

The world as a neural network

Vitaly Vanchurin

Department of Physics, University of Minnesota, Duluth, Minnesota, 55812
Duluth Institute for Advanced Study, Duluth, Minnesota, 55804

E-mail: vvanchur@d.umn.edu

Abstract.

We discuss a possibility that the entire universe on its most fundamental level is a neural network. We identify two different types of dynamical degrees of freedom: “trainable” variables (e.g. bias vector or weight matrix) and “hidden” variables (e.g. state vector of neurons). We first consider stochastic evolution of the trainable variables to argue that near equilibrium their dynamics is well approximated by Madelung equations (with free energy representing the phase) and further away from the equilibrium by Hamilton-Jacobi equations (with free energy representing the Hamilton’s principal function). This shows that the trainable variables can indeed exhibit classical and quantum behaviors with the state vector of neurons representing the hidden variables. We then study stochastic evolution of the hidden variables by considering D non-interacting subsystems with average state vectors, $\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^D$ and an overall average state vector $\bar{\mathbf{x}}^0$. In the limit when the weight matrix is a permutation matrix, the dynamics of $\bar{\mathbf{x}}^\mu$ can be described in terms of relativistic strings in an emergent $D + 1$ dimensional Minkowski space-time. If the subsystems are minimally interacting, with interactions described by a metric tensor, then the emergent space-time becomes curved. We argue that the entropy production in such a system is a local function of the metric tensor which should be determined by the symmetries of the Onsager tensor. It turns out that a very simple and highly symmetric Onsager tensor leads to the entropy production described by the Einstein-Hilbert term. This shows that the learning dynamics of a neural network can indeed exhibit approximate behaviors described by both quantum mechanics and general relativity. We also discuss a possibility that the two descriptions are holographic duals of each other.

Contents

1	Introduction	1
2	Neural networks	3
3	Thermodynamics of learning	5
4	Entropic mechanics	6
5	Quantum mechanics	8
6	Hamiltonian mechanics	10
7	Hidden variables	11
8	Relativistic strings	13
9	Emergent gravity	15
10	Holography	17
11	Discussion	18

1 Introduction

Quantum mechanics is a remarkably successful paradigm for modeling physical phenomena on a wide range of scales ranging from 10^{-19} meters (i.e. high-energy experiments) to 10^{+26} meters (i.e. cosmological observations.) The paradigm is so successful that it is widely believed that on the most fundamental level the entire universe is governed by the rules of quantum mechanics and even gravity should somehow emerge from it. This is known as the problem of quantum gravity that so far has not been solved, but some progress had been made in context of AdS/CFT [1–3], loop quantum gravity [4–6] and emergent gravity [7–9]. Although extremely important, the problem of quantum gravity is not the only problem with quantum mechanics. The quantum framework also starts to fall apart with introduction of observers. Everything seems to work very well when observers are kept outside of a quantum system, but it is far less clear how to describe macroscopic observers in a quantum system such as the universe itself. The realization of the problem triggered an ongoing debate on the interpretations of quantum mechanics, which remains unsettled to this day. On one side of the debate, there are proponents of the many-worlds interpretation claiming that everything in the universe (including observers) must be governed by the Schrödinger equation [10], but then it is not clear how classical probabilities would emerge. On the other side of the debate, there are proponents of the hidden variables theories [11], but there it is also unclear what is the role of the complex wave-function in a purely statistical system. It is important to emphasize that a working definition of observers is necessary not only for settling some philosophical debates, but for understanding the results of real physical experiments and cosmological observations. In particular, a self-consistent and paradoxes-free definition of

observers would allow us to understand the significance of Bell’s inequalities [12] and to make probabilistic prediction in cosmology [13]. To resolve the apparent inconsistency (or incompleteness) in our description of the physical world, we shall entertain an idea of having a more fundamental theory than quantum mechanics. A working hypothesis is that on the most fundamental level the dynamics of the entire universe is described by a microscopic neural network which undergoes learning evolution. If correct, then not only macroscopic observers but, more importantly, quantum mechanics and general relativity should correctly describe the dynamics of the microscopic neural network in the appropriate limits.¹

In this paper we shall first demonstrate that near equilibrium the learning evolution of a neural network can indeed be modeled (or approximated) with the Madelung equations (see Sec. 5), where the phase of the complex wave-function has a precise physical interpretation as the free energy of a statistical ensemble of hidden variables. The hidden variables describe the state of the individual neurons whose statistical ensemble is given by a partition function and the corresponding free energy. This free energy is a function of the trainable variables (such as bias vector and weight matrix) whose stochastic and learning dynamics we shall study (see Sec. 4). Note that while the stochastic dynamics generically leads to the production of entropy (i.e. second law of thermodynamics) the learning dynamics generically leads to the destruction of entropy (i.e. second law of learning). As a result in the equilibrium the time-averaged entropy of the system remains constant and the corresponding dynamics can be modeled using quantum mechanics. It is important to note that the entropy (and entropy production) that we discuss here is the entropy of either hidden or trainable variables which need not vanish even for pure states. Of course, one can also discuss mixed states and then the corresponding von Neumann entropy gives an additional contribution to the total entropy.

