions imposed by this scenario only two structural parameters are considered in this case. Namely, the influence of the effective complexity of the networks, and its depth. The input space dimension is only considered to assess its influence in the effective complexity of networks, with regards the thermodynamic potentials the results are identical to those obtained in scenario I (remember that the internal energy does not depend on the input space dimension, and the entropy grows monotonically with the dimensionality of the input space).

The graph of figure (13) represent the evolution of the entropy for the network architectures shown in table 3 as a function of their effective complexity for an input space dimension g_0 belonging to the set $2, 5, 8$. The most relevant aspect of this graph is that the entropy does not increase monotonically with W because of the strong influence of the topology but particularly as a result of the strong influence of the number of units in the first hidden layer of the networks. For example, it is important to remember that for $g_0 = 8$ the shallow network 100x1 is the network with the smallest effective complexity but attains an entropy value larger than the networks 10x90x1 (expansive topological scheme), and the compressive autoencoder architecture 47x6x47x1 respectively.

Similarly, the compressive autoencoder 47x6x47x1 posess a lower number of adaptive weights compared to the the network 10x90x1 but attains a higher entropy value. Indeed, when increasing the dimensionality of the input space there is always a value of g_0 for which the entropy of the network with the largest number of neurons in the first hidden layer, that is, g_1 surpass the value attained by the entropy of any network architecture possesing an inferior value of , g_1 with respect to the former. Having said this, it is important to remember that in the previous case studies the effect of the number of hidden layers in the thermodynamic behavior of the networks for hierarchical networks (the most common topological scheme in deep learning applications) evidenced a tendency of the entropy to decrease to the extent the number of hidden layers of the networks increased, and at the same time a progressive reduction of the energy stored by the networks. Furthermore, the heat capacity of the networks progressively decrease to the extent the depth of the networks increase. A similar behavior is observed when conserving the total number of neurons N. For example, considering the hierarchical topological scheme it can be deduced from table 3 that the entropy decreases to the extent the depth of the network increases. Indeed, the decrease of the entropy is not monotonically due to the strong influence of the number of neurons of the first hidden layer of the networks. Similarly, the reduction of the energy stored by the networks with their depth is not monotonically, and is bounded by their ground state, i.e., when the whole neurons comprising the network structure are not active. It is important to note that distributing a fixed number of neurons over an increasing number of layers leads to hierarchical networks with a succesive lower numbers of adaptive weights, and thus to a reduction of the values attained by the entropy.

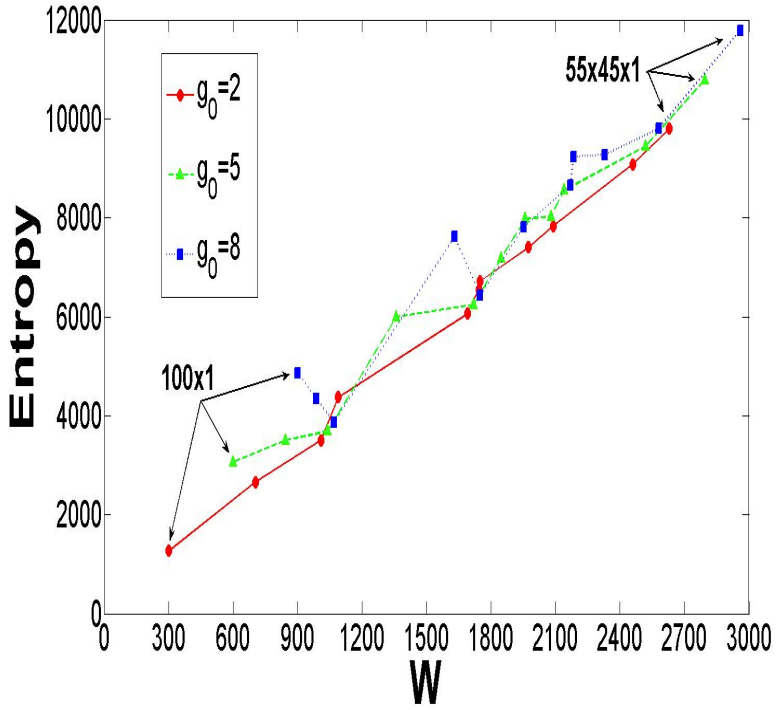

Figure 13. Graphical illustration the evolution of the entropy for the network architectures shown in table 3 as a function of their effective complexity for an input space dimension g_0 belonging to the set $2, 5, 8$. The networks with the lowest (100x1) and the highest (55x45x1) effective complexity for the values of g_0 represented are marked in the graph. The most relevant aspect of this graph is that the entropy does not increase monotonically with the effective complexity of the networks W because of the strong influence of the topology but particularly as a result of the effect exerted by the number of units in the first hidden layer of the networks (i.e., as larger is their number as higher is the entropy).

To further corroborate the aforementioned hypothesis two additional multiplex networks M_{10} and M_{20} with architectures $\overbrace{10 \times 10 \times \cdots \times 10}^{10} \times 1$ ($W = 990$), and $\overbrace{5 \times 5 \times \cdots \times 5}^{20} \times 1$ ($W = 520$) were selected to calculate their entropy, internal energy, and the specific heat. The entropy, internal energy, and specific heat of the network M_{10} reads respectively $S = 3777.27$,$U = 139.25$, and $C_v = 7.81 \times 10^{20}$, whereas for the network M_{20} those values read $S = 1964.3$,$U = 131.75$, and $C_v = 3.9 \times 10^{20}$, that is, thus corroborating the aforementioned hypothesis.

Summarizing, for hierarchical networks distributing a fixed number of neurons over an increasing number of layers leads to a progressive reduction of the effective complexity of the networks, and thus to a reduction of the values attained by the entropy that are implicitly evidencing a reduction of the degree of disorder of the equilibrium states represented by the networks, and thus to its information-storage capacity. Furthermore, this behavior is also accompanied by a successive decrease of both the heat capacity and the energy stored by the networks.

4.4.4. Thermodynamic Interpretation

According to the model presented in section 3 each of the multiplex networks considered in this section (i.e., those of table 3) correspond to the equilibrium states of a physical system composed of 101 artificial neurons with arbitrary topology but constrained to have an input space dimension equal to eight ($g_0 = 8$). For this system the total number of equilibrium states is huge (within the interval $[2^{99}, 2^{100}]$). The physical situation under study, that is, a system with a fixed number of particles makes the canonical ensemble the most adapted statistical ensemble to the analysis of its macroscopic properties. These equilibrium states are characterized as before by an energy value (internal energy) and its fluctuations with respect to its typical value that is shown in table 3. Furthermore, the numerical values of the entropy, free energy, and the specific heat are also available. To understand the macroscopic thermodynamic behavior of this physical system it is assumed as in previous sections the existence of an hypothetical process (or external force) that perturbates the equilibrium states represented by the set of network architectures of table 3 by making changes in their topology so as to pass from one multiplex network architecture to any other of those defined in the table. The kind of theoretical transformations studied hereafter (from one equilibrium state to another) are those that follow the law of increase of the entropy.

Having said this, it is important to note that when conserving the total number of neurons N, the degrees of freedom of the system vary at the different equilibrium states that the system may visit since the topology of the network and its effective complexity is changing when the system passes from one equilibrium state to another. The restrictions imposed by this scenario together with the fact that the transformations of state are drawn according to the law of increase of entropy lead to the fact that in most of the cases the accesible volume of the phase space gets expanded as a result of an increase of the number of adaptive weights, that is, the kind of degrees of freedom that are directly accesible to the learning algorithm. In other words, the nature of the degrees of freedom that are mainly responsible of the accesible volume of the phase space in the transformations of state considered in this scenario are linked to the number of adaptive weights W of the networks, and thereby the resulting degree of disorder of the networks, and its information encoding capacity. Furthermore, the main goal of this scenario is again to understand the role of the topology in the structure of the phase space of the system, and particularly the stability of the equilibrium states represented by the networks together with their thermal properties.

Two main possibilities may be considered, the first involving a transformation where the topological scheme of the networks is conserved (e.g., hierarchical vs. hierarchical) and the second concerning transformations leading to a complete change of the topology (e.g., Compressive autoencoder vs. hierarchical). As occurred in the previous scenario (when conserving the effective complexity of the networks), the influence of the topology in the resulting accesible volume of the phase space is more accused when the transformation of state leads to a change in the topological

scheme of the networks. Independently of this fact the main contribution to the total number of accesible microstates is due to the difference in the number of adaptive weights between the two equilibrium states involved in the transformation. Indeed, the set of architectures presented before correspond to the set of topological schemes described in the previous sections, also playing with the number of hidden layers to understand the influence of this structural parameter in this particular scenario.

From the set of network architectures of table 3, the equilibrium states represented by the networks 10x90x1 (state 1 with $W = 1070$), 47x6x47x1 (state 2 with $W = 987$), 100x1 (state 3 with $W = 900$), 10x20x70x1 (state 4 with $W = 1750$), 90x10x1 (state 5 with $W = 1630$), 35x30x20x10x5x1 (state 7 with $W = 2185$), 70x20x10x1 (state 8 with 2170), 20x60x20x1 (state 10 with $W = 2580$), and 55x45x1 (state 11 with $W = 2960$) a subset of them is used hereafter to understand (under the restrictions imposed by this scenario) the thermodynamic behavior of the model. The aforementioned set of network architectures is sorted by the values attained by the entropy and using the cardinality obtained to denote each state. As stated before, when perturbing the system it is assumed that its evolution from one state to another occurs according to the second law of thermodynamics.

Furthermore, the cardinality imposed by the values attained by the entropy is in direct correspondence with the accessibility of the equilibrium states [38]. For example, the equilibrium state 5 represented by the hierarchical network 90x10x1 is accesible from states 1 and 3 respectively since $S_1 < S_5$ and $S_3 < S_5$, or the equilibrium state 11 is accessible from state 5, and so forth.

Assuming that the physical system is found in the equilibrium state 1 represented by the expansive network architecture 10x90x1. The system is then perturbated by rearranging neurons and connections (i.e., by doing work on the system) passing to the equilibrium state 3 corresponding to the shallow network 100x1. The result of this transformation $T_{1\to3}$ is an increase of the internal energy of the body (i.e., the neural network) since $\Delta U = U_3 - U_1 = 16.18 \pm 6.93 > 0$. The gain of entropy obtained is $\Delta S = S_3 - S_1 = 1003.98$, but at the same time, the free energy of the system decreases $\Delta F = F_3 - F_1 = -987.8$. Surprisingy, the accesible volume of the phase space expands in spite of the decrease in the total number of adaptive weights evidencing the strong influence of the topology (there is a loss of $\Delta W = -170$ adaptive weights in this transformation) passing from a number of accesible states equal to $M(1)_{10\times90\times1} = 1.6 \times 10^{1794}$ to $M(1)_{100\times1} = 7.14 \times 10^{2241}$, that is, a number that is 457 orders of magnitude larger. Furthermore, using equation (49) it can be verified that $\frac{(S_3-S_1)-(U_3-U_1)}{M(1)_{100\times1}-M(1)_{10\times90\times1}} > 0$, that is, the equilibrium state represented by the shallow network is a stable state. Of particular interest is the fact that the heat capacity of the body increases in this transformation passing from a value $C_v = 7.8 \times 10^{20}$ to $C_v = 7.8 \times 10^{21}$, that is, a slight increase of one order of magnitude larger.

