Foundations of Physics (2021) 51:94
https://doi.org/10.1007/s10701-021-00503-3

Emergent Quantumness in Neural Networks

Mikhail I. Katsnelson[1] · **Vitaly Vanchurin**[2,3]

Received: 18 December 2020 / Accepted: 17 September 2021 / Published online: 28 September 2021

Abstract

It was recently shown that the Madelung equations, that is, a hydrodynamic form of the Schrödinger equation, can be derived from a canonical ensemble of neural networks where the quantum phase was identified with the free energy of hidden variables. We consider instead a grand canonical ensemble of neural networks, by allowing an exchange of neurons with an auxiliary subsystem, to show that the free energy must also be multivalued. By imposing the multivaluedness condition on the free energy we derive the Schrödinger equation with "Planck's constant" determined by the chemical potential of hidden variables. This shows that quantum mechanics provides a correct statistical description of the dynamics of the grand canonical ensemble of neural networks at the learning equilibrium. We also discuss implications of the results for machine learning, fundamental physics and, in a more speculative way, evolutionary biology.

1 Introduction

Despite the obvious success of quantum mechanics in description of our physical world, its conceptual status is still a subject of hot debates, see Refs. [1–5], to name just a few contemporary books; more references can be found in the recent papers [6, 7]. As a result, many so-called "no-go theorems" were constructed (e.g. Bell's inequalities [8]) to rule out the existence of a hidden classical world beyond quantum mechanics [9]. Recently the idea of "emergent quantumness" was reincarnated in the programs like "the world as a matrix" [10], "the world as a cellular automaton"

✉ Vitaly Vanchurin
vvanchur@d.umn.edu

Mikhail I. Katsnelson
m.katsnelson@science.ru.nl

[1] Institute for Molecules and Materials, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

[2] Department of Physics, University of Minnesota, Duluth, MN 55812, USA

[3] Duluth Institute for Advanced Study, Duluth, MN 55804, USA

[11, 12] and "the world as a neural network" [7, 13]. An alternative approach was suggested in the "logical inference" program [14–16], where quantum mechanics is considered as a purely phenomenological way to describe the results of repeated experiments assuming that (1) we cannot control all the details relevant for these experiments, (2) our description should be as robust as possible, and (3) it should follow some "axioms of rational thinking". From this point of view, the question on the existence of the hidden world beyond quantum is claimed to be irrelevant: whatever this world is, we are forced, by the properties of our mind, to describe the reality via something similar to quantum theory.

At a phenomenological level, the "neural network" [7] and the "logical inference" [14–16] approaches are not contradictory, and may in some sense be dual to each other. Indeed, if our mind could be modeled as a neural network, then it is not too unreasonable to expect that the principles of work of the neural network [13] might be used to derive the postulated axioms of the logical inference [17]. This possibility is supported by the fact that in both approaches one is able to derive Schrödinger equation [7, 14] by combination of some entropic variational principle aimed to provide the most robust and efficient description of the external world with some form of the Hamilton–Jacobi equations (see also Refs. [18, 19] for other derivations based on entropic principles.) There is, however, an important flaw in these constructions, explicitly mentioned in Ref. [14]. What is actually derived, in both cases, is not the Schrödinger, but the Madelung [20] hydrodynamic equation which is known to be different from the Schrödinger equation [21]. The key difference is in the global topology. In the Madelung form of the Schrödinger equation, we introduce the "fluid density" which is related, in quantum language, to the modulus of the wave function, and the "fluid velocity" which is related to a gradient of the phase of the wave function. However, in the Schrödinger equation the phase is defined modulo 2π (we glue the plane to a cylinder) and in the Madelung equation this condition is lost. Without it, the Madelung hydrodynamics describes only a very special kind of hydrodynamic flows, that is, curlless (without vortices) whereas the crucial point of quantum physics, the quantization (in particular, discreteness of atomic energy levels) is associated with discreteness of circulation, like in superconductors and superfluids [22, 23].[1] In some inexplicit way, this change of topology simplifies the description allowing to pass from the *nonlinear* Madelung equations to the *linear* Schrödinger equation adding extremely powerful machinery of vectors and operators in a Hilbert space. Phenomenologically, it can be justified by introducing a new principle of "separation of conditions" [6], logically independent from the logical inference approach. However, due to extreme importance of this point one needs to have a more detailed understanding of its origin. In this paper we will show that the neural network approach gives a natural way to understand this transition and thus allowing to understand deeper the origin and meaning of one of the most fundamental physical constants, namely, the Planck constant.

[1] MIK thanks Grigory Volovik for emphasizing this connection at our old discussions of the logical inference approach.

The other point of a great interest is a very controversial, but tempting, idea of "emergent quantumness", that is, some quantum-like behavior of systems which are difficult to believe to be quantum per se [24]. Some authors use quantumness just as a metaphor to describe cultural phenomena [25] or genotype-phenotype duality in biological evolution [26], while others suggest the relevance of the true quantum phenomena in human brains [27, 28]. On a more practical level, this line of thinking may be related to a much more pragmatic and solid concept of "quantum annealing" [29–34]. For example, if some optimization problem can be stated as the problem of finding a ground state of a complicated *classical* system, then it is often convenient to add to the system some "quantumness" because the quantum tunneling would prevent a self-trapping of the optimization process in one of metastable states. We will show here that the neural network approach gives a natural explanation of the emergent quantumness and provides a solid formal background for speculations on quantum behavior of non-quantum (e.g. macroscopic and even biological) systems.

The paper is organized as follows. In Sect. 2 we apply the principle of stationary entropy production to derive a functional which governs the emergent dynamics of neural networks. In Sect. 3 we argue that the dynamics can be described by the Schrodinger equation if and only if the free energy of the hidden variables is a multivalued function. In Sect. 4 we construct a grand canonical ensemble of neural networks and show that the corresponding free energy is multivalued. In Sect. 5 we discuss in more detail some fundamental issues such as the difference between pure and mixed states, role of measurements and relations to path-integral formulation of quantum mechanics [35–37]. In Sect. 6 we discuss implications of the main results for machine learning, physics and biology.