The situation changes dramatically, whenever some of the degrees of freedom are not thermalized. While it should in principle be possible to model the thermalized degrees of freedom using quantum theory, the non-thermalized degrees of freedom are not likely to follow exactly the rules of quantum mechanics. We shall discuss two non-equilibrium limits: one which can nevertheless be described using classical physics (e.g. Hamiltonian mechanics) and the other one which can be described using gravitational physics (e.g. general relativity). The classical limit is relevant when the non-equilibrium evolution of the trainable variables is dominated by the entropy destruction due to learning, but the stochastic entropy production is negligible. The dynamics of such a system is well approximated by the Hamilton-Jacobi equations with free energy playing the role of the Hamilton’s principal function (see Sec. 6). The gravitational limit is relevant when even the hidden variables (i.e. state vectors of neurons) have not yet thermalized and the stochastic entropy production governs the non-equilibrium evolution of the system (see Sec. 9). In the long run all of the degrees of freedom must thermalize and then quantum mechanics should provide a correct description of the learning system.

It is well known that during learning the neural network is attracted towards a network with a low complexity, a phenomenon also known as dimensional reduction or what we call the second law of learning [16]. An example of a low complexity neural network is the one described by a permutation weight matrix or when the neural network is made out of one-dimensional chains of neurons.² If the set of state vectors can also be divided into non-interacting subsets (or subsystems) with average state vectors, $\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^D$ and an overall

¹The idea of using neural networks to describe gravity was recently explored in Ref. [14] in context of quantum neural networks, in Ref. [15] in context of AdS/CFT and in Ref. [16] in context of emergent gravity.

²A similar phenomenon was recently observed in context of the information graph flow [35].

average state vector $\bar{\mathbf{x}}^0$, then the dynamics of $\bar{\mathbf{x}}^\mu$ can be described with relativistic strings in an emergent $D + 1$ dimensional space-time (see Sec. 8). In general, the subsystems would interact and then the emergent space-time would be described by a gravitational theory such as general relativity (see Sec. 9). Note that, in either case, the main challenge is to figure out exactly which degrees of freedom have already thermalized (and thus can be modeled with quantum mechanics) and which degrees of freedom are still in the process of thermalization and should be modeled with other methods such as Hamiltonian mechanics or general relativity. In addition, we shall discuss yet another method which is motivated by the holographic principle and is particularly useful when the bulk neurons are still in the process of equilibration, but the boundary neurons have already thermalized (see Sec. 10).

The paper is organized as follows. In Sec. 2 we review the theory of neural networks and in Sec. 3 we discuss a thermodynamic approach to learning. In Sec. 4 we derive the action which governs dynamics of the trainable variables by applying the principle of stationary entropy production. The action is used to study the dynamics near equilibrium in Secs. 5 (which corresponds to quantum limit) and further away from equilibrium in Sec. 6 (which corresponds to classical limit). In Sec. 7 we study a non-equilibrium dynamics of the hidden variables and in Sec. 8 we argue that in certain limits the dynamics can be described in terms of relativistic strings in the emergent space-time. In Sec. 9 we apply the principle of stationary entropy production to derive the action which describes equilibration of the emergent space-time (which corresponds to gravitational limit) and in Sec. 10 we discuss when the gravitational theory can have a holographic dual description as a quantum theory. In Sec. 11 we summarize and discuss the main results of the paper.

2 Neural networks

We start with a brief review of the theory of neural networks by following the construction that was introduced in Ref. [16]. The neural network shall be defined as a neural septuple $(\mathbf{x}, \hat{P}_{in}, \hat{P}_{out}, \hat{w}, \mathbf{b}, f, H)$, where $\mathbf{x} \in \mathbb{R}^N$, is the state vector of neurons, \hat{P}_{in} and \hat{P}_{out} are the projection operators to subspaces spanned by respectively, N_{in} , input and, N_{out} , output neurons, $\hat{w} \in \mathbb{R}^{N \times N}$, is a weight matrix, $\mathbf{b} \in \mathbb{R}^N$ is a bias vector, $f : \mathbb{R} \rightarrow \mathbb{R}$ is an activation function and $H : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^{N \times N} \rightarrow \mathbb{R}$ is a loss function. This definition is somewhat different from the one usually used in the literature on machine learning, but we found that it is a lot more useful for analyzing physical theories in context of a microscopic neural network that we are interested in here. We shall not distinguish between different layers and so all N neurons are connected into a single neural network with connections described by a single $N \times N$ weight matrix, \hat{w} . The matrix can be viewed as an adjacency matrix of a weighted directed graph with neurons representing the nodes and elements of the weight matrix representing directed edges. However, we will distinguish between two different types of neurons: the boundary neurons, $N_\partial = N_{in} + N_{out}$, and the bulk neurons, $N_\not\partial = N - N_\partial$. Similarly, the boundary and the bulk projection operators are defined respectively as $\hat{P}_\partial = \hat{P}_{in} + \hat{P}_{out}$ and $\hat{P}_{\not\partial} = \hat{I} - \hat{P}_\partial$.