Moreover, if now one assumes that the system is found again in the equilibrium state 1 but now passes to the equilibrium state represented by the compressive autoencoder network 47x6x47x1 (state 2), that is, the transformation $T_{1\to2}$. The gain of entropy obtained in this case is $\Delta S = S_2 - S_1 = 489.17$. There is also a reduction of both the internal energy of the body $\Delta U = U_2 - U_1 = -67.91 \pm 5.31$, and the free energy $\Delta F = F_3 - F_2 = -557.08$. Furthermore, as occured before the accesible volume of the phase space expands in spite of the decrease in the total number of adaptive weights (there is a loss of $\Delta W = -83$ adaptive weights in this transformation) evidencing again the influence of the topology, and particularly the importance of the number of units in the first hidden layer of the networks together with the dimension of the input space, passing in this case from a number of accesible states equal to $M(1)_{10\times90\times1} = 1.6 \times 10^{1794}$ to $M(1)_{47\times6\times47\times1} = 1.71 \times 10^{1957}$ that is, a number that is 163 orders of magnitude larger. Furthermore, using equation (49) it can be verified that $\frac{(S_2-S_1)-(U_2-U_1)}{M(1)_{47\times6\times47\times1}-M(1)_{100\times1}} > 0$, that is, the equilibrium state represented by the compressive autoencoder network is also a stable state. Of particular interest is the fact that the heat capacity of the body increases substantially when compared to the previous transformation ($T_{1\to3}$) in this transformation passing from a value $C_v = 7.8 \times 10^{20}$ to $C_v = 1.42 \times 10^{75}$, that is, a 55 orders of magnitude larger, suggesting that learning in this kind of architecture will be in general harder compared to the shallow

network 100x1 or the compressive autoencoder architecture 47x6x47x1, thus leading in average to worst learning and generalization performance.

Similarly, if the system is found in the state 3 (the shallow network 100x1) and is perturbated so as to pass to the equilibrium represented by hierarchical network 90x10x1 (state 5) the result of this transformation is a reduction of the internal energy of the body since $\Delta U = U_5 - U_3 = -142.39 \pm 9.03 < 0$. The gain of entropy obtained is $\Delta S = S_5 - S_3 = 2747.02$, and at the same time, the free energy of the system decreases $\Delta F = F_5 - F_3 = -2889.41$. The accesible volume of the phase space expands as a result of the increase in the total number of adaptive weights (there is a gain of $\Delta W = 700$ adaptive weights in this transformation) passing from a number of accesible states $M(1)_{100\times1} = 7.14 \times 10^{2241}$ to $M(1)_{90\times10\times1} = 2.28 \times 10^{3331}$, that is, a number that is 1090 orders of magnitude larger. Furthermore, using equation (49) it can be verified that $\frac{(S_5-S_3)-(U_5-U_3)}{M(1)_{90\times10\times1}-M(1)_{100\times1}} > 0$, that is, the equilibrium state represented by the hierarchical network is a stable state. Of particular interest is the fact that the heat capacity of the body increases substantially in this transformation passing from a value $C_v = 7.8 \times 10^{21}$ to $C_v = 1.82 \times 10^{144}$, that is, an exponential gain of 123 orders of magnitude. In other words, the internal energy ofthe system decreases accompanied with a substantial increase in the capacity for heat absortion, both facts evidencing better learning and generalizaion performance for the equilibrium state represented by the hierarchical network 90x10x1.

A similar behavior is obtained for the transformation $T_{6\to8}$, that is from the network 35x25x20x10x8x2x1 to 70x20x10x1 (both hierarchical networks but with a different number of hidden layers), that is, an expansion the accesible volume of the phase space, that is, passing from a number of accesible states $M(1)_{35\times25\times20\times10\times8\times2\times1} = 2.46 \times 10^{2862}$ to $M(1)_{90\times10\times1} = 4 \times 10^{2943}$ accompanied with an increase on the total number of adaptive weights $\Delta W = 217$ but with one important difference: the variation in internal energy of the system$\Delta U = U_8 - U_6 = 7.94 \pm 8.89$ cannot be determined with precision due to the fact that in ths transformtion of state the energy fluctuations are comparable to the gain in internal energy of the body. Independently of this fact, the substantial increase of the heat capacity of the system (i.e., the theoretical body) as a result of this transformation is clearly evidencing the influence of the depth of the networks in the thermodynamic properties of hierarchical networks. In other words, it is expected better learning and generalization properties in average for the network 70x20x10x1 compared to the former.

To study the influence of the number of hidden layers in hierarchical networks it is assumed now that the system passes from the equilibrium state represented by the network 90x10x1 (state 5) to the equilibrium state 6 represented by the six hidden layers network 35x25x20x8x2x1, that is, the transformation $T_{5\to6}$. The gain of entropy obtained is $\Delta S = S_6 - S_5 = 1047.48$. In this case there is also a reduction of both the internal energy of the body $\Delta U \pm \Delta U = U_6 - U_5 = -7.98 \pm 8.89$, and the free energy $\Delta F = F_6 - F_5 = -1055.46$. Furthermore, the accesible volume of the phase space gets expanded as a result of the increase in the total number of adaptive weights (there is a gain of $\Delta W = 555$ adaptive weights in this transformation) passing from a number of accesible states equal to $M(1)_{90\times10\times1} = 2.28 \times 10^{3331}$ to $M(1)_{35\times25\times20\times10\times8\times2\times1} = 6.66 \times 10^{3408}$, that is, a number that is 77 orders of magnitude larger. Furthermore, using equation (49) it can be verified that $\frac{(S_6-S_5)-(U_6-U_5)}{M(1)_{90\times10\times1}-M(1)_{35\times25\times20\times10\times8\times2\times1}} > 0$, that is, the equilibrium state represented by the six hidden layer network 35x25x20x10x8x2x1 is a stable state. Of particular interest is the fact that the heat capacity of the body decreases in this transformation passing from a value $C_v = 1.82 \times 10^{144}$ to $C_v = 3.51 \times 10^{56}$, that is, a reduction of 88 orders of magnitude. This behavior is clearly evidencing the influence of the depth of the networks in the thermodynamic properties of hierarchical networks, that is, the progressive decrease of the entropy and the heat capacity when increasing the depth of the networks, both facts affecting somehow its learning and generalization performance.

Finally, an identical behavior is also obtained for the transformation $T_{5\to11}$ when compared to the previous one, that is from the network 90x10x1 to 55x45x1 (hierarchical networks with an identical number of hidden layers, and a number of accesible microstates $M(1)_{55\times45\times1} = 1.25 \times 10^{5182}$) but with an important difference: The internal energy of the system increases in this case $\Delta U = U_{11} - U_5 =$

55.4 ± 7.91. This fact together with the reduction of the heat capacity of the system are evidencing again from a learning point of view the generation (in average) of energy loss landscapes with a more complex geometrical structure compared to those generated by the network 90x10x1, both facts affecting the learning and generalization performance of this network (i.e.,55x45x1) in spite of attaining a larger entropy value.

Summarizing, in this scenario the transformations of state that follow the law of increase of the entropy give rise to an expansion of the accesible volume of the phase space of the system mostly as a result of the increase of the effective complexity of the networks. However, there are very specific cases where the increase of the entropy is not accompanied with an increase on the effective complexity of the networks. Surprisingly, the aforementioned transformations are those leading that change the topological scheme of the networks, thus evidencing again the strong influence of the topology in the thermodynamic behavior of the networks. Furthermore, as occurred in the previous two scenarios, the topology is playing again an important role in the structure of the phase space of the networks, and thereby in the resulting geometrical complexity of the loss landscapes generated by those networks when facing any learning task.

4.4.5. Experimental Results

The graphs of figure (14) represent the generalization performance for a subset of the multiplexes architectures studied in section 4.4.1 (see table 3) when using as training algorithm standard gradient descent (i.e., Backpropagation). The dataset used for training the networks is an artificial database *Gauss8d.dat* used in [12] to study the effect of the inpùt dimension on the classification error rates. The set of 3 hidden layer architectures are underparameterized (i.e., the total number of adaptive weights is smaller than the total number of patterns contained in the dataset. A ten fold crossvalidation procedure [25] models was used again to assess the generalization performance of the networks using ten averages per fold.

Each graph is composed of two parts. The top part represents the evolution of the generalization performance of the network sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The botton part represents the average generalization (the standard deviation is represented as an error bar) obtained when averaging the generalization at the sampling points described before. It is important to remember that the goal is to assess the typical behavior of the network architectures in terms of generalization performance, that is, which is the typical expected behavior of the generalization independently of the number of epochs used to train the networks.

The network architectures of figure (14) are underparameterized for the dataset *Gauss8D.dat* (i.e., $W < 5000$), and were selected principally to compare first the learning and generalization performance of the shallow network 100x1 against deep networks of two hidden layers (i.e., networks with just an extra hidden layer) under the presence of the curse of dimensionality problem), secondly to understand the differences in the learning and generalization of the networks 90x10x1 and 55x45x1, that is, networks with and identical topological scheme (hierarchical in this case) but with substantial differences on the number of adaptive weights, and thus in their resulting entropy values (see table 3). It is important to remember that under the restrictions imposed by this scenario (i.e., networks with an identical number of units N) the differences on the observed entropy values are mainly due to the number of adaptive weights of the networks.

When comparing the generalization performance obtained with standard gradient descent techniques, the worst results are obtained for the expansive network architecture 10x90x1, followed by the shallow network 100x1, that is, the networks storing larger amounts of energy and with lower heat capacities compared to the hierarchical networks 90x10x1 and 55x45x1, thereby evidencing once again a more complex geometrical structure of the loss landscape (i.e., the landscape of empirical risk) [52,69] both facts affecting the resulting learning and generalization performance of the network.

Of particular interest is the fact that the network 10x90x1 attains a lower entropy value compared to the shallow network in spite of having a larger number of accesible microstates but also a larger

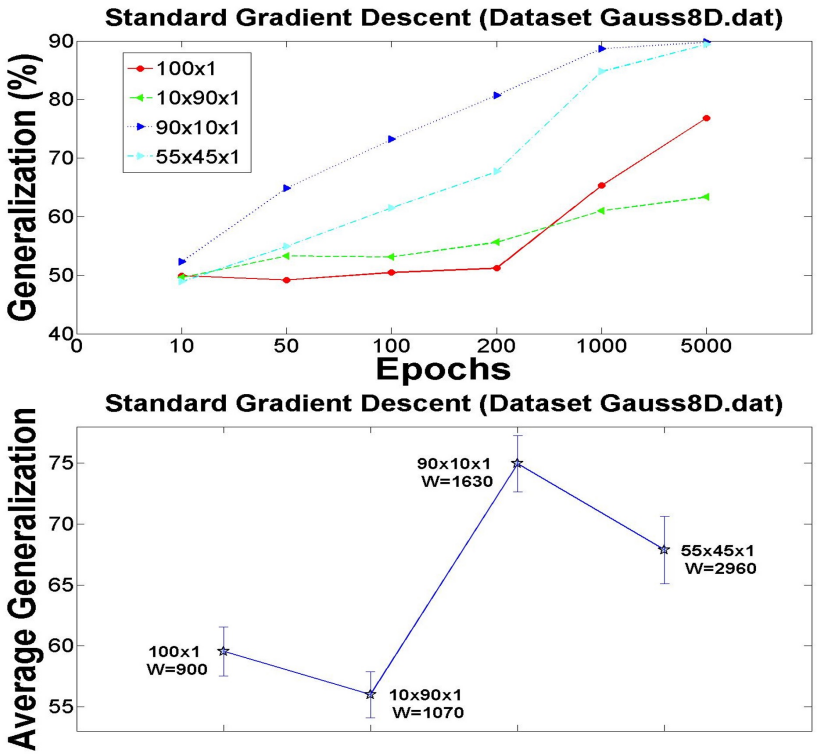

Figure 14. Graphical illustration of the generalization performance obtained for the group of networksM_1 and M_2 of table 3 with an identical number of units using as learning algorithm backpropagation with the dataset *Gauss8d.dat* [12]. The top part represents the evolution of the generalization performance of the networks sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The botton part represents the average generalization (its standard deviation is represented as an error bar) obtained when averaging the generalization at the aforementioned sampling points using a ten fold cross-validation procedure [25]. The goal is to assess again the typical behavior of the generalization independently of the number of epochs used to train the networks. Quantitatively networks storing larger amounts of energy and attaining smallest heat capacities (i.e., 100x1 and 10x90x1) are those presenting the worst learning and generalization performance for standard gradient descent. The generalization performance improves with the value attained by the entropy of the networks (see table 3) excepting for the hierarchical network 55x45x1 evidencing the existence of more complex geometrical structure of its error loss landscape.

number of adaptive weights W as a result of the strong influence of the topology. It is important to remember that the total number of units N of these networks is identical, thus, the larger amount of energy stored by the networks 100x1 (shallow network) and 10x90x1 (the expansive network architecture) are in this case evidencing the interplay between the topology and the number of adaptive weights W.