2 Stationary Entropy Production

Consider a learning system described by a coupled dynamics of *trainable* variables, $\mathbf{q}$, and non-trainable or *hidden* variables, $\mathbf{x}$. In "epistomological" kind of approaches [6, 14–16] one can identify the trainable variables with characteristics of a human mind whereas the hidden variables represent an external world, but this identification is not needed for our formal consideration which we will try to keep as general as possible. In context of artificial neural networks the trainable variables determine the weight matrix and bias vector, and the hidden variables represent the state vector of neurons [13]. It is assumed that on the shortest time-scales the dynamics of the trainable variables undergoes diffusion

$$\begin{aligned}\frac{\partial p(t,\mathbf{q})}{\partial t} &= \sum_k \frac{\partial}{\partial q_k}\left(D\frac{\partial p(t,\mathbf{q})}{\partial q_k} - \frac{dq_k}{dt}p(t,\mathbf{q})\right)\\ &= \sum_k \frac{\partial}{\partial q_k}\left(D\frac{\partial p(t,\mathbf{q})}{\partial q_k} - \gamma\frac{\partial F(t,\mathbf{q})}{\partial q_k}p(t,\mathbf{q})\right)\end{aligned} \tag{1}$$

and the dynamics of hidden variables is only described through its free energy

$$\begin{aligned}\frac{d}{dt}F(t,\mathbf{q}) =& \frac{\partial F(t,\mathbf{q})}{\partial t} + \sum_k \frac{dq_k}{dt}\frac{\partial F(t,\mathbf{q})}{\partial q_k}\\ =& \frac{\partial F(t,\mathbf{q})}{\partial t} + \gamma \sum_k \left(\frac{\partial F(t,\mathbf{q})}{\partial q_k}\right)^2\end{aligned} \tag{2}$$

where the trainable variables experience a classical drift in the direction of the gradient of the free energy,

$$\frac{dq_k}{dt} = \gamma \frac{\partial F(t,\mathbf{q})}{\partial q_k}. \tag{3}$$

(See Refs. [7, 13] for details.) We shall assume that the drift γ and diffusion D coefficients are constants (independent of $\mathbf{q}$), but their numerical values depend on a learning algorithm. For example, if a neural network is trained using the stochastic gradient descent, then γ and D depend on the learning rate and the mini-batch size [38]. Note that the system under consideration is supposed to be, initially, purely classical and subjected by stochastic and, moreover, dissipative dynamics described by equations (1)–(3). One can think in particular on conventional neural network algorithms realized at normal classical computers, nothing specifically quantum is assumed yet.

To describe the dynamics on longer time-scales we employ the principle of stationary entropy production:

Principle of Stationary Entropy Production The path taken by a system is the one for which the entropy production is stationary.

The principle was first introduced in Ref. [19] as a generalization of both, the maximum entropy principle [39, 40] and the minimum entropy production principle [41, 42]. In context of the neural networks the entropy production must be maximized in an optimal neural architecture [7, 13]. The rationale behind it is simple: if less information is used for optimizing the network for past data, then more entropy is available for optimizing the network for future data and, therefore, a larger space of solutions can be explored. With this respect the principle can be thought of as a formalization of the Occam's razor principle.[2]

The Shannon entropy of the trainable variables is given by

$$S_q(t) = -\int d^K q \;\; p(t,\mathbf{q}) \log\left(p(t,\mathbf{q})\right) \tag{4}$$

and the total entropy production can be calculated from (1),

[2] We thank Nikolay Mikhailovsky for pointing out the connection to Occam's razor principle.

$$\begin{aligned}\frac{dS_q(t)}{dt} =& - \int d^K q\, p \frac{\partial \log(p)}{\partial t} - \int d^K q\ \log(p) \frac{\partial p}{\partial t} \\ =& - \frac{d}{dt} \int d^K q\, p - \int d^K q\ \log(p) \frac{\partial p}{\partial t} \\ =& - \int d^K q\ \log(p) \sum_k \frac{\partial}{\partial q_k} \left(D \frac{\partial p}{\partial q_k} - \gamma \frac{\partial F}{\partial q_k} p \right) \\ =& D \int d^K q\ \sum_k \frac{1}{p} \left(\frac{\partial p}{\partial q_k} \right)^2 - \gamma \int d^K q\ \sum_k \frac{\partial p}{\partial q_k} \frac{\partial F}{\partial q_k}.\end{aligned} \tag{5}$$

The two terms represent the entropy production due to stochastic and learning dynamics of the trainable variables. At a learning equilibrium (i.e. $\nabla F \sim 0$) the entropy production of the trainable variables due to learning is subdominant and the total entropy production of trainable variables (5) can be approximated as

$$\frac{dS_q(t)}{dt} \approx \int d^K q\ \sqrt{p} \left(-4D \sum_k \frac{\partial^2}{\partial q_k^2} \right) \sqrt{p}. \tag{6}$$

Note that this term is nothing but Fisher information which determines metrics in the information space [43] and plays an important role in the logical inference approach to quantum mechanics [14–16] and in the information theory approach to emergent gravity [44]. Further away from the equilibrium the entropy production due to learning cannot be ignored and the dynamics is better described by a classical Hamiltonian mechanics with the free energy, F, identified with the Hamilton's principle function (see [7] for details on both classical and quantum limits).

The problem of optimization of the entropy production (6) subject to a constraint (2) can be solved using the method of Lagrange multipliers by defining a functional [7]

$$\begin{aligned}\mathcal{S}[p, F, \lambda] =& \int_0^T dt \frac{dS_q}{dt} + \lambda \int_0^T dt d^K q\, p \left(\frac{\partial F}{\partial t} + \gamma \sum_k \left(\frac{\partial F}{\partial q_k} \right)^2 + \frac{V}{\epsilon} \right), \\ =& \int_0^T dt\, d^K q\, \sqrt{p} \left(-4D \sum_k \frac{\partial^2}{\partial q_k^2} + \lambda \frac{\partial F}{\partial t} + \lambda\gamma \sum_k \left(\frac{\partial F}{\partial q_k} \right)^2 + \lambda \frac{V}{\epsilon} \right) \sqrt{p},\end{aligned} \tag{7}$$

where the total time-averaged free energy production pre unit time step ϵ is

$$V(\mathbf{q}) \equiv - \left\langle \epsilon \frac{d}{dt} F(t, \mathbf{q}) \right\rangle_t. \tag{8}$$

Again, the time step ϵ here is just a parameter of our neural network related to the rate of computation at a given realization of the neural network algorithm and is not related to any fundamental physical constants such as Planck time, etc. We postpone the discussion of the fundamental physics till the last section, meanwhile it is better to keep in mind just some more or less standard computations on more or less standard computers.

Note that the logical inference approach [14] leads to a technically very similar formulation, the principle of robustness of description shown to be equivalent to the minimum of Fisher information, but the second requirement, correctness of the Hamilton-Jacobi equations at the average, was postulated as a property of our world, in spirit of Bohr's correspondence principle. The neural network approach [7] provides us an explicit model with the phenomenologically desired properties. Of course, strictly speaking, one cannot exclude existence of other models which lead to more or less the same phenomenology, but this model seems to be, in some sense, the most natural.