The state vector of neurons, $\mathbf{x} \in \mathbb{R}^N$, or just state vector, evolves in discrete time-steps according to equation

$$\mathbf{x}(t + 1) = \mathbf{f}(\hat{w}\mathbf{x}(t) + \mathbf{b}) \quad (2.1)$$

which can also be written in terms of components³

$$x_i(t+1) = f(w_{ij}x_j(t) + b_i). \quad (2.2)$$

A crucial simplification of the dynamical system (2.1) was to assume that the activation map $\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ acts separately on each component (2.2) with some activation function $f(x)$. Logistic function $f(x) = (1 + \exp(x))^{-1}$ and rectified linear unit $f(x) = \max(0, x)$ are some important examples of the activation function, but we shall use the hyperbolic tangent $f(x) = \tanh(x)$ which is also widely used in machine learning. The main reason is that the hyperbolic tangent is a smooth odd function with a finite support which greatly simplifies analytical calculations that we shall carry out in the paper.

The main problem in machine learning, or the main learning objective, is to find a bias vector, \mathbf{b} , and a weight matrix, \hat{w} , which minimize some suitably defined loss function $H(\mathbf{x}, \mathbf{b}, \hat{w})$. In what follows we shall consider two loss functions: the “bulk” loss and the “boundary” loss. The bulk loss function is defined as a local sum over all neurons

$$\begin{aligned} H(\mathbf{x}, \mathbf{b}, \hat{w}) &= \frac{1}{2} (\mathbf{x} - \mathbf{f}(\hat{w}\mathbf{x} + \mathbf{b}))^T (\mathbf{x} - \mathbf{f}(\hat{w}\mathbf{x} + \mathbf{b})) + V(\mathbf{x}) \\ &= \frac{1}{2} (x_i - f(w_{ij}x_j + b_i))(x_i - f(w_{ik}x_k + b_i)) + \sum_i V(x_i). \end{aligned} \quad (2.3)$$

The first term represents the sum over squares of local errors or, equivalently, differences between the state of a neuron before, x_i , and after, $f(w_{ij}x_j + b_i)$, a single execution of the activation map. The second term represents a local objective such as a binary classification of the signal x_i . For example, if $V(x_i) = -\frac{m}{2}x_i^2$, then the values of x_i closer to lower- and upper-bounds are rewarded and values in-between are penalized. Although the bulk loss is much easier to analyze analytically, in practice it is often more useful to define the boundary loss function by summing over only boundary neurons,

$$H_{\partial}(\mathbf{x}, \mathbf{b}, \hat{w}) = H(\hat{P}_{\partial}\mathbf{x}, \hat{P}_{\partial}\mathbf{b}, \hat{P}_{\partial}^T \hat{w} \hat{P}_{\partial}). \quad (2.4)$$

In fact the boundary loss is usually used in supervised learning, but, as was argued in [16], the bulk loss is more suitable for unsupervised learning tasks.

Instead of following the dynamics of the individual states, which might be challenging, one can use the principle of maximum entropy [17, 18] to derive a canonical ensemble of states [16]. The corresponding canonical partition function is

$$\mathcal{Z}(\beta, \mathbf{b}, \hat{w}) = \int d^N x e^{-\beta H(\mathbf{x}, \mathbf{b}, \hat{w})} \quad (2.5)$$

and the free energy is

$$F(\beta, \mathbf{b}, \hat{w}) = -\frac{1}{\beta} \log \mathcal{Z}(\beta, \mathbf{b}, \hat{w}). \quad (2.6)$$

At a constant “temperature”, $T = \beta^{-1} = \text{const}$, the ensemble can evolve with time either due to internal (or what we shall call hidden) dynamics of the state vector, $\mathbf{x}(t)$, or due to the external (or what we shall call training) dynamics of the bias vector, $\mathbf{b}(t)$, and weight matrix,

³Summations over repeated indices are implied everywhere in the paper unless stated otherwise. For example, $w_{ij}x_j = \sum_j w_{ij}x_j$, $\frac{\partial^2 F}{\partial q_k^2} = \sum_k \frac{\partial^2 F}{\partial q_k^2}$ and $\left(\frac{\partial F}{\partial q_k}\right)^2 = \sum_k \left(\frac{\partial F}{\partial q_k}\right)^2$.

$\hat{w}(t)$. The partition function for the bulk loss function (2.3) with a mass-term potential, $V(x_i) = -\frac{m}{2}x_i^2$, and a hyperbolic tangent activation function, $f(x) = \tanh(x)$, was calculated in [16] using Gaussian approximation. The result is

$$\mathcal{Z}(\beta, \mathbf{b}, \hat{w}) \approx (2\pi)^{N/2} \det \left(\hat{I}(1 - \beta m) + \beta \hat{G} \right)^{-1/2} \quad (2.7)$$

where

$$\hat{G} \equiv \left(\hat{I} - \hat{f}'\hat{w} \right)^T \left(\hat{I} - \hat{f}'\hat{w} \right) \quad (2.8)$$

and \hat{f}' is a diagonal matrix of first derivatives of the activation function,

$$f'_{ii} \equiv \left(\frac{df(y_i)}{dy_i} \right)_{y_i = w_{ij}\langle x_j \rangle + b_i}. \quad (2.9)$$