Moreover, the generalization performance appears to improve with the value of the entropy attained by the networks (see table 3) excepting for the hierarchical network 55x45x1 (i.e.,the network attaining the largest entropy value). Indeed, the hierarchical network 90x10x1 is the network attaining the best learning and generalization performance for standard gradient descent. However, the reduced heat capacity of the network 55x45x1 compared to 90x10x1 together with the fact that the network 55x45x1 stores a larger amount of energy compared to the network 90x10x1 appear to suggest a more complex geometrical structure of the loss landscape in spite of having a larger number of accesible microstates (i.e., a larger accesible volume of the phase space), and thus a larger storage capacity compared to the network 90x10x1.

To test this hypothesis the generalization capability of these networks was also assessed using the scaled conjugate gradients algorithm [11,30] (see appendix C), corroborating the predictions of the theoretical model, that is, using a learning protocol with better convergence properties leads to the fact that the learning and generalization performance obtained are now in direct correspondence with entropy values attained by the networks excepting for the shallow network 100x1 whose learning and generalization performance ouperform those of hierarchical networks in spite of the fact of attaining a lower entropy value, and a lower number of accesible microstates. Surprisingly, when using a learning protocol with better covergence properties the two networks achieving the best learning and generalization performance are those storing the largest amount of energy. However, it is important to emphasize that these set of networks are underparameterized for the dataset used (i.e., $W < 5000$), the curse of dimensionality phenomenon is present, and especially the fact that the phase space and the VC dimension can play independent roles in the process of generalization [47]. Furthermore, as opposed to the previous case studies, when conserving the total number of artificial neurons, the kind of degrees of freedom that are mainly responsible of the accesible volume of the phase space of the system are the total number of adaptive weights W, that is, the kind of degrees of freedom that are directly configurable by the learning protocol suggesting that the nature of the degrees of freedom mainly responsible of the accesible volume of the phase space might influence the resulting learning and generalization performance of these models. This issue is further analyzed in detail in the second part of this research.

4.4.6. Summary of Results

When conserving the total number of artificial neurons of the networks the principal conclusions that can be extracted when combining both the theoretical and experimental results are very similar to those obtained in the previous case studies and they can be synthesized as follows:

1. Quantitatively networks leading to better generalization performance are those with larger information-storage capacities. From the whole set of network topologies, hierarchical networks (excepting shallow networks) are those leading to higher entropy values, and thus to better generalization performance. However, a larger capacity (i.e., a higher entropy value) might be not enough to provide better learning and generalization performance used because of the influence of the topology in the geometrical complexity of the loss landscape.
2. The second law of thermodynamics permits to compare the generalization performance of different network architectures (i.e., equilibrium states) comparing entropy variations between equilibrium states that are adiabatically accesible. Specifically, if A and B are two network architectures with an identical number of neurons if $S_A > S_B$ quantitatively the typical generalization performance of Network A will be better compared to Network B if both states are stable excepting those transformations accompanied with a reduction of the heat capacity

indicating the existence of a more complex geometrical structure of the loss landscape. In that case, the improvement on generalization performance is limited by the convergence properties of the learning algorithm used.

3. When conserving the total number of neurons of the networks the transformations of state that follow the law of increase of the entropy always give rise to an expansion of the accesible volume of the phase space of the system mostly as a result of the increase of the effective complexity of the networks. However, there are very specific cases where the increase of the entropy is not accompanied with an increase on the effective complexity of the networks as a result of the strong influence of the topology in the structure of the phase space of the system.
4. Keeping fixed the total number of neurons of the networks while increasing the effective complexity leads to networks with larger entropy and internal energy values. This fact may lead to equilibrium states with limited stability and/or with harder energetic landscapes from a learning point of view as a result of the increased amount of energy stored by these networks. Equilibrium states leading to the generation of more complex geometrical structures of the loss landscape (either stable or metastable) are characterized by storing larger amounts of energy compared to stable equilibrium states, and also by attaining lower heat capacities.
5. For hierarchical networks distributing a fixed number of neurons over an increasing number of layers leads to a progressive reduction of the effective complexity of the networks, and thus to a reduction of the values attained by the entropy that are implicitly evidencing a reduction of the degree of disorder of the equilibrium states represented by those networks, and thus to their information-storage capacity. Furthermore, this behavior is also accompanied by a successive decrease of both the heat capacity and the energy stored by the networks.

5. Thermodynamic Efficiency

The theory of thermal conductivity deals with the transport of heat (i.e., how quickly a body can absorb heat) whereas the concept of heat capacity (or specific heat) is a thermodynamic concept that expresses the idea how much heat a body it can absorb. It is important to note that heat capacity is a thermal property of a body whereas thermal conductivity is a transport property both of them varying considerably with temperature. In solid state materials heat conduction is strongly linked to geometrical properties of the lattices of atoms defining its structure [65] such as lattice contributions (phonons) and electric carriers (electrons and holes) between others. Indeed, materials with both high and very low thermal conductivities are nowadays of great theoretical and technological interest.

The heat capacity of a body is proportional to the variations of internal energy with respect to the temperature. Therefore, if a system fluctuates about its equilibrium value to a greater extent, it is plausible to state that it will have a larger heat capacity, as more of its energy levels can be more easily accessed. In other words, the heat transfer through the body (i.e., its thermal conductivity) should increase because of the existence of larger energy fluctuations.

Having said this, the theoretical analysis performed in section 4 permits to interpret machine learning properties of deep learning machines in terms of thermodynamic concepts. Particularly, it was shown the link between the energy stored by the networks (and its specific heat) with the geometrical complexity of the loss landscape (i.e, the landscape of the empirical risk), but especially, the analogy between the learning phase of this models and the transport of heat through the neural network (the theoretical body). It is important to remember that in a supervised learning context presenting a pattern to the network and changing the values of its adaptive weights is equivalent to bring energy into the network without delivering work thus changing the average energy of the system due to the changes in the probabilities of ocurrence of the microstates, and the variation of internal energy is identified with the infinitesimal heat absorbed by the network.

Thus, as higher is the heat capacity of a network as higher (and faster) will be its capacity for absorving the infinitesimal heat generated whenever a pattern is presented to the network during the training phase, and thus, as better will be its thermodynamic efficiency due to lower

(or reduced) dissipation processess. It is important to remember that deep learning machines are viewed as theoretical bodies whose thermodynamic properties are determined from the proposed theoretical model, and the energetics of any kind of information processing system (including deep learning machines) are subject to the laws of thermodynamics, and particularly, the second law of thermodynamics limits the thermodynamic efficiency with which information can be processed.

Moreover, the basic principle of the thermodynamics of information processing holds that any logically irreversible manipulation of information (such as the erasure of information or the merging of two computation paths) must be accompanied by a corresponding entropy increase in the non-information-bearing degrees of freedom of the information-processing apparatus or its environment, that is, dissipation [8]. Thus, from an information-processing point of view the nonlinear operation performed by the artificial neurons is logically irreversible since different computation paths (i.e., the inputs to the neuron) are merged into a unic path (i.e., the output of the neuron).

Taking all these consideration into account, it is plausible to state that the computational times associated to the learning and recall phases of deep learning machines might be influenced by its thermodynamic efficiency.

The graphs of figures (15), and (16) present a graphical comparison for the group of architectures used in section 4.2 between the values attained by the specific heat for these set of networks with the required training time to complete 200 epochs with the datasets used to assess the generalization performance for the group of architectures studied in section 4.2. Again the qualifiers specific heat and heat capacity are used indistinctly hereafter (remember that both concepts are equivalent for networks with an identical structural complexity).

Each graph is composed of two parts. The top part represents the logaritm of the heat capacity for the group of networks used before (i.e., those of table I). The botton part represents the average running times (and its standard deviation is represented as an error bar) obtained when averaging 1000 training phases of 200 epochs each using standard gradient descent with the same datasets used to assess the generalization performance. The software simulations concerning the graphs (15) were performed on computer running windows 7 as operating system and with an intel core i5 CPU at 2.27 Ghz excepting for the graph of figure (16) that was performed on a computer running windows XP with a single pentium *M* microprocessor runnning at 1.6 Ghz. In both cases, it is important to note that the measures of running times must be carried without the presence of programs and/or background processes leading to any extra CPU load that might bias the measures (e.g., an antivirus software).

Moreover, the goal is to assess the likelihood of the aforementioned hypothesis concerning the relationship between the computational time required for learning when fixing the computational budget, that is, the number of epochs used for the learning phase, and the thermodynamic efficiency with which information can be processed by deep learning machines during the training phase of these models, especially taking into consideration that the computational complexity measured in terms of the number of multiplications, additions, storage and nonlinear operations involved in the learning phase of networks with an identical structural complexity is exactly the same.

From the inspection of the graph of figure (15), it can be observed that there exists an inverse correlation between the specific heat associated to a network and the computational time required for learning. Specifically, as higher is the heat capacity of a network as lower is the computational time required for training. In this case, being faster the shallow network compared to the deep networks of three hidden layers. The standard deviation of the running times is quite similar for the three networks in this case, although slightly higher for the network 11x7x3x1.

An identical behavior is also observed for the set of network architectures of figure (16). Quantitatively those networks with larger heat capacities are those exhibiting faster computational training times. Furthermore, those networks with a hierarchical scheme (or partially hierarchical such as the expansive autoencoder architecture 16x211x39x1) are those leading to lower computational times, thereby evidencing not only the strong influence of the topology in the learning and generalization capabillities of these models but also its influence in the resulting thermal conductivity of these models.