It is convenient to rewrite the functional (7) as

$$\mathcal{S}[p,F,\hbar] = \frac{\lambda}{\epsilon}\int_0^T dt\, d^K q\, \sqrt{p}\left(-\frac{\hbar^2}{2m}\sum_k \frac{\partial^2}{\partial q_k^2} + \frac{\partial(\epsilon F)}{\partial t} + \frac{1}{2m}\sum_k \left(\frac{\partial(\epsilon F)}{\partial q_k}\right)^2 + V\right)\sqrt{p} \tag{9}$$

where

$$m \equiv \frac{\epsilon}{2\gamma}, \tag{10}$$

and

$$\hbar \equiv \epsilon\sqrt{\frac{4D}{\gamma\lambda}}. \tag{11}$$

The constant (11) will play the role of the Planck constant in the further consideration but currently it is just a combination of some parameters characterizing our neural network, and it is not assumed to be either microscopic or fundamental. Generally speaking, different neural networks can have different "Planck constants", and one can, in principle, even assume a variable "constant" during the computation.

The main difference between (7) and (9) is that instead of solving the equations for p, F and λ, we are now solving them for p, F and $\hbar$. The optimal solutions are obtained by setting all possible variations of (9) to zero

$$\frac{\partial}{\partial\hbar}\mathcal{S}[p,F,\hbar] = -\int_0^T dt\, d^K q \frac{\hbar}{m}\sqrt{p}\sum_k \frac{\partial^2}{\partial q_k^2}\sqrt{p} = 0 \tag{12}$$

$$\frac{\delta}{\delta F}\mathcal{S}[p,F,\hbar] = -\frac{\partial}{\partial t}p - \frac{1}{m}\sum_k \frac{\partial}{\partial q_k}\left(\frac{\partial(\epsilon F)}{\partial q_k}p\right) = 0 \tag{13}$$

$$\frac{\delta}{\delta p}\mathcal{S}[p,F,\hbar] = -\frac{\hbar^2}{2m}\frac{1}{\sqrt{p}}\sum_k \frac{\partial^2\sqrt{p}}{\partial q_k^2} + \frac{\partial(\epsilon F)}{\partial t} + \frac{1}{2m}\sum_k\left(\frac{\partial(\epsilon F)}{\partial q_k}\right)^2 + V = 0. \tag{14}$$

3 Schrödinger Dynamics

In the previous section we derived the functional (9) which describes the total entropy production of the trainable variables $\mathbf{q}$ subject to a constraint imposed on the dynamics of free energy of the hidden variables $\mathbf{x}$. The stationary solutions for probability density, $p(t, \mathbf{q})$, free energy, $F(t, \mathbf{q})$ and "Planck constant", $\hbar$, are given by equation (12), which represents conservation of entropy

$$\left\langle \frac{d}{dt} S_q \right\rangle_t = 0, \tag{15}$$

and by equations (13) and (14), which are the Madelung hydrodynamic equations [20]

$$\frac{\partial}{\partial t} p = - \sum_k \frac{\partial}{\partial q_k} \left(u_k p \right) \tag{16}$$

$$\frac{\partial}{\partial t} u_j = - \sum_k u_k \frac{\partial}{\partial q_k} u_j - \frac{1}{m} \frac{\partial}{\partial q_j} \left(V - \frac{\hbar^2}{2m} \sum_k \frac{\partial^2 \sqrt{p}}{\partial q_k^2} \right) \tag{17}$$

with velocity of the fluid

$$u_k \equiv \frac{1}{m} \frac{\partial}{\partial q_k} (\epsilon F). \tag{18}$$

It is well known that the Madelung equations can be derived from the Schrödinger equation

$$-i\hbar \frac{\partial}{\partial t} \Psi = \left(\frac{\hbar^2}{2m} \sum_k \frac{\partial^2}{\partial q_k^2} - V \right) \Psi \tag{19}$$

where the wave function is defined as

$$\Psi \equiv \sqrt{p} \exp\left(\frac{iF\epsilon}{\hbar} \right). \tag{20}$$

This implies that the solutions of (16) and (17) are also the solutions of the Schrödinger equation, (19), and so it is expected that the time-averaged change in entropy is zero (12), but the opposite is not true. The core of the problem is that the quantum phase in (20) must be a multivalued function, but the free energy F of hidden variables may or may not be single-valued. Therefore, in order to establish an equivalence between quantum mechanics and neural networks, we must consider statistical ensembles for which the free energy of hidden variables would be multivalued,

$$F \cong F + \mu n \qquad \forall n \in \mathbb{Z}. \tag{21}$$

This can be accomplished by constructing a statistical ensemble of neural networks for which the discrete shift, μn, is not observable. In the following section we shall consider one such ensemble, the grand canonical ensemble with chemical potential, μ, for which the exact number of neurons is unobservable and, therefore, the free energy is multivalued (21). Note that the proportionality of thermodynamic potentials to the number of particles in a thermodynamic limit (a very large number of degrees of freedom), which is crucially important for our whole construction, can be proven mathematically rigorously for a broad class of continuous and lattice models of statistical mechanics [45].

If we assume for a moment that the multivaluedness condition (21) is satisfied, then (9) can be rewritten as

$$\mathcal{S}[p, F, \hbar] = \frac{\lambda}{\epsilon}\int_0^T dt\, d^K q\, p(t, \mathbf{q})\left(\frac{\hbar^2}{2m}\sum_k \frac{\partial \varphi^*}{\partial q_k}\frac{\partial \varphi}{\partial q_k} - i\hbar\frac{\partial \varphi}{\partial t} + V\right) \tag{22}$$

where

$$\varphi \equiv \log\sqrt{p} + i\epsilon\frac{F + \mu n}{\hbar}. \tag{23}$$

If $n \in \mathbb{Z}$ is indeed unobservable, then $\mathcal{S}[p, F, \hbar]$ should not depend on n which is guaranteed only if

$$\mu = \frac{2\pi\hbar}{\epsilon}m, \tag{24}$$

for some $m \in \mathbb{Z}$. If that would not be true, then by studying changes in the action (22) we would be able to extract information about n in a conflict with our assertion that n is unobservable. In other words, $\mathcal{S}[p, F, \hbar]$ must be invariant under transformation $F \to F + \mu n$ for all $n \in \mathbb{Z}$ which is assured if we impose the condition (24). To prove that $m = \pm 1$ we must now look at other discrete transformations of F. If $m \neq \pm 1$, then there are transformations

$$F \to F + \mu\frac{n}{m} \qquad\qquad \forall n \in \mathbb{Z} \tag{25}$$

which leave $\mathcal{S}[p, F, \hbar]$ invariant, but $n/m \notin \mathbb{Z}$, e.g. $n = 1$, $m = 2$ and $n/m = 1/2$. If this is the case, then the parameter μ in (21) was not chosen correctly to describe the unobservability in F. Instead the parameter must be rescaled $\mu \to \pm\mu/m$ and then (25) reduces to (21) and equation (24) becomes