3 Thermodynamics of learning

Given the partition function, the average loss can be calculated by a simple differentiation,

$$U(\beta, \mathbf{b}, \hat{w}) = \langle H(\mathbf{x}, \mathbf{b}, \hat{w}) \rangle = -\frac{\partial}{\partial \beta} \log(\mathcal{Z}(\beta, \mathbf{b}, \hat{w})) = \frac{\partial}{\partial \beta} (\beta F(\beta, \mathbf{b}, \hat{w})). \quad (3.1)$$

If the neural network was trained for a long time, then the weight matrix and the bias vector are in a state which minimizes (at least locally) the average loss function and then its variations with respect to \hat{w} and \mathbf{b} must vanish,

$$\begin{aligned} \frac{\partial U(\beta, \mathbf{b}, \hat{w})}{\partial w_{ij}} &= \frac{\partial^2}{\partial w_{ij} \partial \beta} (\beta F(\beta, \mathbf{b}, \hat{w})) = 0 \\ \frac{\partial U(\beta, \mathbf{b}, \hat{w})}{\partial b_i} &= \frac{\partial^2}{\partial b_i \partial \beta} (\beta F(\beta, \mathbf{b}, \hat{w})) = 0. \end{aligned} \quad (3.2)$$

We shall call this state, the state of the learning equilibrium. An important property of the equilibrium, which follows from (3.2), is that the total free energy must decompose into a sum of two terms

$$F(\beta, \mathbf{b}, \hat{w}) = A(\beta) - \frac{1}{\beta} C(\mathbf{b}, \hat{w}). \quad (3.3)$$

Likewise, the total entropy must also decompose into a sum of two terms,

$$S_x(\beta, \mathbf{b}, \hat{w}) = \beta^2 \frac{\partial}{\partial \beta} F(\beta, \mathbf{b}, \hat{w}) = \beta^2 \frac{\partial}{\partial \beta} \left(A(\beta) - \frac{1}{\beta} C(\mathbf{b}, \hat{w}) \right) = S_0(\beta) + C(\mathbf{b}, \hat{w}) \quad (3.4)$$

where the first term is the familiar thermodynamic entropy

$$S_0(\beta) = \beta^2 \frac{\partial A(\beta)}{\partial \beta} = \beta(U(\beta) - A(\beta)). \quad (3.5)$$

and the second term, $C(\mathbf{b}, \hat{w})$, is related to the complexity of the neural network (see Ref. [16]).

As the learning progresses, the average loss, $U(\beta)$, decreases, the temperature parameter, β^{-1} , decreases and, thus, one might expect that the thermodynamic entropy, S_0 , should also decrease. However, it is not the thermodynamic entropy, S_0 , but the total entropy, S_x ,

(whose exponent describes accessible volume of the configuration space for \mathbf{x}) should decrease with learning. We call it the second law of learning:

Second Law of Learning: *the total entropy of a learning system can never increase during learning and is constant in a learning equilibrium,*

$$\frac{d}{dt}S_x \leq 0. \quad (3.6)$$

In the long run the system is expected to approach an equilibrium state with the smallest possible total entropy, S_x , which corresponds to the lowest possible sum of the thermodynamic entropy, $S_0(\beta)$, and of the complexity function $C(\mathbf{b}, \hat{w})$.

For a system transitioning between equilibrium states at constant temperature, $T = 1/\beta$, variations of the free energy must vanish, $dF = 0$, and then equation (3.3) takes the form of the first law,

$$dA - TdC = dU - TdS_x = dU - TdS_0 - TdC = 0, \quad (3.7)$$

or what we call the first law of learning:

First Law of Learning: *the increment in the loss function is proportional to the increment in the thermodynamic entropy plus the increment in the complexity*

$$dU = TdS_x = TdS_0 + TdC. \quad (3.8)$$

4 Entropic mechanics

So far the neural networks were analyzed by considering statistical ensembles of the state vectors, \mathbf{x} , but the bias vector, \mathbf{b} , and weight matrix, \hat{w} , were treated deterministically. The next step is to promote \mathbf{b} and \hat{w} to stochastic variables in order to study their near-equilibrium dynamics. In the next section we will show that the training dynamics of \mathbf{b} and \hat{w} can be approximated by Madelung equations with \mathbf{x} playing the role of the hidden variables. For this reason, we shall refer to the bias vectors and weight matrices as “trainable” variables and to the state vectors as “hidden” variables. This does not mean that the trainable variables are the quantized versions of the corresponding classical variables, but only that their stochastic evolution near equilibrium can often be described by quantum mechanics.