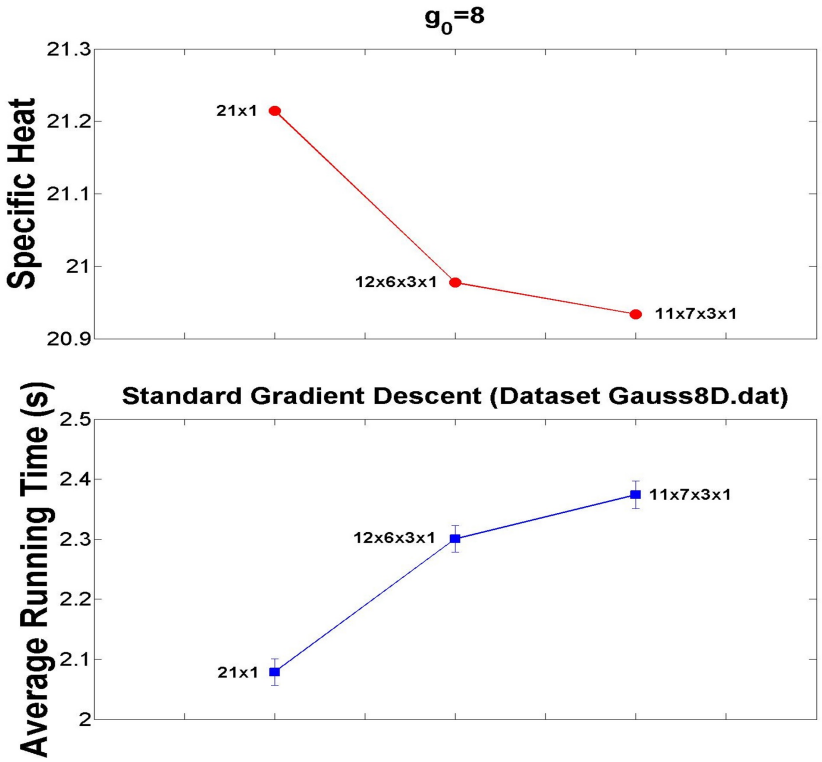

Figure 15. Graphical comparison between the values attained by the specific heat (top part of the graph) for the group of networks M_{1a}, and M_{3a} of table 1 (a shallow network and two deep networks with three hidden layers with an identical structural complexity) and the averaged computational times (bottom part of the graph) obtained after averaging 1000 learning phases of 200 Epochs each (its standard deviation is represented as an error bar) using the backpropagation algorithm with the dataset *Gauss8D.dat* [12]. Deep learning machines are viewed as theoretical bodies whose thermodynamic properties are determined by the proposed mathematical model. The graph shows the existence of an inverse correlation between the specific heat associated to a network and the computational time required for learning, thus evidencing the influence of the thermal conductivity of these models in the computational times associated to their learning and recall phases. In other words, fixing the computational budget for learning (i.e., the number of Epochs) as higher is the heat capacity of a network as lower will be the computational time required for learning.

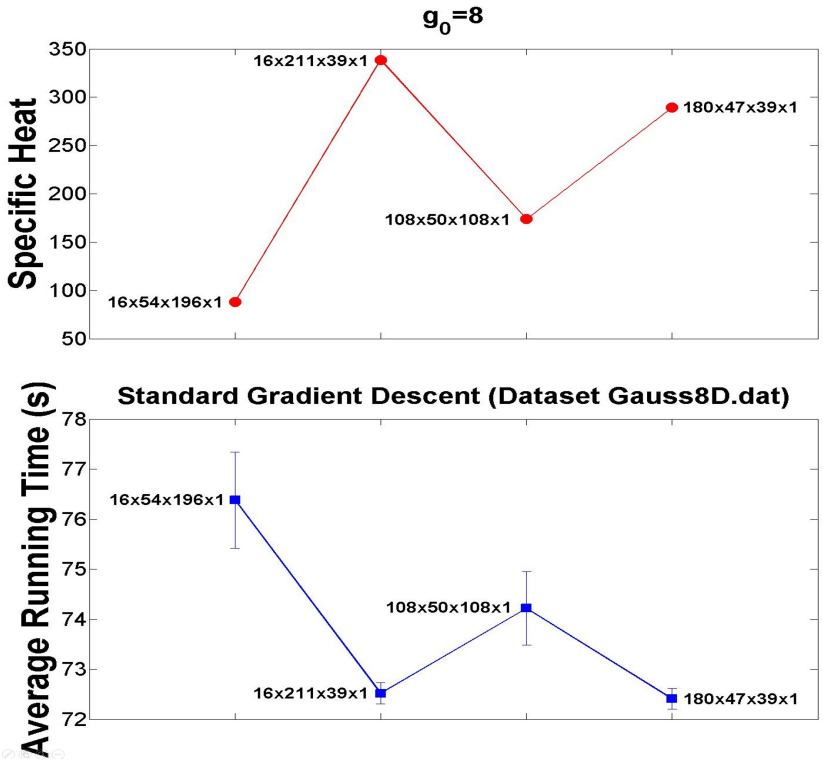

Figure 16. Graphical comparison between the values attained by the specific heat (top part of the graph) for the group of networks M_{3b} of table 1 (four deep networks with three hidden layers with an identical structural complexity following different topological schemes) and the averaged computational times (bottom part of the graph) obtained after averaging 1000 learning phases of 200 Epochs each (its standard deviation is represented as an error bar) using the backpropagation algorithm with the dataset *Gauss8D.dat* [12]. Quantitatively the graph shows again the existence of an inverse correlation between the heat capacity associated to a network and the computational time required for learning, thus evidencing the influence of the thermal conductivity of these models in the computational times associated to their learning and recall phases. Of particular interest is the fact that those networks presenting larger energy fluctuations are those presenting lower computational times. In other words, fixing the computational budget for learning (i.e., the number of Epochs) as higher is the heat capacity of a network as lower will be the computational time required for learning.

Of particular interest is the fact that these networks are those presenting larger energy fluctuations. Indeed, networks with larger heat capacities and larger energy fluctuations are those presenting smaller deviations with respect to the average learning time.

Having said this, it is important to note that the influence of the topology in the thermodynamic efficiency of these models is so strong that its effects can be even appreciated in the rest of scenarios considered (i.e., case studies II and III) where there is exists an intrinsic difference on the computational complexity due either to the difference on neuron numbers (case study II) or because of the differences on the number of adaptive weights of the networks (case study III), leading in both cases to differences on the associated computational complexity, that is, the number of multiplications, additions, and non-linear operations involved during the learning phase.

For example, the difference on neuron numbers of the networks 5x96x116x1 ($N = 218.$ and $C_v = 8.66 \times 10^{154}$), and 216x44x12x1 ($N = 273$, and $C_v = 5.81 \times 10^{346}$) is $\Delta N = 55$, both of them with an identical effective complexity for $g_0 = 8$. Taking into consideration that the hierarchical network 216x44x12x1 possesses a larger number of neurons longer average learning times are expected because in a digital computer the nonlinear operations performed by the articial neurons takes more CPU clock cycles compared to simpler operations like multiplications or additions. In other words, in general when conserving the effective complexity of the networks, increasing neuron numbers leads to an increase of the information-storage capacity of the networks but at the cost of a worst thermodynamic efficiency (remember that the nonlinear operation performed by the artificial neurons is a logically irreversible operation according to the basic principle of the thermodynamics of information processing).

However, in spite of the substantial difference on neuron numbers ($\Delta N = 55$) between both architectures following an identical procedure as that described before for the group of networks of figure (15) (using the same dataset for the learning phase), the computational times obtained are very similar, that is, an average running time of $22,48s$ ($\sigma = 2.39$ for the former and $22,58s$ ($\sigma = 0.53$) for the hierarchical network 216x44x12x1 evidencing once again the strong influence of the topology in the resulting thermodynamic efficiency of these computational models.

Moreover, for the third case study (networks with an identical number of units but with different effective complexity) a similar experiment was performed using the network architectures 90x10x1 ($W = 1360$ for $g_0 = 5$) and 10x90x1 ($W = 1040$ for $g_0 = 5$) but using the dataset *phoneme.dat* [12] that contains samples in a five dimensions space ($g_0 = 5$). In this case, there exists a difference of $\Delta W = 320$ on the number of adaptive weights between both networks. Surprisingly, the computational times obtained are lower for the hierarchical network 90x10x1, that is, an average runnning time of $7s$ ($\sigma = 0.15$) for the former and $7.52s$ ($\sigma = 0.22$)for the expansive network architecture 10x90x1, thus corroborating once again the link between the topology and the thermodynamic efficiency of these models.

Finally, with regards the recall phase of these models, the multiplex networks M_{2a}, and M_{2b} of table 1 where selected to test their thermodynamic efficiency. Specifically, the computational times associated to the result of averaging 100 simulations were measured. Each simulation consisted in the following steps: firstly, running a learning phase of 200 epochs using the dataset *Clouds.dat* and the standard Backpropagation algorithm as learning protocol, initializing the adaptive weights of the networks from random configurations at each simulation. Secondly, to better appreciate the temporal differences in the computations, the time spent by the networks in 500 recall phases (i.e, 500 passes of the training dataset) was computed after each learning phase. For both groups, the computational times obtained appears to corroborate that the influence of the topology in the thermodynamic effciency with which information can be processed by these models is so strong that its effects are also tangible in the recall phase. More especifically, the computational times obtained for the first group M_{2a} composed of the hierarchical network 301x11x1, and the expansive topological scheme network 12x300x1, were respectively $38.39s$ ($\sigma = 0.27$), and 98.47 ($\sigma = 20.36$), whereas for the second group M_{2b} the average

computational times obtained were 24.47s ($\sigma = 0.22$) for the network 150x24x1, and 30.5 ($\sigma = 3.96$) for the network 25x149x1.

Summarizing, the influence of the topology of deep learning machines in its resulting computational properties is so strong that it also affects the thermodynamic efficiency with which information can be processed by these computational models during its learning and recall phases.

6. Discussion

Deep learning machines have given rise to an active and ongoing debate [23]. While their incredible success throughout an important number of disciplines have been broadly recognized, their scientific foundations are still poorly understood . The theoretical analysis presented before permits to interpret machine learning properties of deep learning machines such as their learning and generalization performance, or the complexity of the loss landscape in terms of thermodynamic concepts. Thereby, allowing the use of the well known machinery of Thermodynamics to understand the behavior of deep learning machines, and particularly:

1. When fixing the structural complexity of the networks the theoretical information storage capacity of the networks should be identical, however, this appears not to be the case due to the strong influence of the topology in the structure of the phase space of the networks. Networks with an identical structural complexity may exhibit strong differences in the number of accesible states, and thus in their resulting degree of disorder as well as in their information storage capacity. Hierarchical networks appear to present a higher degree of disorder, as well as a higher storage capacity for encoding information compared to the rest of possible topological schemes. The transformations of state according to the law of increase of entropy always lead to an expansion of the accesible volume of the phase space of the system. Thus, metastable states are linked only to those transformations giving rise to an increase of the internal energy of the body. Equilibrium states (i.e., deep learning machines) with associated larger entropy values are those with a higher probability of occurrence, however, at the same time, networks possessing a similar number of accesible states may lead to substantial differences on the entropy values as a result of differences in the probabilities of ocurrence of the accesible states that are conditioned by the particularities of the topology. Furthermore, this behavior is suggested to be responsible of changes in the resulting geometrical complexity of the loss landscape that face any statistical classification algorithm during the learning phase of these models making the learning task easier or harder. Finally, under the strong restrictions imposed by this scenario shallow networks appear to present a higher degree of disorder but also a higher information storage capacity compared to deep networks. The internal energy content is also higher for shallow networks compared to deep networks.
2. When fixing the effective complexity of the networks the theoretical information storage capacity of the networks increases with the total number of neurons but also its resulting degree of disorder. Metastable states are linked to those transformations of state leading to an increase of the internal energy of the body accompanied with an expansion of the accesible volume of the phase space, but also to those transformations leading to a decrease (or loss) of the internal energy of the body that is accompanied with a contraction of the accesible volume of the phase space. Furthermore, the accesible volume of the phase space gets compressed or expanded depending on the neuron numbers but especially on the particularities of the topology of the networks. The probabilities of occurrence of the accesible states may also vary in a way that can be independent (or not) of the way the accesible volume of the phase space gets expanded or contracted, and this behavior is suggested to be responsible of the geometrical complexity of the loss landscape that face any statistical classification algorithm during the learning phase of these models making the learning task easier or harder depending on its characteristics. Compared to deep networks, shallow networks present a higher degree of disorder but also a higher information storage capacity, although in this case the disorder and its larger storage capacity is principally due to

their larger amount of neurons. The internal energy content is also higher for shallow networks compared to deep networks although again is principally due to the larger amount of neurons that present shallow networks under the restrictions imposed by this scenario.