$$\hbar = \pm\frac{\mu\epsilon}{2\pi}. \tag{26}$$

By imposing the condition (26) on $\hbar$ in (22) we arrive at the Schrödinger action

$$\mathcal{S}[\Psi] = \frac{\lambda}{\epsilon}\int_0^T dt\, d^K q\left(\frac{\hbar^2}{2m}\sum_k \frac{\partial \Psi^*}{\partial q_k}\frac{\partial \Psi}{\partial q_k} - i\hbar\Psi^*\frac{\partial \Psi}{\partial t} + V\Psi^*\Psi\right) \tag{27}$$

where the wave function Ψ is given by (20). Therefore, if the multivaluedness condition (21) is satisfied, then the Planck constant must be given by (26) and then the Schrödinger action (27) provides a correct statistical description of the learning dynamics of neural networks.

4 Grand Canonical Ensemble

Consider a neural network at a learning equilibrium described by a temperature parameter, T, and, in addition, with a possible access to a reservoir of auxiliary neurons described by a chemical potential, μ. What this means is that the learning algorithm is such that the system can either increase (i.e. neurogenesis) or decrease (i.e. neurodegeneration) the total number of active neurons, N. It is not immediately clear that such an algorithm would be present in an optimal learning system, but this is something that we will discuss shortly. Meanwhile, the very fact that the exact number of active neurons (or hidden variables) N is unknown suggests that the system should be modeled with a grand canonical ensemble. The corresponding thermodynamic potential is the grand (or Landau) potential

$$\Omega(\mathbf{q}, T, \mu) = F - \mu N \tag{28}$$

where

$$\mu = \frac{\partial F}{\partial N}. \tag{29}$$

For a system kept at an equilibrium with constant temperature, T, and chemical potential, μ, the fundamental thermodynamic relations is

$$d\Omega = dF - \mu dN = 0. \tag{30}$$

The relation (30) can be regarded as a generalization of the first law of learning that was introduced in [7, 13] in context of a canonical ensemble of neural networks.

According to the first law (30) the free energy, F, can undergo both continuous transformations due to dynamics of trainable variables, $\mathbf{q}$, and discontinuous transformations due to dynamics of the number of neurons, N. This implies that the free energy must be "quantized" in the following sense

$$F(\mathbf{q}, T, N) = \Omega(\mathbf{q}, T, \mu) + \mu N. \tag{31}$$

Since the exact number of active neurons, N, in the grand canonical ensemble is unknown, the free energy F is only known up to an additive constant μn where $n \in \mathbb{Z}$. If we identify μ in (21) with chemical potential of the grand canonical ensemble, then the unobservability of the number of active neurons implies the multivaluedness condition (21). Strictly speaking, the condition is only satisfied if the integer n in (21) remains smaller than the uncertainty in the number of neurons

$$\Delta N = \sqrt{\langle N^2 \rangle - \langle N \rangle^2} = \sqrt{T \frac{\partial^2 \Omega}{\partial \mu^2}}. \tag{32}$$

In the limit $n \ll \Delta N$ the multivaluedness condition is satisfied and the Schrödinger equation provides a good statistical description of the learning dynamics, but in the opposite limit $n \gtrsim \Delta N$ the Schrödinger description is expected to break down. However, one can argue that in an optimal neural network the parameter ΔN must be maximized which would make the behavior of the system as quantum as possible.

Indeed, for every "macroscopic" solution for trainable variables, $\mathbf{q}$, there is a statistical ensemble of microscopic solutions for hidden variables, $\mathbf{x}$. With this respect the grand canonical ensemble, $\Delta N \neq 0$, provides a clear advantage over canonical ensemble, $\Delta N = 0$, as it allows for a much larger number of microscopic solutions corresponding to different values of the number of active neurons, N. More precisely, if the system has access to $\sim 2\Delta N$ auxiliary neurons, then each of these neurons can either be active or not active, and then the additional entropy is given by

$$\Delta S \sim 2\Delta N = 2\sqrt{T \frac{\partial^2 \Omega}{\partial \mu^2}}. \tag{33}$$

This entropy describes the additional "macroscopic" solutions for trainable variables, $\mathbf{q}$, which can have discontinuous jumps in the free energy (21) due to uncertainty in the total number of neurons (32). Therefore, an optimal neural network must be described by a grand canonical ensemble with the largest possible ΔN for which the multivaluedness condition (21) would be maximally satisfied. This establishes an equivalence between quantum mechanics and an optimal learning system described by a grand canonical ensemble of neural networks.

Whether the free energy F (and the loss function $U = F - TS_x$) is "quantized" (31) and whether it can change discontinuously depends on the learning system. In Ref. [13] it was shown numerically that for the bulk loss function the discontinuous jumps are suppressed, but for the boundary loss function the discontinuous jumps are expected even when the total number of neurons is kept constant. This result agrees very well with our analysis of the grand canonical ensembles and with the first law of learning (30). Indeed, from the point of view of the bulk neurons the total number of neurons is constant, the relevant statistical ensemble is canonical and the discontinuous jumps do not occur,

$$dU = TdS_x. \tag{34}$$

On the other hand, from the point of view of only boundary neurons, some of the bulk neurons can act as a reservoir of auxiliary neurons, the relevant ensemble is grand canonical and the discontinuous jumps are expected,

$$dU = TdS_x + \mu dN. \tag{35}$$

Therefore, in addition to theoretical considerations we also have preliminary numerical results suggesting that the discontinuous jumps in the boundary loss function is a result of the learning dynamics described by a grand canonical ensemble.