Consider a family of trainable variables, $\mathbf{b}(\mathbf{q})$ and $\hat{w}(\mathbf{q})$, parametrized by dynamical parameters q_k 's where $k \in (1, \dots, K)$. Typically the number of parameters K is much smaller than $N + N^2$ (i.e. the number of parameters required to describe a generic vector \mathbf{b} and a generic matrix \hat{w}) and the art of designing a neural architecture is to come up with functions $\mathbf{b}(\mathbf{q})$ and $\hat{w}(\mathbf{q})$ which are most efficient in finding solutions. To make the statement more quantitative, consider an ensemble of neural networks described by a probability distribution $p(t, \mathbf{q})$ which evolves with time according to a Fokker-Planck equation

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial q_k} \left(D \frac{\partial p}{\partial q_k} - \frac{dq_k}{dt} p \right). \quad (4.1)$$

If we assume that the learning evolution (or the drift) is in the direction of the gradient of the free energy,

$$\frac{dq_k}{dt} = \gamma \frac{\partial F}{\partial q_k} \quad (4.2)$$

then

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial q_k} \left(D \frac{\partial p}{\partial q_k} - \gamma \frac{\partial F}{\partial q_k} p \right). \quad (4.3)$$

This may be a good guess on short-time scales when the free energy does not change much, but in general both $p(t, \mathbf{q})$ and $F(t, \mathbf{q})$ can depend on time explicitly and implicitly through variable \mathbf{q} . To describe such dynamics we shall employ the principle of stationary entropy production (see Ref. [31]):

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

The principle can be thought of as a generalization of both, the maximum entropy principle [17, 18] and the minimum entropy production principle [19, 20] which is often used in non-equilibrium thermodynamics. In context of neural networks it is beneficial to have large entropy as it implies a higher rate with which new solutions can be discovered. Then the optimal neural architecture should be the one for which the entropy destruction is minimized or, equivalently, the entropy production is maximized. This justifies the use of the principle in context of the optimal learning systems [16].

The Shannon entropy of the distribution $p(t, \mathbf{q})$ (not to confuse with $S_x(\beta, \mathbf{q})$) is given by

$$S_q(t) = - \int d^K q p(t, \mathbf{q}) \log(p(t, \mathbf{q})). \quad (4.4)$$

and using (4.3) the entropy production is given by

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

which can be simplified (after integrating by parts and ignoring the boundary terms, i.e. by assuming periodic or vanishing boundary conditions),

$$\begin{aligned} \frac{dS_q}{dt} &= \int d^K q \frac{\partial p}{\partial q_k} \left(\frac{D}{p} \frac{\partial p}{\partial q_k} - \gamma \frac{\partial F}{\partial q_k} \right) \\ &= \int d^K q \sqrt{p} \left(-4D \frac{\partial^2}{\partial q_k^2} + \gamma \frac{\partial^2}{\partial q_k^2} F \right) \sqrt{p}. \end{aligned} \quad (4.5)$$

This quantity is a functional of both $p(t, \mathbf{q})$ and $F(t, \mathbf{q})$ and, thus, in addition to modeling the dynamics of the probability distribution we must also model the dynamics of the free energy.

The total rate of change of the free energy is given by

$$\frac{d}{dt} F(t, \mathbf{q}) = \frac{\partial F(t, \mathbf{q})}{\partial t} + \frac{dq_k}{dt} \frac{\partial F(t, \mathbf{q})}{\partial q_k} = \frac{\partial F(t, \mathbf{q})}{\partial t} + \gamma \left(\frac{\partial F(t, \mathbf{q})}{\partial q_k} \right)^2 \quad (4.6)$$

where the first term represents the change of the free energy due to dynamics of hidden variables, \mathbf{x} , and the second term represents the change in the free energy due to dynamics

of trainable variables, \mathbf{b} and \hat{w} . In what follows, it will be convenient to denote the time-averaged rate of change of free energy as

$$\left\langle \frac{d}{dt} F(t, \mathbf{q}) \right\rangle_t \equiv -V(\mathbf{q}). \quad (4.7)$$

Then, according to the principle of stationary entropy production, the dynamics of $p(t, \mathbf{q})$ and $F(t, \mathbf{q})$ must be such that the entropy production is extremized subject to a constraint

$$\frac{\partial F}{\partial t} + \gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + V = 0. \quad (4.8)$$

The optimization problem can be solved by defining the following ‘‘action’’,

$$\mathcal{S}_q[p, F] = \int_0^T dt \frac{d\mathcal{S}_q}{dt} + \mu \int_0^T dt d^K q p \left(\frac{\partial F}{\partial t} + \gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + V \right), \quad (4.9)$$

where μ is a Lagrange multiplier, and then the ‘‘equations of motion’’ are obtained by setting variations of the action to zero,

$$\frac{\delta \mathcal{S}_q}{\delta p} = \frac{\delta \mathcal{S}_q}{\delta F} = 0. \quad (4.10)$$

5 Quantum mechanics

In the previous section we developed a stochastic description of the trainable variables \mathbf{q} which describe the weight matrix $\hat{w}(\mathbf{q})$ and the bias vector $\mathbf{b}(\mathbf{q})$. We argued that on short time-scales the dynamics of the probability distribution $p(t, \mathbf{q})$ and of the free energy $F(t, \mathbf{q})$ is given by equations (4.3) and (4.6), but on longer time-scales an approximate dynamics can be obtained using the principle of stationary entropy production. The corresponding ‘‘action’’ is given by (4.9) which can be rewritten using (4.5),

$$\mathcal{S}_q[p, F] = \int_0^T dt d^K q \sqrt{p} \left(-4D \frac{\partial^2}{\partial q_k^2} + \gamma \frac{\partial^2}{\partial q_k^2} F + \mu \frac{\partial F}{\partial t} + \mu \gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + \mu V \right) \sqrt{p}. \quad (5.1)$$