3. When fixing the total number of neurons the information storage capacity of the networks increases with the effective complexity of the networks but also their resulting degree of disorder. When passing from one equilibrium state to another according to the law of increase of the entropy the accesible volume of the phase space always get expanded, thus, metastable states are linked only to those transformations leading to an increase of the internal energy of the body. As opposite to the previous two scenarios shallow networks may or may not present a higher degree of disorder and a larger information storage capacity compared to deep networks since its behavior is strongly linked to the particularities of its structural parameters such as the total number of units, and/or the input space dimension. Furthermore, independently of the interplay between the aforementioned structural parameters shallow networks are always the network topology storing the largest amount of energy.

6.1. Implications to the Main Set of Theory Questions

The preliminary implications of the model with respect to the three main sets of theory questions [51] are discussed hereafter and further developed in the second part of this research. It is important to emphasize that the most important implication of the proposed theoretical model is to show that independently of the learning protocol and/or the dataset used to train deep learning machines the role of the topology in the resulting computational properties of these models (expressed in terms of thermodynamics concepts) is more important than has previously been assumed.

With regards to the power of the architecture (the first question) is related to the entropy. Complex learning functions need higher amounts of information for being described, and thus need networks with higher storage capacities. Networks with higher entropy values have the potential to provide better learning and generalization performance since as higher is the entropy of those networks as higher will be the complexity of the space of functions that they are able to represent. However, a larger entropy values is a neccessary but not sufficient condition for a network to provide better generalization performance because of the strong influence of the internal energy in the generalization performance. Thus, independently of the function to be learnt shallow networks will typically present a disadvantage compared to deep networks as a result of storing larger amounts of energy given the relationship between the energy stored by the networks and the geometrical complexity of the loss landscape.

The learning process (second question) , and particularly the structure of the landscape of empirical risk affects the thermodynamic efficiency of these models (not forgetting that the topology rules the thermodynamic signature of these models, and thus their thermodynamic efficiency) and is related to the amount of energy stored by the networks and the value attained by the specific heat. Quantitatively networks that store larger amounts of energy and attain lower heat capacities lead to more complex geometrical structure of the loss landscapes resulting in harder learning phases evidenced by a lower speed of learning. More specifically, minima are easier to be found in those networks attaining larger heat capacities and lower internal energy values independently of the existence or not of overparameterization conditions. In other words, the minimal number of epochs needed to attain a prescribed learning and/or generalization error (the speed of learning) will be typically larger for those networks that attain larger internal energy values and lower heat capacities.

Generalization (third question) is linked to the entropy and the amount of energy stored by the networks, and thereby to the geometrical structure of the loss landscape, that is the reason why minima found by Stochastic Gradient Descent generalize so well. Both shallow and deep networks are universal, that is they can approximate arbitrarily well any continuous function of n variables on a compact domain.

However, compared to deep networks the extreme topology of shallow networks (concentrating all the units in a single layer) intrinsically store larger amounts of energy attaining at the same time (in most cases) lower heat capacities both facts affecting not only the geometrical complexity of the loss landscape but also its thermodynamic efficiency. Furthermore, deep networks are less affected by overfitting compared to shallow networks because shallow networks store larger amounts of energy compared to deep networks, thus leading typically to more complex geometrical structures of the loss landscape. Overfitting is discussed in detail in the second part of this research although it can be advanced that is not the result of an excessively large number of degrees of freedom but a combination of the geometrical structure of the loss landscape, the accesible volume of the phase space, and the nature of the degrees of freedom involved in the accesible volume of the phase space. For example, the networks 150x24x1 ($N = 175$)and 558x5x3x1 ($N =$) possess and identical effective complexity for $g_0 = 2$ in spite of a difference of 392 neurons. The larger number of neurons of the network 558x5x3x1 leads to the fact that this network stores a larger amount of energy ($U = 705.1$) compared to the network 150x24x1 ($U = 253.3$) in spite of the fact that both possess an identical number of adaptive weights ($W = 3924$). In other words, the accesible volume of the phase space for network 558x5x3x1 is larger compared to the network 150x24x1 but such a difference is mainly because of the difference on neuron numbers that are acting as "noise", that is, increasing the amount of energy stored by the system thus leading to a more complex geometrical structure of the loss landscape resulting in larger overfitting effects. Furthermore, the overfitting affects are hypothesized to be larger under the absence of overparameterization conditions.

In summary, quantitatively the best generalization performance is obtained for those networks maximizing the quotient entropy versus internal energy. Hierarchical networks constitute the topological scheme that maximize the generalization performance criterion, and with the highest thermodynamic efficiency compared to any other topological scheme. Although it is important to highlight that excessively increasing the depth of hierarchical networks penalizes their thermodynamic efficiency, and thus the expected computational times associated to their learning and recall phases.

Having said this, according to [51] deep hierarchical networks have the theoretical guarantee, which shallow networks do not have, that they can avoid the curse of dimensionality for an important class of problems, corresponding to compositional functions. To avoid the curse of dimensionality deep networks need to achieve a rate of convergence independent of the input dimensionality g_0 something that is only possible providing increasing smoothness of the unknown underlying function with increasing g_0.

According to the model the energy stored by deep learning machines (shallow or deep) is independent of the input dimensionality. However, hierarchical networks compared to others constitute the only topological scheme attaining the lowest internal energy values, and the energy stored by the networks together with the level of disorder progressively decrease when increasing their depth. When increasing the input dimensionality the increasing smoothness provided by deep neural networks was shown to be a combination of their depth and the amount of energy stored by these networks, thus it is plausible to hypothesize that the results obtained in [51] are more general than expected, that is deep networks have the guarantee to avoid the curse of dimensionality beyond the specific case of compositional functions.

7. Conclusions

Given the basic rules for assembling the elements conforming an artificial neuron, the scheme of computations carried out by the neuronal model together with its interactions when embedded in an artificial neural network, a theoretical model grounded in Statistical Mechanic was derived.The theoretical model was aimed at characterizing the physical system represented by artificial fully connected feedforward neural networks. Each artificial network is considered as an equilibrium state that is characterized by a thermodynamic signature that is unique according to the model. Furthermore, the evolution of the system from one state to another is assumed that occurs according to the law of

increase of the entropy. The macroscopic thermodynamic behavior of these models is studied using the theoretical model through a set of theoretical transformations of state that follow the second law of thermodynamics which permit to study the changes of energy of the system, how the structure of the phase space varies when the system passes from one equilibrium state to another, and also the stability of the equilibrium states. In other words, the theoretical model permits to interpret machine learning properties of deep learning machines such as their learning and generalization performance, or the complexity of the loss landscape in terms of thermodynamic concepts. Thereby, allowing the use of the well known machinery of Thermodynamics to understand the behavior of deep learning machines. To this end, the emphasis was put on the analysis of three scenarios that take into consideration the different possibilities that arise when considering the structural parameters defining the complexity of a network: The total number of neurons and the total number of adaptive weights. In the first scenario the structural complexity of the networks is fixed (i.e, the total number of units and adaptive weights). In the second scenario the effective complexity of the networks (i.e., the total number of adaptive weights) is fixed but the total number of units is permitted to vary. Finally, in the third scenario the total number of neurons is fixed while the effective complexity of the networks is permitted to vary. Each scenario impose different restrictions on the structural parameters of the networks leading to important differences in the structure of the phase space of the networks, network storage capacities, and thus in their learning an generalization performance. Finally, the thermodynamic efficiency with which information can be processed by these models was also assessed. The results of these analysis suggested the following conclusions:

1. The influence of the topology of these models in both their thermodynamic efficiency , and in their resulting learning and generalization performance goes beyond what has previously been assumed. The learning and generalization performance of these models cannot be explained only by the complexity of the learning task (i.e, the probability distribution of the input patterns) and/or the learning protocol used but rather the speciic topological characteristics of the neural architecture considered.
2. Deep learning machines (shallow or deep) are characterized by a thermodynamic signature that is unique to each neural architecture. This signature is shown to be related with the the power of the architecture (e.g., its capacity), the complexity of the geometrical structure of the loss landscape, its learning and generalization capabilities, and the thermodynamic efficiency with which information can be processed. Furthermore, it is shown that it permits in combination with the law of increase of the entropy to estimate and compare the capabilities of different neural architectures.

In summary, deep learning machines are viewed as theoretical bodies characterized by a thermodynamic signature. This thermodynamic signature is specific to each network (shallow or deep), and permits to characterize the thermodynamic macroscopic behavior of these models which is implicitly linked to their learning and generalization performance but also to the thermodynamic efficiency with which information can be processed by these machine learning models.

Perhaps the most important implications of the model is the introduction of a reasonable explanation for a persistent and unresolved question raised after the universal approximation theorem concerning under which conditions may deep networks be more powerful than shallow networks. The results provided by the model suggest that typically the equilibrium states represented by shallow networks store larger amounts of energy, and present lower heat capacities compared to deep networks as a result of its extreme topology (all the units concentrated in a single layer) evidencing the existence of a more complex geometrical structure of the loss landscape and poorer thermodynamic efficiencies but also leading in some cases to equilibrium states that are stable within certain limits (metastable), thereby having a profound influence in the resulting learning and generalization performance of these models as well as in their thermodynamic efficiency.

However, because of the interplay between the structural parameters that govern the topology of these models (e.g, the dimensionality of the input space) the properties of these models may change

leading to shallow networks with better learning and generalization performance compared to deep networks.

Undoubtedly, deep learning machines have given rise to an active and ongoing debate principally driven by the difficulties that has been found to understand their scientific foundations. Many researchers have attempted to understand the apparently unreasonable properties of these models motivating the proliferation of fierce debates in the last few years. Clearly, the large number of parameters characterizing these models, and understanding their complex interactions have contributed to the difficulties associated with their study. The present study is expected to provide not only some answers to the current debate between shallow versus deep learning machines capabilities debate but also to contribute from a novel perspective to the theory of deep learning machines.

"

Supplementary Materials: The following are available online at www.mdpi.com/link, Figure S1: title, Table S1: title, Video S1: title.

Acknowledgments: The author would like to thank Elka Korutcheva for inviting me to give a seminar of early stages of this research at the Departament of Fundamental Physics of the Faculty of Sciences of the UNED University (Madrid) in May 2017 and for the fruitful discussions that took place with her in combination with Javier Rodriguez-Lagunas during the seminar. The author furthermore thanks Javier Rogriguez-Lagunas for performing the regression fit shown in figure (10), and Cyril Furtlehner for short but helpful discussions on Thermodynamics. The initial part of this research was supported by DARPA under the supervision of Hava Siegelmann and Robert Kozma, and performed during a postdoctoral research position at the College of Information and Computer Sciences (CICS) of the University of Massachussets, Amherts (UMASS) in 2016.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VC Vapnik-Chevornenkis dimension

Appendix A. Definitions and Notations

The goal of these sections is to provide the definitions of the mathematical concepts and notations used throughout this paper (section A), the derivations of the most important mathematical expressions that appear in the paper (section B), and some supporting information (tables and figures) that complement the information presented in this study (section C).