Using the "quantization" of the free energy (31) and the optimal value for the Planck constant (26), the wave function (20) can written as

$$\Psi(t, \mathbf{q}) = \sqrt{p(t, \mathbf{q})} \exp\left(\frac{i\Omega(t, \mathbf{q})\epsilon}{\hbar}\right). \tag{36}$$

As an example, consider a grand potential which can be expressed as a sum of a fixed time-independent term and a time-dependent term, i.e.

$$\Omega(t, \mathbf{q}) = \Omega_0(\mathbf{q}) + \Omega_1(t, \mathbf{q}). \tag{37}$$

In such limit the Schrödinger action (19) can be rewritten as

$$\begin{aligned} \mathcal{S}[\tilde{\Psi}] =& \frac{\lambda}{\epsilon} \int_0^T dt\, d^K q \\ & \left(-\frac{\hbar^2}{2m} \tilde{\Psi}^* \sum_k \left(\frac{\partial}{\partial q_k} + i\frac{eA_k}{\hbar} \right) \left(\frac{\partial}{\partial q_k} + i\frac{eA_k}{\hbar} \right) \tilde{\Psi} - i\hbar \tilde{\Psi}^* \frac{\partial \tilde{\Psi}}{\partial t} + V \tilde{\Psi}^* \tilde{\Psi} \right) \end{aligned}$$

where the new wave function is now defined as

$$\tilde{\Psi}(t, \mathbf{q}) \equiv \sqrt{p(t, \mathbf{q})} \exp\left(\frac{i\Omega_1(t, \mathbf{q})\epsilon}{\hbar}\right) \tag{38}$$

and

$$A_k(\mathbf{q}) \equiv \frac{\epsilon}{e} \frac{\partial \Omega_0(\mathbf{q})}{\partial q_k}. \tag{39}$$

Then upon variation of the action with respect to the new wave function we get

$$-i\hbar \frac{\partial}{\partial t} \tilde{\Psi} = \left[\frac{\hbar^2}{2m} \left(\frac{\partial}{\partial q_k} + i\frac{eA_k}{\hbar} \right) \left(\frac{\partial}{\partial q_k} + i\frac{eA_k}{\hbar} \right) - V \right] \tilde{\Psi}. \tag{40}$$

Note that the result is only valid when the time-independent term, $\Omega_0(\mathbf{q})$, is fixed (i.e. is not varied), which is exactly the limit in which the Schrödinger equation (40) provides a good description of a quantum particle with "charge" e in an external field described by the "vector potential ", $\mathbf{A}$.

5 Quantum Superpositions

In the previous sections we modeled the learning dynamics of neural networks using either Madelung (16), (17) or Schrödinger (19) equations. In both cases the dynamics was described in a position basis $\mathbf{q}$ which is the preferred basis in our entire construction. The main difference between Madelung and Schrödinger equations is that the Schrödinger dynamics is linear and thus can be expressed in any orthonormal basis without loosing the generality. What is less clear is how to put the system in an arbitrary initial state or how to measure the system with respect to arbitrary measurement operators. In particular, it is not clear how to start the dynamics in a superposition of position eigenstates or how to perform measurements using non-diagonal (in the position basis) measurement operators.

The concept of measurement plays a central role in quantum physics, as especially emphasized in Bohr's complementarity principle [46, 47]. What we deal with is never a "quantum system by itself" but a result of its interaction with some measurement devices, and we can choose different descriptions choosing different sets of devices. For example, we can measure either coordinate of a particle by a local detector which clicks when the particle interacts with it or momentum of the particle via its wave properties, using to this aim diffraction lattices, etc. For the case of neural networks the coordinate representation is special. It is very straightforward to measure the set of $\mathbf{q}$ at a given time instant, this information is directly available at the computer. At the derivation of Schrödinger equation in logical inference approach [14] the space is supposed to be filled by coordinate detectors as well whereas measurement of momenta is not so easily realizable even as a gedanken experiment. Note however that the quantum mechanics allows a purely space-time formulation which was realized in Feynman's path integral approach [35]. Mathematically speaking, the solution of the Cauchy problem for the Schrödinger equation can be presented as a path integral, via splitting of the evolutionary operator into many elementary factors with the further use of Trotter decomposition formula [36]. All interference phenomena, energy quantization and other specifically quantum phenomena follow immediately from this representation [35–37]. Importantly, the trajectories giving the main contribution to the path integral are continuous but not continuously differentiable [48] which means impossibility of simultaneous measurements of coordinates and velocities. Coming back to the operator language, one can discuss the measurements of noncommutative coordinate operators at different time instants rather than the measurements of noncommutative coordinate and velocity operators at the same time instant. Therefore furthe we will discuss only measurements of the coordinates $\mathbf{q}$.

In artificial neural networks numerical values of the trainable variables $\mathbf{q}$ are only known up to numerical precision and so after measurement in a position basis the system can only be in one of a finite number of states, i.e. $\mathbf{q} \in \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_M\}$. Using the bra-ket notations the position eigenstates are given by

$$|\mathbf{q}\rangle \in \{|\mathbf{q}_1\rangle, |\mathbf{q}_2\rangle, ..., |\mathbf{q}_M\rangle\}, \tag{41}$$

and the most general initial state can be expressed as a linear superposition

$$|\Psi(0)\rangle = \sum_{i=1}^{M} \Psi(0, \mathbf{q}_i)|\mathbf{q}_i\rangle. \tag{42}$$

It is certainly possible to use a random number generator to set the initial state to be in a state $|\mathbf{q}_i\rangle$ with probability $p_i = |\Psi(0, \mathbf{q}_i)|^2$, but then the system would not be in a pure state (42). Such states are known as mixed states that are usually described by a density matrix

$$\hat{\rho} = \sum_{i=1}^{M} p_i |\mathbf{q}_i\rangle\langle\mathbf{q}_i|. \tag{43}$$

It seems that the only way for a system to remain in a pure state is through unitary evolution which can in general be time-dependent. Then to prepare a superposition state (42) at time $t = 0$ we can pre-evolve it starting from a position eigenstate $|\mathbf{q}_j\rangle$ at time $t = -t_-$, i.e.

$$|\Psi(0)\rangle = e^{-it_-\hat{H}_-/\hbar}|\mathbf{q}_j\rangle \tag{44}$$

where $\hat{H}_-$ is the pre-evolution Hamiltonian operator. Note that $\hat{H}_-$ emerges from a microscopic loss function and a training dataset which need not be the same as for $\hat{H}$ which governs the main part of the evolution

$$|\Psi(T)\rangle = e^{-iT\hat{H}/\hbar}|\Psi(0)\rangle. \tag{45}$$

It is important to emphasize that although the loss functions, training datasets and emergent Hamiltonians for pre-evolution and main evolution can differ, the neural architectures, described by trainable $\mathbf{q}$ and non-trainable $\mathbf{x}$ variables, must remain the same. Of course, this does not guarantee that we can use the pre-evolution to prepare all possible superposition states, but by modifying t_- (or pre-evolution time interval) and $\hat{H}_-$ (or pre-evolution loss function and training dataset) a larger variety of pure initial states can be realized. Also note that on top of realizing superposition states one can still use a random number generator to create a mixed state by starting the pre-evolution with (43) and then the density matrix at time $t = 0$ would be

$$\hat{\rho}(0) = e^{-it_-\hat{H}_-/\hbar}\left(\sum_{i=1}^{M} p_i |\mathbf{q}_i\rangle\langle\mathbf{q}_i|\right)e^{it_-\hat{H}_-/\hbar} \tag{46}$$

which need not be diagonal.