The five terms on the right hand side represent:

- (1) $-4D \frac{\partial^2}{\partial q_k^2}$, entropy production due to stochastic dynamics of q_k 's,
- (2) $\gamma \frac{\partial^2 F}{\partial q_k^2}$, entropy production due to learning dynamics of q_k 's,
- (3) $\mu \frac{\partial F}{\partial t}$, free energy production due to dynamics of x_i 's
- (4) $\mu \gamma \left(\frac{\partial F}{\partial q_k} \right)^2$, free energy production due to learning dynamics of q_k 's,
- (5) μV , the (negative of) total time-averaged free energy production.

Note that the entropy production due to stochastic dynamics is usually positive (due to the second law of thermodynamics), but the entropy production due to learning dynamics is usually negative (due to the second law of learning). While the learning entropy production is expected to dominate the dynamics far away from an equilibrium, the stochastic entropy production is expected to give the main contribution near equilibrium.

From (5.1) the equations of motion (4.10) are obtained by setting variations to zero,

$$\frac{\delta \mathcal{S}_q[p, F]}{\delta F} = \gamma \frac{\partial^2}{\partial q_k^2} p - \mu \frac{\partial}{\partial t} p - 2\mu\gamma \frac{\partial}{\partial q_k} \left(\frac{\partial F}{\partial q_k} p \right) = 0 \quad (5.2)$$

$$\frac{\delta \mathcal{S}_q[p, F]}{\delta p} = -\frac{4D}{\sqrt{p}} \frac{\partial^2 \sqrt{p}}{\partial q_k^2} + \gamma \frac{\partial^2 F}{\partial q_k^2} + \mu \frac{\partial F}{\partial t} + \mu\gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + \mu V = 0. \quad (5.3)$$

It is convenient to define a velocity vector

$$u_k \equiv 2\gamma \frac{\partial}{\partial q_k} F. \quad (5.4)$$

and then (5.2) can be expressed as a Fokker-Planck equation

$$\frac{\partial}{\partial t} p = -\frac{\partial}{\partial q_k} (u_k p) + \boxed{\frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} p} \quad (5.5)$$

and (5.3) as a Navier-Stokes equation (after differentiating with respect to $\frac{\partial}{\partial q_j}$)

$$\frac{\partial}{\partial t} u_j + u_k \frac{\partial}{\partial q_k} u_j + \boxed{\frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} u_j} = -2\gamma \frac{\partial}{\partial q_j} \left(V - \frac{4D}{\mu\sqrt{p}} \frac{\partial^2 \sqrt{p}}{\partial q_k^2} \right). \quad (5.6)$$

Several comments are in order. First of all, the Fokker-Planck equation (5.5) differs from the “stochastic” Fokker-Planck equation (4.3). This is a consequence of our assumption that (4.3) is only valid on very short time scales, while, according to the principle of stationary entropy production, equations (5.5) and (5.6) must be valid on much longer time-scales. Secondly, if $\mu > 0$ then the kinetic viscosity in the Navier-Stokes equation (5.6), $-\frac{\gamma}{\mu}$, is negative which is a consequence of the second law of learning. And finally, if we neglect the entropy production due to learning (i.e. $\gamma \frac{\partial^2 F}{\partial q_k^2}$ in (5.1)), then the resulting equations of motion would be the same as (5.5) and (5.6), but with terms in boxes set to zero. These are the well known Madelung equations which are equivalent to the Schrödinger equation

$$-i\sqrt{\frac{4D}{\gamma}} \frac{\partial}{\partial t} \Psi = \left(4D \frac{\partial^2}{\partial q_k^2} - V \right) \Psi \quad (5.7)$$

for the wave-function defined as

$$\Psi \equiv \sqrt{p} \exp \left(i\sqrt{\frac{\gamma}{4D}} F \right). \quad (5.8)$$

Moreover, in this limit the action (5.1) takes the form of the Schrödinger action

$$\mathcal{S}_q[\Psi] = \int_0^T dt \int d^K q \Psi^* \left(-4D \frac{\partial^2}{\partial q_k^2} + V - i\sqrt{\frac{4D}{\gamma}} \frac{\partial}{\partial t} \right) \Psi. \quad (5.9)$$

Therefore, we conclude that near equilibrium, i.e. when the first term in (5.1) is much larger than the second term, our system can be modeled by quantum mechanics.