Definition A.1 (Combinatorial class). A combinatorial class is a finite or denumerable set in which a size function is defined, satisfying the following conditions: (1) the size of an element is a non-negative integer. (2) The number of elements of any given size is finite. If A is a class, the size of an element $\alpha \in A$ is denoted as $|\alpha|$.

Definition A.2 (Counting sequence). The counting sequence of a combinatorial class A is the sequence of integers $(A_n)_{n\geq 0}$ where A_n is the number of objects in class A that have size n.

Definition A.3 (Generating function). The generating function of a sequence A_n is the formal power series, where $w_n = 1$ for the ordinary case and $w_n = n!$ for the exponential case.

$$A(u) = \sum_{n\geq 0}^{+\infty} A_n \frac{u^n}{w_n} = \sum_{\alpha \in A} \frac{u^{|\alpha|}}{w_n} \tag{A1}$$

Definition A.4 (Coefficient extraction). We generally let $A_n = w_n[u^n]A(u)$ denote the operation of extracting the coefficient of u^n in the formal power series (A.1).

$$A_n = w_n[u^n]\left(\sum_{n\geq 0}^{+\infty} A_n \frac{u^n}{w_n}\right) = \frac{1}{2\pi i}\oint_C w_n A(u)\frac{du}{u^{n+1}} \tag{A2}$$

Definition A.5 (Integer composition). A composition of an integer n is a sequence $(x_1, x_2, x_3, \ldots, x_k)$ of integers (for some k) such that:

$$n = x_1 + x_2 + ... + x_k \quad \text{and} \quad x_j \geq 1 \quad 1 \leq j \leq k \tag{A3}$$

The x_i are called the summands or the parts and the quantity n is called the size.

$$n = x_1 + x_2 + ... + x_k \quad \text{and} \quad x_1 \geq x_2 \geq ... \geq x_k \tag{A4}$$

The x_i are called the summands or the parts and the quantity n is called the size.

Definition A.6 (Binomial Convolution). Let $a(u)$, $b(u)$ and $c(u)$ be exponential generating functions, with $a(u) = \sum_{n\geq 0}^{+\infty} a_n \frac{u^n}{n!}$, and so on. The binomial convolution formula is:

$$if \quad a(u) = b(u)c(u), \quad then \quad a_n = \sum_{k=0}^{n}\binom{n}{k} b_k c_{n-k} = b_n \oplus c_n \tag{A5}$$

where $\binom{n}{k} = n!/(k!(n-k)!)$ represents a binomial coefficient.

Definition A.7 (Cartesian Product of Combinatorial classes). The cartesian product construction applied to two combinatorial classes B and C forms ordered pairs, $A = B \times C \quad \Leftrightarrow \quad A = \{\alpha = (\beta, \gamma) | \beta \in B, \gamma \in C\}$, with the size of a pair $\alpha = (\beta, \gamma)$ being defined by $|\alpha|_A = |\beta|_B + |\gamma|_C$. By considering all possibilities, the counting sequences corresponding to A, B, C are related by the convolution relation:

$$A_n = \sum_{k=0}^{n} B_k C_{n-k} \tag{A6}$$

Furthermore, the convolution relation is the formula for a product of two power series: $A(u) = B(u)C(u)$.

Definition A.8 (Admissible Combinatorial Constructions). Let Φ be an m-ary construction that associates to any collection of classes $B^1, B^2, \cdots, B^m$ a new class $A = \Phi[B^1, B^2, \cdots, B^m]$. The construction Φ is admissible if and only if the counting sequence A_n of A only depends on the counting sequences $B_n^1, B_n^2, \cdots, B_n^m$ of $B^1, B^2, \cdots, B^m$. For such an admissible construction, there then exists a well-defined operator Ψ acting on the corresponding ordinary generating functions $A(u) = \Psi[B^1(u), B^2(u), \cdots, B^m(u)]$.

- **Cartesian product:** This construction $A = B \times C$ forms all possible ordered pairs in accordance with definition A.7. The size of a pair is obtained additively from the size of components.
- **Sequence construction:** If B is a combinatorial class then the sequence class $A = SEQ(B)$ is defined as the infinite sum:

$$A = SEQ(B) = \{\epsilon\} + B + (B \times B) + (B \times B \times B) + \cdots \tag{A7}$$

The construction $A = SEQ(B)$ defines a proper class satisfying the finiteness condition for sizes if and only if B contains no object of size 0, with ϵ being a neutral structure (of size 0) which plays a similar role to that of the "empty" word in formal language theory. In other words, we have $A = \{(\beta_1, \beta_2, \cdots, \beta_l) | l \geq 0, \beta_j \in B\}$, where the neutral structure corresponds to $l = 0$. It follows that the size of an object $\alpha \in A$ is to be taken as the sum of the sizes of its components: $\alpha = (\beta_1, \beta_2, \cdots, \beta_l) \Rightarrow |\alpha| = |\beta_1| + \cdots + |\beta_l|$. Therefore, the ordinary generating function associated to the sequence construction is as follows:

$$A = SEQ(B) \quad \Rightarrow \quad A(u) = 1 + B(u) + B^2(u) + \cdots = \frac{1}{1 - B(u)} \tag{A8}$$

Similarly, sequences whose number of components are exactly k are expressed as SEQ_k and it admits the translation into ordinary generating functions:

$$A = SEQ_k(B) = \overbrace{B \times B \times \cdots \times B}^{k \text{ times}} \quad \Rightarrow \quad A(u) = B(u)^k \tag{A9}$$

Appendix B. Derivation of the Most Important Equations

Equation (5) (The generating function of the integer composition modeling the operation of the linear combiner of an artificial neuron). Let us denote as $C_p(n)$ the counting sequence of the combinatorial class C_p describing the composition with summands bounded in number and size. The number of summands of the composition is equal to the number of inputs to the artificial neuron, i.e., n, whereas its values are restricted to the set $1, 2, \ldots, mp$. To express this fact in terms of the language of admissible combinatorial constructions, the first step is to define the combinatorial class associated to the integers I. Let the size of each integer its value, then the counting sequence I_n associated to the class I is $I_n = 1$ for $n \geq 1$ corresponding to the fact that there is exactly one object in I for each size $n \geq 1$. If integers are represented by small balls, one has $I = \{1, 2, 3, \cdots\} = \{\bullet, \bullet\bullet, \bullet\bullet\bullet, \cdots\}$. Taking these considerations into account, the combinatorial description of the class C_p can be expressed as follows:

$$C_p = SEQ_n(I^{\{1..mp\}}) = \overbrace{I^{\{1..mp\}} \times I^{\{1..mp\}} \times I^{\{1..mp\}} \times \cdots \times I^{\{1..mp\}}}^{j \text{ times}} \tag{A10}$$

Thus, from the symbolic language description of the combinatorial class C_p the obtention of the the generating function $C_p(z)$ is straightforward:

$$C_p(z) = [I^{\{1..mp\}}(z)]^n = [z + z^2 + z^3 + \ldots + z^{mp}]^n = \frac{z^n}{(1-z)^n}(1 - z^{mp})^n \tag{A11}$$

Equations (11) and (12) (The coefficients associated to the generating function of the combinatorial class X modeling the energy states of an artificial neuron). Let us first to rewrite the expression of the generating function $X(z)$:

$$X(z) = \Big(\sum_{k=n}^{\delta} C_p(k) \Big) z^{\mu} + \Big(\sum_{k=1+\delta}^{nmp} C_p(k) \Big) z^{\mu+\Delta} \tag{A12}$$

The ordinary generating function of the combinatorial class C_p describing the integer composition that models the linear combiner of a neuron can be rewritten in terms of the product of two generating functions $A(z)$ and $B(z)$, i.e, $C_p(z) = A(z)B(z)$ that read:

$$A(z) = \sum_{l \geq 0}^{+\infty} a_l z^l = \frac{z^n}{(1-z)^n} \qquad \text{and} \qquad B(z) = \sum_{l \geq 0}^{+\infty} b_l z^l = (1 - z^{mp})^n \tag{A13}$$

Thus, the coefficients of the generating function $C_p(z)$ can be expressed in terms of the convolution $C_p(k) = \sum_{l=0}^{k} b_l c_{k-l}$. In turn, using the definition A.4, the coefficients a_l can be written as:

$$a_l = [z^l]A(z) = \binom{l-1}{n-1} \tag{A14}$$

whereas, the coefficients b_l using the expression of the Newton binom $(1-z^{mp})^n = \sum_{l=0}^{n} \binom{n}{l}(-1)^l z^{mpl}$ can be written as:

$$b_l = [z^l]B(z) = \binom{n}{l}(-1)^l \qquad l = 0, mp, 2mp, 3mp, \ldots \tag{A15}$$

Thus, the expression for the coefficients $C_p(k)$ can be finally written as:

$$C_p(k) = \sum_{l=0}^{k}(-1)^l \binom{n}{l}\binom{k-mpl-1}{n-1} \tag{A16}$$

Substituting the coefficient (A16) in expression (A12) one

$$\lambda^1_{m,p,n} = \sum_{k=n}^{\delta}\sum_{l=0}^{k}(-1)^l \binom{n}{l}\binom{k-mpl-1}{n-1} \tag{A17}$$

and

$$\lambda^2_{m,p,n} = \sum_{k=\delta+1}^{nmp}\sum_{l=0}^{k}(-1)^l \binom{n}{l}\binom{k-mpl-1}{n-1} \tag{A18}$$

Equations (16) and (23) (The partition function of a feedforward fully connected deep neural network with L layers and sigmoidal units). The generating function $L_k(z)$ of the combinatorial class L_k that modeled the possible energy values of a generic layer of a deep network composed of g_k units is a functional of the generating functions of the units that it contains. However, for the case of a feedforward fully connected deep neural network the generating functions of the artificial neurons of the layer are identical, thus, the equation (14) can be rewritten as:

$$L_k(z) = \prod_{j=1}^{g_k} X_j(z) = (X_k(z))^{g_k} = z^{\mu g_k}\left(\lambda^1_{m,g_{k-1}} + \lambda^2_{m,g_{k-1}} z^{\Delta}\right)^{g_k} \tag{A19}$$

Substituting the resulting expression of equation (A19) in equation (13) one gets:

$$M(z) = X_1(z)^{g_1}\prod_{k=2}^{L} X_k(z)^{g_k} = z^{\mu g_1}\left(\lambda^1_{m,p,g_0} + \lambda^2_{m,p,g_0} z^{\Delta}\right)^{g_1}\prod_{k=2}^{L} z^{\mu g_k}\left(\lambda^1_{m,g_{k-1}} + \lambda^2_{m,g_{k-1}} z^{\Delta}\right)^{g_k} \tag{A20}$$

Therefore, equation (15) can be finally obtained by simply evaluating equation (A20) in $z = e^{-\beta}$. Finally, to derive a recursive expression for the equation of the partition function (15) the first step is to express the generating function of the network recursively:

$$
\begin{aligned}
M_1(z) &= [X_1(z)]^{g_1} \\
M_2(z) &= [X_1(z)]^{g_1}[X_2(z)]^{g_2} = M_1(z)[X_2(z)]^{g_2} \\
M_3(z) &= [X_1(z)]^{g_1}[X_2(z)]^{g_2}[X_3(z)]^{g_3} = M_2(z)[X_3(z)]^{g_3} \\
&\bullet \\
&\bullet \\
&\bullet \\
M_k(z) &= \prod_{i=1}^{k-1} M_i(z)[X_k(z)]^{g_k} = M_{k-1}(z)[X_k(z)]^{g_k}
\end{aligned}
\tag{A21}
$$

Using the recursion (A21) together with the fact that $Z_k = M_k(z)|_{z=e^{-\beta}}$ the equation (23) is finally obtained.