What about measurement operators? Can we use a similar method to (effectively) measure the system using non-diagonal measurement operators,

$$\hat{O}_m \equiv \sum_{i,j=1}^{M} O_{ij}^{(m)}|\mathbf{q}_i\rangle\langle\mathbf{q}_j|? \tag{47}$$

If the quantum description is correct then the probability of observing a given measurement operator must be given by

$$p(m) = \langle\Psi(T)|\hat{O}_m^{\dagger}\hat{O}_m|\Psi(T)\rangle \tag{48}$$

where

$$\sum_m \hat{O}_m^\dagger \hat{O}_m = \hat{I}. \tag{49}$$

For diagonal measurement operators (denoted by $\hat{D}_m$'s) the probabilities are given by

$$p(m) = \sum_{i=1}^{M} \left| D_{ii}^{(m)} \right|^2 |\Psi(T, \mathbf{q}_i)|^2 \tag{50}$$

and by measuring positions $\mathbf{q}_i$'s, probabilities $|\Psi(T, \mathbf{q}_i)|^2$'s can be calculated and $p(m)$'s can be verified against theoretical predictions. However, for non-diagonal measurement operators

$$p(m) = \sum_{i=1}^{M} \left| \sum_{j=1}^{M} O_{ij}^{(m)} \Psi(T, \mathbf{q}_j) \right|^2, \tag{51}$$

and the knowledge of probabilities $|\Psi(T, \mathbf{q}_i)|^2$ is not sufficient for calculating $p(m)$'s or for performing measurements of the corresponding operators. On the other hand, what one can do is to post-evolve the state $|\Psi(T)\rangle$ to $|\Psi(T + t_+)\rangle$ using some post-evolution Hamiltonian $\hat{H}_+$, i.e.

$$|\Psi(T + t_+)\rangle = e^{-it_+\hat{H}_+/\hbar} |\Psi(T)\rangle, \tag{52}$$

and then measure it using some set of diagonal operators $\hat{D}_m$'s. Then the probabilities of measuring $\hat{D}_m$'s would be given by (50), which can be calculated and verified against theoretical predictions, but the same probabilities can also be expressed as

$$p(m) = \langle \Psi(T + t_+) | \hat{D}_m^\dagger \hat{D}_m | \Psi(T + t_+) \rangle \tag{53}$$

$$= \langle \Psi(T) | e^{it_+\hat{H}_+/\hbar} \hat{D}_m^\dagger e^{-it_+\hat{H}_+/\hbar} e^{it_+\hat{H}_+/\hbar} \hat{D}_m e^{-it_+\hat{H}_+/\hbar} | \Psi(T) \rangle \tag{54}$$

$$= \langle \Psi(T) | \hat{O}_m^\dagger \hat{O}_m | \Psi(T) \rangle, \tag{55}$$

i.e. as probabilities of effectively measuring non-diagonal operators

$$\hat{O}_m = e^{it_+\hat{H}_+/\hbar} \hat{D}_m e^{-it_+\hat{H}_+/\hbar}. \tag{56}$$

Of course, there is no guarantee that we would be able to effectively measure all possible sets of the measurement operators since the procedure is still limited by possible choices of the post-evolutionary Hamiltonian $\hat{H}_+$ (or loss function and training dataset) and time interval t_+.

6 Discussion

In this paper we analyzed the emergent macroscopic dynamics of neural networks, but deliberately omitted specifications of the microscopic dynamics. The only important microscopic ingredient was that there are two types of degrees of freedom which correspond respectively to trainable and hidden variables. In the emergent picture the trainable variables were identified with the variables of the wave function, but the hidden variables were only described at the level of statistical ensembles. In fact, it was not even important whether the hidden variables are actually non-trainable or only appear to evolve as non-trainable variables, but what was important is that their statistics can be described with a grand canonical ensemble. In this respect one can argue that the emergent quantumness is a generic macroscopic prediction of any learning system with a coupled dynamics of these two types of variables, i.e. trainable and hidden.

It is also worth emphasizing that the trainable variables were assumed to be continuous and the configuration space was assumed to be flat. More generally some of the variables (hidden or trainable) can be either discrete or the configuration space can be curved. In such cases the derivatives would be replaced with either finite differences or with covariant derivatives, but this would not alter the main conclusion of the paper. In fact, we expect that in more realistic models of neural network (which could give rise to emergent quantum field theories and gravity) the trainable variables must include features of both discrete variables and curved spaces. On the other hand, the *discreteness* of the neural network (i.e. the number of neurons is a discrete integer) is a crucial point of the whole construction and for the main result, i.e. the emergence of quantumness. This result has some immediate and important implications for machine learning, physics and biology that we shall discuss next.

6.1 Machine Learning

Machine learning is perhaps the simplest example of a learning system which has some apparent advantages over physical and biological systems. Artificial neural networks are well defined mathematically and, as such, provides an excellent experimental platform for testing the new ideas numerically. There are at least three (not unrelated) ideas which follow directly from our results. According to our analysis an optimal learning system should be based on an algorithm which allows for the number of hidden variables to vary. One way to allow the number of hidden variables to change is to develop an algorithm designed specifically for addition and removal of the auxiliary neurons. In this respect it might be useful to develop a *neurodegeneration* method for removing neurons that causes the smallest increase in the loss function and a *neurogenesis* method for adding auxiliary neurons that causes the largest decrease in the loss function.

Another possibility is to develop an algorithm in which the change in the number of hidden variables would occur dynamically. We expect that this is what might actually be happening behind the scenes in deep learning. If correct, this suggests a simple and intuitive explanation of why the deep neural networks (with many hidden layers) are very effective in learning. The reason might be that the bulk neurons on the hidden layers can act as a reservoir of auxiliary neurons and the corresponding ensemble becomes grand canonical. In such a limit the correct effective description of the learning dynamics of the boundary system is quantum with all of the computational advantages which come with it. Perhaps a lot more importantly, the emergent quantumness implies that one might be able to design an artificial neural network which can mimic the behavior of a quantum computer. Of course, such an artificial quantum computer would not be quantum per se, but one could still make it maximally quantum by designing an algorithm which maximizes the uncertainty in the number of active neurons.