6 Hamiltonian mechanics

The next step is to consider a non-equilibrium dynamics of the trainable variables which is relevant when the second term in (5.1) is much larger than the first term. This corresponds to a limit when the entropy destruction is dominated by the learning dynamics and the stochastic entropy production is negligible. The corresponding Fokker-Planck equation remains the same as before (5.5), but the Navier-Stokes equation (5.6) is greatly simplified

$$\frac{\partial}{\partial t} u_j + u_k \frac{\partial}{\partial q_k} u_j + \frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} u_j = -2\gamma \frac{\partial}{\partial q_j} V. \quad (6.1)$$

In this limit the dynamics of the free energy F does not depend on the probability distribution p and thus equation (6.1) decouples from (5.5) and can be solved separately. In terms of the free energy the equation of motion (5.3) is

$$-\frac{\partial F}{\partial t} = V + \gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + \frac{\gamma}{\mu} \frac{\partial^2 F}{\partial q_k^2} \quad (6.2)$$

which can be thought of as a Hamilton-Jacobi equation for the Hamilton's principle function F and a Hamiltonian function

$$H \left(q_k, \frac{\partial F}{\partial q_k}, \frac{\partial^2 F}{\partial q_k^2} \right) = V + \gamma \left(\frac{\partial F}{\partial q_k} \right)^2 + \frac{\gamma}{\mu} \frac{\partial^2 F}{\partial q_k^2}. \quad (6.3)$$

Note, however, that in classical mechanics the Hamiltonian function depends only on q_k 's and $\frac{\partial F}{\partial q_k}$'s, but in our case it also depends on one more variable $\sum_k \frac{\partial^2 F}{\partial q_k^2}$.

From equations (4.2) and (5.4) we get

$$\frac{dq_j}{dt} = \gamma \frac{\partial F}{\partial q_j} = \frac{1}{2} u_j \quad (6.4)$$

and then (6.2) can be rewritten as

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + \frac{dq_k}{dt} \frac{\partial F}{\partial q_k} = -\frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} F - V. \quad (6.5)$$

In the limit when the entropy production (due to both learning and stochastic dynamics) is negligible, i.e. $\left| \frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} F \right| \ll |V|$, equations (6.4) and (6.5) can be used to obtain classical equations of motion

$$\frac{d^2 q_j}{dt^2} = -\gamma \frac{\partial V}{\partial q_j}. \quad (6.6)$$

In the opposite limit, $|V| \ll \left| \frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} F \right|$, the equation for free energy (6.5) takes the following form

$$\frac{\partial F}{\partial t} = -\gamma \left(\frac{\partial F}{\partial q_k} \right)^2 - \frac{\gamma}{\mu} \frac{\partial^2}{\partial q_k^2} F. \quad (6.7)$$

which has a simple time-independent (i.e. $\frac{\partial F}{\partial t} = 0$) solution given by,

$$F = C_0 + \frac{1}{\mu} \sum_k \log(C_k + \mu q_k) \quad (6.8)$$

where C_0 and C_k 's are arbitrary coefficients. Note that $\frac{\partial F}{\partial t} = 0$ corresponds to a limit when the change in the free energy production due to dynamics of x_i 's is negligible or in other words when the training dataset is not dynamical (as is often the case in machine learning).

The solution (6.8) has an exact form of the free energy for a canonical ensemble (2.7),

$$F = \frac{1}{2\beta} \log \det((1 - \beta m) + \beta \hat{G}) - \frac{N}{2\beta} \log(2\pi) = \frac{1}{2\beta} \sum_i \log((1 - \beta m) + \beta \lambda_i) - \frac{N}{2\beta} \log(2\pi), \quad (6.9)$$

with $\mu = 2\beta$ and the dynamical variables q_i set to the eigenvalues λ_i of the operator \hat{G} . In this limit the average loss is

$$U = \frac{\partial(\beta F)}{\partial \beta} = \frac{1}{2} \sum_i \frac{\lambda_i}{1 + \beta \lambda_i} = \lambda_i \frac{\partial F}{\partial \lambda_i}, \quad (6.10)$$

where for simplicity we have set the mass parameter to zero, $m = 0$. This equation can be thought of as a virial theorem for our learning system where $\frac{\partial F}{\partial \lambda_i}$ is the ‘‘force’’ acting on a ‘‘particle’’ at position λ_i . More generally, the eigenvalues λ_i 's could be arbitrary functions of q_i 's and time t and then

$$\begin{aligned} \frac{\gamma}{\mu} \sum_k \frac{\partial^2 F}{\partial q_k^2} &= \frac{\gamma}{\mu} \sum_{i,j,k} \frac{\partial^2 F}{\partial \lambda_i \partial \lambda_j} \frac{\partial \lambda_i}{\partial q_k} \frac{\partial \lambda_j}{\partial q_k} = -\frac{\gamma\beta}{2\mu} \sum_{i,k} ((1 - \beta m) + \beta \lambda_i)^{-2} \left(\frac{\partial \lambda_i}{\partial q_k} \right)^2 \\ &= -\frac{2\gamma\beta}{\mu} \sum_{i,k} \left(\frac{\partial F}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial q_k} \right)^2 = -\frac{2\gamma\beta}{\mu} \sum_{i,j,k} \frac{\partial F}{\partial \lambda_i} \left(\frac{\partial \lambda_i}{\partial q_k} \delta_{ij} \frac{\partial \lambda_j}{\partial q_k} \right) \frac{\partial F}{\partial \lambda_j} \\ &= -\frac{2\gamma\beta}{\mu} \sum_{i,j,k,m,n} \frac{\partial F}{\partial q_m} \left(\frac{\partial q_m}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial q_k} \delta_{ij} \frac{\partial \lambda_j}{\partial q_k} \frac{\partial q_n}{\partial \lambda_j} \right) \frac{\partial F}{\partial q_n} \\ &= -\frac{2\gamma\beta}{\mu} \sum_{i,k,m,n} \frac{\partial F}{\partial q_m} \left(\frac{\partial q_m}{\partial \lambda_i} \left(\frac{\partial \lambda_i}{\partial q_k} \right)^2 \frac{\partial q_n}{\partial \lambda_i} \right) \frac{\partial F}{\partial q_n} \end{aligned} \quad (6.11)$$