Equation (25) (The partition function of a feedforward fully connected deep neural network of L layers with sigmoidal units excepting the output layer L that is composed of units with linear activation functions). Using expressions (A21) and (5) the equation of the partition function one gets:

$$
Z = M^L(z)|_{z=e^{-\beta}} = M^{L-1}(e^{-\beta})[X_L(e^{-\beta})]^{g_L} \tag{A22}
$$

where the expression for $X_L(z)$ gets now the form:

$$
X_L(z) = Cp(z) = \frac{z^{g_{L-1}}}{(1-z)^{g_{L-1}}}(1-z^{2m})^{g_{L-1}} \tag{A23}
$$

Similarly, from expression (25) the equations of the free energy, the entropy, the internal energy and its fluctuations, and the specific heat can be easily derived leading respectively to the expressions:

$$
\begin{aligned}
F &= g_L g_{L-1}(m+\frac{1}{2}) - \frac{g_L g_{L-1}}{\beta} log\Big(\frac{\sinh(\beta m)}{\sinh(\frac{\beta}{2})}\Big) - \mu \sum_{i=1}^{L} g_i - \\
&- \frac{g_1}{\beta}\log\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right) - \frac{1}{\beta}\sum_{i=2}^{L} g_i \log\left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right)
\end{aligned}
\tag{A24}
$$

$$
\begin{aligned}
S &= g_L g_{L-1} log\Big(\frac{\sinh(\beta m)}{\sinh(\frac{\beta}{2})}\Big) - \frac{\beta m g_L g_{L-1}}{\tanh(\beta m)} + \frac{\beta g_L g_{L-1}}{2\tanh(\frac{\beta}{2})} + \\
&+ g_1 \log\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right) + \sum_{i=2}^{L} g_i \log\left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right) + \\
&+ \beta\Delta e^{-\Delta\beta}\frac{g_1\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} + \beta\Delta e^{-\Delta\beta}\sum_{i=2}^{L}\frac{g_i\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}}
\end{aligned}
\tag{A25}
$$

$$
\begin{aligned}
U &= g_L g_{L-1}(m+\frac{1}{2}) - \frac{m g_L g_{L-1}}{\tanh(\beta m)} + \frac{g_L g_{L-1}}{2\tanh(\frac{\beta}{2})} + \\
&+ \mu\sum_{i=1}^{L-1} g_i + \beta\Delta e^{-\Delta\beta}\frac{g_1\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0}e^{-\Delta\beta} + \lambda^2_{m,p,g_0}} + \\
&+ \beta\Delta e^{-\Delta\beta}\sum_{i=2}^{L-1}\frac{g_i\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}}e^{-\Delta\beta} + \lambda^2_{m,g_{i-1}}}
\end{aligned}
\tag{A26}
$$

$$
\begin{aligned}
(\Delta U)^2 &= \Delta^2 e^{-\Delta\beta} g_1 \frac{\lambda^1_{m,p,g_0}\lambda^2_{m,p,g_0}}{\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right)^2} + \\
&+ \Delta^2 e^{-\Delta\beta} \sum_{i=2}^{L-1} g_i \frac{\lambda^1_{m,g_{i-1}}\lambda^2_{m,g_{i-1}}}{\left(\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}\right)^2} + \\
&+ m^2\left(1 - \frac{1}{\tanh(\beta m)^2}\right) - \frac{1}{4}\left(1 - \frac{1}{\tanh(\frac{\beta}{2})^2}\right)
\end{aligned} \tag{A27}
$$

Finally, from expression of the entropy (A25) the equations of the specific heat can be easily derived leading respectively to the following set of equations:

$$C_v = -\beta\frac{\delta S}{\delta\beta} = \textstyle\sum_{i=1}^{5} C_{v_i} \tag{A28}$$

whereas the terms C_{v_i} $(1 \leq i \leq 5)$ read respectively:

$$C_{v_1} = \Delta\beta e^{-\Delta\beta} g_1 \frac{\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}} \tag{A29}$$

$$C_{v_2} = \Delta\beta e^{-\Delta\beta} \sum_{i=2}^{L-1} g_i \frac{\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_{i-1}} + e^{-\Delta\beta}\lambda^2_{m,g_{i-1}}} \tag{A30}$$

$$
\begin{aligned}
C_{v_3} &= -\Delta\beta(1-\Delta\beta)e^{-\Delta\beta} g_1 \left(\frac{\lambda^1_{m,p,g_0}\lambda^2_{m,p,g_0}}{\left(\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}\right)^2} + \left(\frac{\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}}\right)^2\right) - \\
&- g_1\left(\frac{\Delta\beta e^{-\Delta\beta}\lambda^2_{m,p,g_0}}{\lambda^1_{m,p,g_0} + e^{-\Delta\beta}\lambda^2_{m,p,g_0}}\right)^2
\end{aligned} \tag{A31}
$$

$$
\begin{aligned}
C_{v_4} &= \Delta\beta(1-\Delta\beta)e^{-\Delta\beta} \sum_{i=2}^{L-1} g_i \left(\frac{\lambda^1_{m,g_i}\lambda^2_{m,g_i}}{\left(\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}\right)^2} + \left(\frac{\lambda^2_{m,g_{i-1}}}{\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}}\right)^2\right) - \\
&- \sum_{i=2}^{L-1} g_i \left(\frac{\Delta\beta e^{-\Delta\beta}\lambda^2_{mg_i}}{\lambda^1_{m,g_i} + e^{-\Delta\beta}\lambda^2_{m,g_i}}\right)^2
\end{aligned} \tag{A32}
$$

$$C_{v_5} = \beta^2 g_L g_{L-1}\left(\frac{1}{4\sinh(\frac{\beta}{2})^2} - \frac{m^2}{\sinh(\beta m)^2}\right) \tag{A33}$$

Appendix C. Support Graphs

The following graphs show the generalization performance obtained for the set of network architectures studied in the Case studies I and III using a training algorithm with better convergence properties than standard backpropagation algorithm. Specifically, the figure (A1) correspond to the generalization performance obtained for the set of architectures used in the first case study (networks with identical structural complexity) for the scaled conjugate gradients algorithm. Similarly, the figure (A2) show the generalization performance obtained when using the scaled conjugate gradients algorithm for the set of test architectures used in case study III.

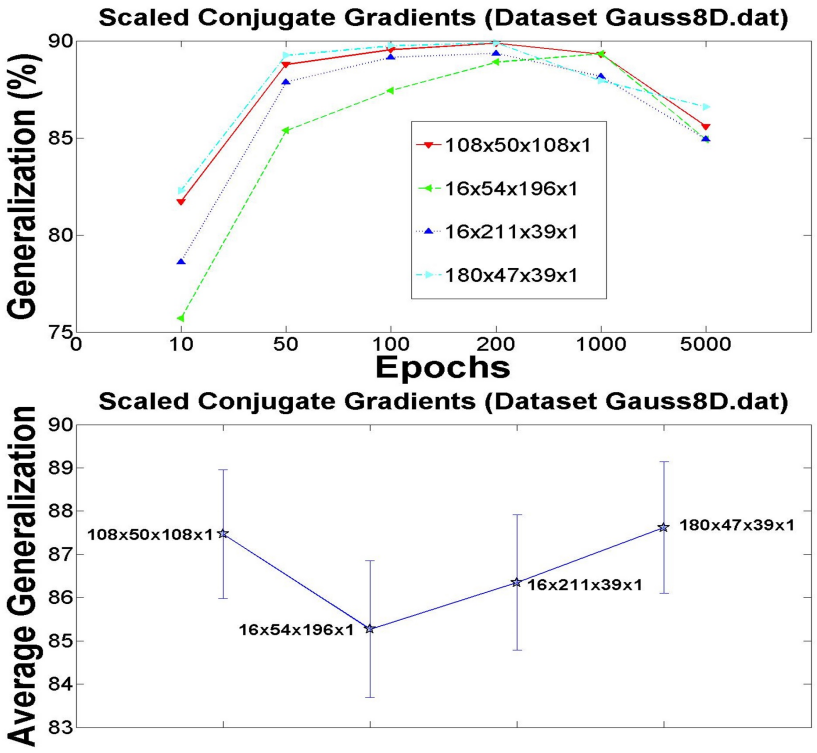

Figure A1. Graphical illustration of the generalization performance obtained for the group of networks M_{3b} of table 1 with an identical structural complexity using as learning algorithm the scaled conjugate gradients algorithm [11,30] with the dataset *Gauss8d.dat* [12]. The top part represents the evolution of the generalization performance of the networks sampled at 10, 50,100,200, 1000 and 5000 epochs respectively. The botton part represents the average generalization (the standard deviation is represented as an error bar) obtained when averaging the generalization at the sampling points described before. The goal is to assess the typical behavior of the generalization independently of the number of epochs used to train the networks. Compared to the graph (7) the generalization performance is now in direct correspondence with the entropy value attained by each network (see table 1), thus, evidencing the existence of a more complex geometrical structure of the loss landscape for the networks 108x50x108x1, and 16x54x196x1 that is circumvented in this case because of the fact of using a learning protocol with better convergence properties.

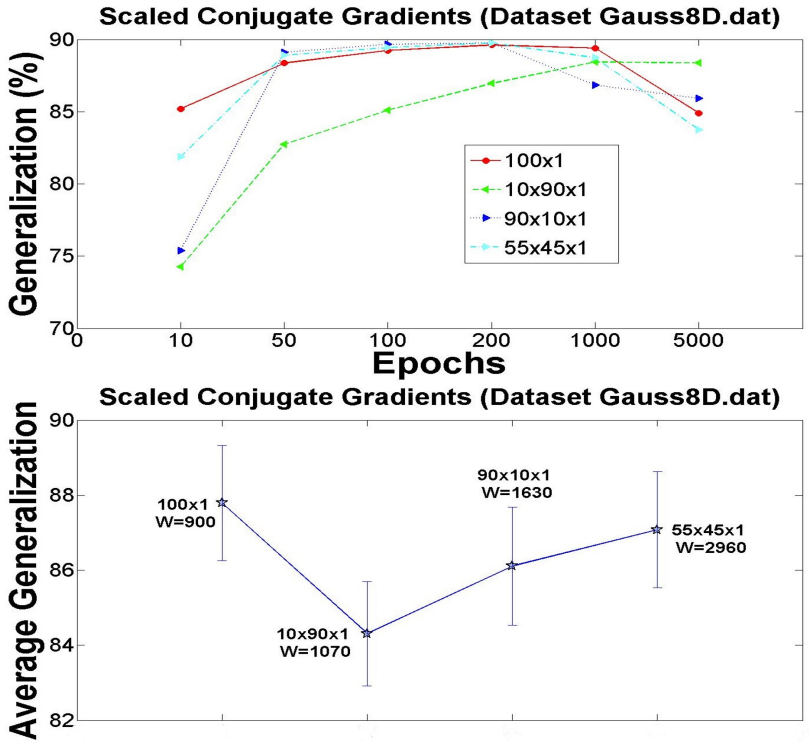

Figure A2. Graphical illustration of the generalization performance obtained for the group of networks M_1 and M_2 of table 3 with an identical number of units using as learning protocol the scaled conjugate gradients algorithm with the dataset *Gauss8d.dat* [12]. The top part represents the evolution of the generalization performance of the networks. The botton part represents the average generalization (and its standard deviation represented as an error bar) following in both cases an identical procedure as described in figure (14). Compared to the graph (14) the generalization performance is now in direct correspondence with the entropy value attained by each network (see table 3) excepting for the shallow network 100x1 thus, corroborating the existence of a more complex geometrical structure of the loss landscape for the network 55x45x1 that is circumvented in this case because of the fact of using a learning protocol with better convergence properties. Surprisingly, the shallow network outperforms the average generalization of networks attaining higher entropy values.