6.2 Physics

We can now try to tackle the somewhat more difficult problem of modeling physical systems using neural networks. Indeed, if quantum mechanics provides a good description of the physical world and a good description of the neural networks, then why cannot the physical word be a neural network? This was precisely the questions that was asked in Ref. [7] where not only quantum mechanics, but also gravity and observers were described as emergent phenomena. (See Refs. [49–52] for other approaches to emergent gravity). In this paper we concentrated mostly on the emergent quantum behavior of the neural networks, but our results have some interesting implications for both gravitational and biological systems. Our main quantum result is that the correct statistical ensemble of hidden variables is the grand canonical ensemble, where the chemical potential is what determines the valued of the physical "Planck constant". In the quantum limit the learning system satisfies the following two conditions:

(1) The system is at a learning equilibrium (i.e. small gradient of the free energy) and
(2) The quantum phase is multivalued (i.e. large uncertainty in the number of neurons). However, since the emergent quantum behavior is only approximate, it is also important to identify the systems in which significant deviations from the quantum behavior are expected. For example, for a canonical ensemble of hidden variables the conditions (1) can be satisfied, but the condition, (2) is badly violated and then the system is better described with the Madelung equations. In an opposite limit, when the condition (2) is satisfied, but the condition (1) is violated, the system is better described with the Hamilton-Jacobi equations (see Ref. [7] for details).

Perhaps a more surprising aspect of the learning dynamics is that one might be able to derive a couple of dual descriptions of the very same system. In a "boundary" description one keeps track of only a small number of trainable variables which have already thermalized, i.e. satisfying the condition (1), and the rest of the variables are treated as hidden variables whose total number is unknown, i.e. satisfying the condition (2). In a “bulk” description one keeps track of most of the trainable variables which can be very far from the true equilibrium, i.e. violating the condition (1), and the total number of hidden variables is small and its fluctuations are also small, i.e. violating the condition (2). For the boundary system the correct emergent description is quantum, but for the bulk system the correct description is mostly classical and in some cases gravitational [7]. This resonates well with a holographic conjecture which states that a gravitational system in the bulk should have a quantum dual description on the boundary [53–55]. Indeed, if the microscopic neural network is being trained by processing the new training data through its boundary (as for example in the deep feedforward neural networks), then the boundary neurons should be the first to thermalize. Moreover, from the point of view of the boundary neurons the ensemble of hidden variables is in a grand canonical equilibrium, and then the emergent dynamics is quantum. On the other hand, the bulk neurons would be further away from the equilibrium and their emergent dynamics would be mostly classical and, perhaps, in some limits gravitational [7]. This offers an interesting new perspective on the holographic principle and on gravity as an emergent phenomenon. (See also Ref. [56] for an alternative approach to gravity and holography in context of neural networks.)

The emergent quantumness also provides a new twist to the long-standing problem of derivation of irreversible macroscopic laws from reversible microscopic ones. The basic equations determining the dynamics of neural network are already irreversible, due to the presence of diffusion terms with real diffusion coefficient (contrary to imaginary diffusion coefficient in time-reversal symmetric Schrödinger equation), and it is the reversibility of the microscopic laws turns out to be an emergent phenomenon, due to negative entropy production during learning. The negative entropy production is the direct consequence of the second law of learning, i.e. the total entropy can never increase during learning and is constant in the learning equilibrium [7, 13]. Then, after the equilibrium is reached, any positive entropy production due to diffusion of trainable variables must be balanced by the negative entropy production of either trainable or hidden variables. On the shortest time scales the interplay between positive and negative entropy productions can be expressed explicitly by including respectively the diffusion and drift terms in the Fokker-Planck equation, but on the longer time scales the entropy balance is expressed as an optimization problem which can be described by the Schrödinger equation. Therefore, for the long time scale dynamics there is an emergent time reversal symmetry, but on the short time scales the symmetry must be broken.

6.3 Biology

From a biological perspective, it is very tempting to speculate whether our mind can be modeled as a neural network which undergoes a learning evolution. In that case, its behavior near equilibrium should indeed resemble the quantum systems since it would be described by an effective Schrödinger equation. Of course, the effective Planck constant arising in such an equation has nothing in common with the real physical Planck constant. In this sense, our approach is essentially different from those of Refs. [27, 28] where the relevance of truly quantum processes in our bodies for our mind is suggested. The other point worth to be emphasized is that, in our approach, only a fully optimized neural network has this property of “quantumness”. Of course, it is not obvious at all whether our real brains originated from a real biological evolution on a specific planet are optimal enough to be “quantum”. Any suggestions of such kind would be unavoidably very speculative, but, in our opinion, deserve to be considered.

Speaking more generally on the biological evolution, one should mention Ref. [26] where the question of what kind of physics would be needed to describe evolution is discussed. The key point is a coexistence of two levels (genotype and phenotype) which is essentially entangled in a very unusual way from the point of view of conventional statistical mechanics. Physical carriers of genetic information are (very roughly speaking) macromolecules subjected to thermal fluctuations, electrostatic interactions, with electronic structure determined by quantum mechanics, etc. However, the functionality of the genetic information carriers cannot be adequately described in terms of their physical properties only. They are projected to a phenotype level, and at this level are subjected to selection with the laws which are also in agreement with general laws of physics and chemistry (as we believed), but act on a completely different level of macroobjects. This was compared with the role played by von Neumann measurements due to interaction with macroscopic measuring devices in quantum mechanics [26]. In this regard, the concept of emergent quantumness via neural network approach developed here might be useful for specification and formalization of this, still vague and preliminary, analogy. Anyway, there seems to be a clear association of genotype-phenotype duality in biology to the duality of hidden and trainable variables in neural networks. Another thought-provoking question is the utility of “quantum jumps” in the fitness values in the “quantum-like” evolutionary dynamics, but it obviously goes far beyond the scope of this particular work and deserves a separate consideration.

Acknowledgements VV was supported in part by the Foundational Questions Institute (FQXi). MIK acknowledges a support by NWO via Spinoza Prize.

References

1. Home, D.: Conceptual Foundations of Quantum Physics. Plenum Press, New York (1997)
2. Weinberg, S.: Lectures on Quantum Mechanics. Cambridge University Press, Cambridge (2003)
3. Khrennikov, A.Y.: Contextual Approach to Quantum Formalism. Springer, Berlin (2009)
4. Rauch, H., Werner, S.A.: Neutron Interferometry: Lessons in Experimental Quantum Mechanics, Wave-Particle Duality, and Entanglement. Oxford University Press, Oxford (2015)
5. Landsman, K.: Foundations of Quantum Theory: From Classical Concepts to Operator Algebras. Springer, Berlin (2017)
6. De Raedt, H., Katsnelson, M.I., Willsch, D., Michielsen, K.: Separation of conditions as a prerequisite for quantum theory. Ann. Phys. (NY) **403**, 112–135 (2019)
7. Vanchurin, V.: The world as a neural network. Entropy **22**, 1210 (2020)
8. Bell, J.: On the Einstein Podolsky Rosen Paradox. Physics **1**(3), 195–200 (1964)
9. Bohm, D.: A suggested interpretation of the quantum theory in terms of ‘hidden’ variables. I. Phys. Rev. **85**, 166–179 (1952)