References

1. Anselmi F., & Poggio T. (2016).*Visual Cortex and Deep Networks*. MIT Press 2016.
2. Ba J., and Caruana R. (2014). Do Deep Nets Need to be Deep? In NIPS, pages 2654-2662.
3. Baldi P. (2018). Deep Learning in biomedical data science. *Annual Review of Biomedical Data Science*,1:181:205.
4. Baldi P., & Vershynin R. (2019). The Capacity of Feedforward Neural Networks. *arXiv:1611.03530*.
5. Barkai E., Hansel D., & Kanter I. (1990). Statistical Mechanics of a Multilayered Neural Network. *Physical Review Letters*, Vol 65 (18).
6. Basri R., & Jacobs D. (2016). Efficient Representation of Low-Dimensional Manifolds using Deep Networks. Arxiv preprint: 1602.04723.
7. Bellman R. (1961). *Adaptive Control Processes: A Guided Tour*, Princeton, NJ: Pinceton University Press.
8. Bennett C. H., (1982). The Thermodynamics of Computation. A Review. *International Journal of Theoretical Physics*, 21(12), 905-940.
9. Bianchini M., and Scarselli F. (2014). On the Complexity of neural network classifiers: A comparison between shallow and deep architectures. IEEE Transactions on Neural Networks and Learning Systems, 25(8).
10. Bianconi G., (2009). Entropy of Network Ensembles. *Physical Review E*, vol. 79, 036114. DOI:10.1103/PhysRevE.79.036114.
11. Bishop C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.
12. Blayo F., Cheneval Y., Guérin-Dugué A., et al. (1995), Enhanced Learning for Evolutive Neural Architecture ESPRIT Basic Research Project Number 6891, Deliverable R3-B4-P, Task B4 (Benchmarks).
13. Boccaleti S., Bianconi G., Criado R., et al. (2014). The Structure and Dynamics of Multilayer Networks. Physics Reports 544, 1-122.
14. Bullmore E., Sporns O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci.10, 186-198.
15. Chandler D., (1987). *Introduction to Modern Statistical Mechanics*. Oxford University Press.
16. Chui C., Li X., & Mhaskar H. (1994). Neural networks for localized approximation. *Mathematics and Computation*, vol. 63, no. 208, pp. 607-623.
17. Chui C., Li X., & Mhaskar H. (1996). Limitations of the approximations capabilities of neural networks with one hidden layer.*Advances in Computational Mathematics*, vol. 5, no.1, pp. 233-243.
18. Cohen N., Sharir O., and Shashua A. (2015). On the Expressive Power of Deep Learning: A Tensor Analysis.
19. Comtet, L. (1974). *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Dordrecht, Holland: Reidel Publishing Company.
20. Cybenko G. (1989). Approximation by Superpositions of Sigmoidal Functions. *Mathematics of Control, Signal and Systems*, 2(4):303-314.
21. Delalleau O., & Bengio Y. (2011). Shallow vs. Deep Sum-product Networks. In NIPS, page 666-674.
22. Dogorovtsev S.N, Goltsev A., & Mendes J.F.F. (2008). *Review of Modern Physics*, 80, 1275.
23. Elad M., "Deep, deep trouble." siam news, https://sinews.siam.org/Details-Page/deep-deep-trouble, 2017.
24. Eldan R., & Shamir O. (2016). The Power of Depth for Feed-forward Neural Networks. In *Conference on Learning Theory*, pp 907-940.
25. Efron B. (1979). Bootstrap Methods: Another Look at the Jacknife, *The Annals of Statistics*, Vol. 7, No. 1, pp. 1-26.
26. Engel, A., & Van den Broeck, C. (2001). Statistical mechanics of learning. Cambridge,UK: Cambridge University Press.
27. Flajolet, P., & Sedgewick R. (2009). *Analytic Combinatorics*. Cambridge, UK: Cambridge University Press.
28. Gardner E. (1988). The Space of Interactions in Neural Network Models. *Journal of Physics A*, Vol 65 (18).
29. Hartmann, A.K., & Weigt, R. (2005). *Phase Transitions in Combinatorial Optimization Problems: Basic Algorithms and Statistical Mechanics*. Weinheim, DE: Wiley-Vch Verlag Gmbh & Co.
30. Haykin S, (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.
31. Hermann C. (2005). *Statistical Physics (Including Applications to Condensed Matter)*. New York, USA: Springer.
32. Hornik, K. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359-366.
33. Khadivi P., Tandon R., and Ramakrishnan N. (2016). Flow of Information in Feed-forward Deep Neural Networks. Arxiv preprint: 1603.06220v1.

34. Landau L.D., & Lifshitz E.M. (1958). *Statistical Physics, volume 5 of Course of Theoretical Physics*, Addison-Wesley, Reading. Massachusetts,USA.
35. LeCunn Y., Bengio Y, & Hinton G. (2015). Deep Learning. *Nature* 521, 436-444.
36. Li B., & Saad D., (2018). Exploring the Function Space of Deep-Learning Machines. *Nature*.
37. Liang T., Poggio T., Rakhlin A., & Stokes J. (2019). Fisher-Rao Metric, Geometry, and Complexity of Neural Networks. *Proceedings of the* 22^{nd} *International Conference on Artificial Intelligence and Statistics AISTATS*, 2019,Naha, Okinawa, Japan.
38. Lieb E.H., & Yngvason J. (1999). The Physics and Mathematics of the Second Law of Thermodynamics. *Physics Reports* 310, 1-96.
39. Malzahn D., & Engel A. (1999). Correlations between Hidden Units in Multilayer Neural Networks and Replica Symmetry Breaking. *Physical Review E*, Vol 60, 2097.
40. Mhaskar H. (1996). Neural Networks for optimal approximation of smooth and analytical functions. *Neural Computation*, vol.8, no.1., pp. 164-167.
41. Mhaskar H., Liao Q., & Poggio T. (2016). Learning Functions: When Is Deep Better than Shallow?. *Arxiv preprint: 1603.00988v4*.
42. Mehta P., & Schwab D.J. (2014). An Exact Mapping between the Variational Renormalization Group and Deep Learning. *Arxiv preprint: 1410.3831*.
43. Monasson R., & Zecchina R. (1995). Weight Space Structure and Internal Representations: A Direct Approach to Learning and Generalization in Multilayer Neural Networks. *Physical Review Letters*, Vol 75 (12).
44. Montufar G.F., Pascanu R., Cho K., and Bengio Y. (2014). On the Number of Linear Regions of Deep Neural Networks. In NIPS, pages 2924-2932.
45. Neyshabur B., Bhojanapali S., McAllester D., & Srebro N., (2017). Exploring generalization in deep learning. In NIPS, pp. 5947-5956.
46. Niyogi P., & Girosi F. (1996). On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*,vol. 8., pp. 819-842.
47. Opper M. (1994). Learning and Generalization in Two-Layer Neural Network: The Role of the Vapnik-Chervonenkis Dimension. *Physical Review Letters*, Vol 72 (13).
48. Opper M., & Haussler D., (1995).Bounds for Predictive Errors in the Statistcal Mechanics of Suepervised Learning. *Physical Review Letters*, Vol 75 (20), 3772.
49. Pinkus A. (1999). Approximation theory of the mlp model in neural networks. *Acta Numerica*, vol. 5, no.1, pp.233-243.
50. Mhaskar, H., & Poggio T., (2016). Deep vs. Shallow Networks: An Approximation Theory Perspective. *Arxiv preprint: 1608.03287v1, CBMM Memo No. 054*.
51. Poggio T., Mhaskar, H., Rosasco L., Miranda B., & Liaoi Q. (2017). Theory I: Why and When Can Deep Networks Avoid the Curse of Dimensionality. *Arxiv preprint: 1703.09833v2, CBMM Memo No. 058*.
52. Poggio T., & Liaoi Q. (2017). Theory II: Landscape of the empirical risk in deep learning. *Arxiv preprint: 1703.09833, CBMM Memo No. 066*.
53. Poggio T., Kawaguchi K., Liaoi Q., Miranda B., Rosasco L., Boix X., Hidary J., & Mhaskar H., (2018). Theory III: explaining the non-overfitting puzzle. *Arxiv preprint: 1703.09833, CBMM Memo No. 073*.
54. Poggio T., Kawaguchi K., Liaoi Q., Miranda B., Rosasco L., Boix X., Hidary J., & Mhaskar H., (2019). Theory III: Dynamics and Generalization in Deep Networks. *Arxiv preprint: 1708.01422v3, CBMM Memo No. 090*.
55. Poole B., Lahiri S., Rahgu M., Sohl-Dickstein J., and Ganguli S. (2016). Exponential Expressivity in Deep Neural Networks through Transient Chaos. *Arxiv preprint: 1606.05340v2*.
56. Schmidhuber J. (2015). Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85-117.
57. Schwarze H., & Hertz J., (1993). Generalization in Fully Connected Committee Machines. *Europhysics Letters*, Volume 21, No.7 pp. 785.
58. Silver J., Schrittwieser J., Simonyan k., Antonoglou I., Huang A., Guez A., Hubert T., Baker L., Lai M., Bolton A., , et al. Mastering the game of go without human knowledge (2017). *Nature*,550(7676):354.
59. Lin. M., Tegmark M., & Rolnick D. (2016). Why does deep and cheap learning work so well?. *Arxiv preprint: 1608.08225*.
60. Telgarsky M. (2015). Representation Benefits of Deep Feed-forward Networks. *Arxiv preprint: 1509.08101.*
61. Tishby N., F. C. Pereira, & Bialek W., (1999). The information bottleneck method. in Proceedings of 37th Annual Allerton Conference on Communication, Control and Computing.

62. Tishby N., & Zaslavsky N., (2015). Deep Learning and the Information Bottleneck Principle, in Proceedings of the *2015 IEEE Information Theory Workshop (ITW)*
63. Shwartz-Ziv R., & Tishby N., (2017). Opening the Black Box of Deep Neural Networks via Information. *Arxiv preprint: 1703.00810v3*
64. Piran Z., Shwartz-Ziv R., & Tishby N., (2020). The Dual Information Bottleneck method,*Arxiv preprint: 2006.04641v1*.
65. Tritt, T.M. (Ed.), (2004). *ThermalConductivity: Theory, Properties, and Applications*. New York, NY: Kluwer Academic & Plenum Publishers.
66. Raghu M., Poole B., Kleinberg J., Ganguli S., & Sohl-Dickstein J., (2017). On the Expressive Power of Deep Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70.
67. Reichl, L. E., (1998). *A Modern Course in Statistical Physics*. New York, NY: John Wiley & Sons.
68. Vapnik V., (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag.
69. Baldasi C., Pittorino F., & Zecchina R., (2019). Shaping the learning landscape in neural networks around wide flat minima.*arXiv:1905.07833*.
70. Zhang, C., Bengio S., Hardt M., Recht, B., & Vinyals O. (2016). Understanding deep learning requires rethinking generalization.*arXiv:1611.03530*.