10. Adler, S.: Quantum Theory as an Emergent Phenomenon. Cambridge University Press, Cambridge (2004)
11. 't Hooft, G.: Quantum mechanical behavior in a deterministic model. Found. Phys. Lett. **10**, 105 (1997)
12. 't Hooft, G.: The Mathematical basis for deterministic quantum mechanics. J. Phys. Conf. Ser. **67**, 012015 (2007)
13. Vanchurin, V.: Towards a theory of machine learning. Mach. Learn.: Sci. Technol. **2**, 035012 (2021)
14. De Raedt, H., Katsnelson, M.I., Michielsen, K.: Quantum theory as the most robust description of reproducible experiments. Ann. Phys. **347**, 45–73 (2014)
15. De Raedt, H., Katsnelson, M.I., Michielsen, K.: Quantum theory as plausible reasoning applied to data obtained by robust experiments. Phil. Trans. R. Soc. A **374**, 20150233 (2016)
16. De Raedt, H., Katsnelson, M.I., Michielsen, K.: Logical inference derivation of the quantum theoretical description of Stern-Gerlach and Einstein-Podolsky-Rosen-Bohm experiments. Ann. Phys. (NY) **396**, 96–118 (2018)
17. Cox, R.T.: The Algebra of Probable Inference. Johns Hopkins University Press, Baltimore (1961)
18. Caticha, A.: Entropic dynamics: quantum mechanics from entropy and information geometry. Ann. Phys. **531**(3), 1700408 (2019)
19. Vanchurin, V.: Entropic mechanics: towards a stochastic description of quantum mechanics. Found. Phys. **50**, 40–53 (2019)
20. Madelung, E.: Quantentheorie in hydrodynamischer Form. Z. Phys. **40**, 322–326 (1927). (**in German**)
21. Wallstrom, T.C.: Inequivalence between the Schrödinger equation and the Madelung hydrodynamic equations. Phys. Rev. A **49**, 1613–1617 (1994)
22. Feynman, R.P.: Statistical Mechanics: A Set of Lectures. Taylor and Francis, London (1998)
23. Abrikosov, A.A.: Fundamentals of the Theory of Metals. North-Holland, Amsterdam (1988)
24. Khrennikov, A.: Ubiquitous Quantum Structure: From Psychology to Finance. Springer, Berlin (2010)
25. Irkhin, V.Y., Katsnelson, M.I.: Wings of the Phoenix: Introduction to Quantum Mythophysics. Ural State Univ. Publ., Ekaterinburg (2003). (**in Russian**)
26. Katsnelson, M.I., Wolf, Y.I., Koonin, E.V.: Towards physical principles of biological evolution. Phys. Scr. **93**, 043001 (2018)
27. Penrose, R.: The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics. Oxford University Press, Oxford (1989)
28. Fisher, M.P.A.: Quantum cognition: The possibility of processing with nuclear spins in the brain. Ann. Phys. (NY) **362**, 593–602 (2015)
29. Apolloni, B., Carvalho, M.C., De Falco, D., Diego, D.: Quantum stochastic optimization. Stoc. Proc. Appl. **33**, 233–244 (1989)
30. Finnila, A.B., Gomez, M.A., Sebenik, C., Stenson, C., Doll, J.D.: Quantum annealing: A new method for minimizing multidimensional functions. Chem. Phys. Lett. **219**, 343–348 (1994)
31. Kadowaki, T., Nishimori, H.: Quantum annealing in the transverse Ising model. Phys. Rev. E. **58**, 5355–5363 (1998)
32. Brooke, J., Bitko, D., T. F., Rosenbaum, Aeppli, G.: Quantum annealing of a disordered magnet. Science 284 (1999) 779-781
33. Farhi, E., Goldstone, J., Gutmann, S., Lapan, J., Ludgren, A., Preda, D.: A Quantum adiabatic evolution algorithm applied to random instances of an NP-Complete problem. Science **292**, 472–475 (2001)
34. Childs, A.M., Farhi, E., Preskill, J.: Robustness of adiabatic quantum computation. Phys. Rev. A **65**, 012322 (2001)
35. Feynman, R.P., Hibbs, A.R.: Quantum Mechanics and Path Integrals. McGraw-Hill, New York (1965)
36. Schulman, L.S.: Techniques and Applications of Path Integration. Wiley, New York (1981)
37. Kleinert, H.: Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets, 5th edn. World Scientific, Singapore (2009)
38. Chaudhari, P., Soatto, S.: Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks. In: 2018 Information Theory and Applications Workshop (ITA), San Diego, CA, pp. 1–10 (2018)
39. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. Ser. **II**(106), 620–630 (1957)

40. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. Ser. **II**(108), 171–190 (1957)
41. Prigogine, I.: Etude Thermodynamique des phénoménes irréversibles. Desoer, Liége (1947)
42. Klein, M.J., Meijer, P.H.E.: Principle of minimum entropy production. Phys. Rev. **96**, 250–255 (1954)
43. Amari, S.: Information Geometry and its Applications. Springer, Tokyo (2016)
44. Vanchurin, V.: Covariant Information Theory and Emergent Gravity. Int. J. Mod. Phys. A **33**(34), 1845019 (2018)
45. Ruelle, D.: Statistical Mechanics: Rigorous Results. Imperial College Press, London (1999)
46. Bohr, N.: The quantum postulate and the recent development of atomic theory. Nature **121**, 580–590 (1928)
47. Wheeler, J.A., Zurek, W.H. (eds.): Quantum Theory and Measurement. Princeton University Press, Princeton (1983)
48. Berezin, F.A.: Feynman path integrals in a phase space. Sov. Phys. Usp. **23**, 763–788 (1980)
49. Jacobson, T.: Thermodynamics of space-time: the Einstein equation of state. Phys. Rev. Lett. **75**, 1260 (1995)
50. Padmanabhan, T.: Thermodynamical aspects of gravity: new insights. Rep. Prog. Phys. **73**(4), 046901 (2010)
51. Verlinde, E.P.: On the Origin of Gravity and the Laws of Newton. JHEP **1104**, 029 (2011)
52. De, S., Singh, T.P., Varma, A.: Quantum gravity as an emergent phenomenon. Int. J. Mod. Phys. D **28**(14), 1944003 (2019)
53. Witten, E.: Anti-de Sitter space and holography. Adv. Theor. Math. Phys. **2**, 253–291 (1998)
54. Susskind, L.: The World as a hologram. J. Math. Phys. **36**, 6377–6396 (1995)
55. Maldacena, J.M.: The Large N limit of superconformal field theories and supergravity. Int. J. Theor. Phys. **38**, 1113–1133 (1999)
56. Dvali, G.: Black holes as brains: neural networks with area law entropy. Fortsch. Phys. **66**(4), 1800007 (2018)