where we assumed that $\frac{\partial \lambda_i}{\partial q_j}$ is invertible. This implies that for the canonical free energy (6.9) the Hamiltonian function (6.3) can be written in terms of only first derivatives of the Hamilton's principle function F ,

$$H \left(q_k, \frac{\partial F}{\partial q_k} \right) = V + \gamma \frac{\partial F}{\partial q_m} \left(\delta_{mn} - \frac{2\beta}{\mu} \left(\frac{\partial q_m}{\partial \lambda_i} \left(\frac{\partial \lambda_i}{\partial q_k} \right)^2 \frac{\partial q_n}{\partial \lambda_i} \right) \right) \frac{\partial F}{\partial q_n}, \quad (6.12)$$

and, thus, the system is Hamiltonian although the kinetic term may not be canonical.

7 Hidden variables

We have seen that neural networks can exhibit both quantum (Sec. 5) and classical (Sec. 6) behaviors if the dynamics of the trainable variables \mathbf{q} (or equivalently of the bias vector \mathbf{b} and weight matrix \hat{w}) is followed explicitly, but the dynamics of the hidden variables (or the state vectors \mathbf{x}) was expressed only implicitly through $\frac{\partial F}{\partial t}$. For this reason it was convenient to think of the state vectors \mathbf{x} as hidden random variables whose individual dynamics was shadowed by our statistical description. In this section we shall be interested instead in

a non-equilibrium dynamics of the hidden variables which is relevant, for example, on the time-scales that are much smaller than thermalization time.

Recall that the state of the individual neurons evolves according to (2.1) which can be approximated to the leading order as

$$\bar{x}_i^{(0)}(t+1) \approx \left(\hat{f}'_0 \hat{w} \right)_{ij} \bar{x}_j^{(0)}(t) \quad (7.1)$$

where $\hat{f}'_0 = \hat{f}'$ is the matrix of first derivative of the activation function (2.9). More generally, we can consider D non-interacting subsystems of states vectors (e.g. D separate sets of training data) denoted by $\mathbf{x}^{(d)}$ where $d = 1, \dots, D$. Then the overall distribution of the state vectors is in general multimodal with D local maxima, $\bar{\mathbf{x}}^{(d)}$, and each of these maxima evolves according to

$$\bar{x}_i^{(d)}(t+1) \approx \left(\hat{f}'_d \hat{w} \right)_{ij} \bar{x}_j^{(d)}(t) \quad (7.2)$$

where

$$\bar{\mathbf{x}}^{(0)} = \sum_d \bar{\mathbf{x}}^{(d)}. \quad (7.3)$$

and

$$\left(\hat{f}'_d \right)_{ii} \equiv \left(\frac{df(y_i)}{dy_i} \right)_{y_i = w_{ij} \bar{x}_j^{(d)} + b_i}. \quad (7.4)$$

It is convenient to define a continuous time coordinate τ such that

$$\frac{\partial}{\partial \tau} \bar{x}_i^{(\mu)}(\tau) = \alpha (\bar{x}_i^{(\mu)}(t+1) - \bar{x}_i^{(\mu)}(t)) \quad (7.5)$$

where $\mu = 0, 1, \dots, D$ and α is an auxiliary parameter. Although the different subsystems are represented by different hidden variables $\mathbf{x}^{(d)}$'s, they are all processed by the very same neural network described by the same trainable variable \mathbf{b} and \hat{w} . With this respect the hidden variables are not interacting directly with each other, but they are interacting (minimally) through the trainable variables, \mathbf{b} and \hat{w} . If such (minimal) interactions are negligible, then $\frac{\partial \bar{x}_i^{(c)}}{\partial \tau} \frac{\partial \bar{x}_i^{(d)}}{\partial \tau} \propto \delta_{cd}$ with no summations over index i . Then

$$\frac{\partial \bar{x}_i^{(0)}}{\partial \tau} \frac{\partial \bar{x}_i^{(0)}}{\partial \tau} = \sum_d \frac{\partial \bar{x}_i^{(d)}}{\partial \tau} \frac{\partial \bar{x}_i^{(d)}}{\partial \tau} \quad \text{for all } i \quad (7.6)$$

or

$$\eta_{\mu\nu} \frac{\partial \bar{x}_i^{(\mu)}}{\partial \tau} \frac{\partial \bar{x}_i^{(\nu)}}{\partial \tau} = 0 \quad \text{for all } i \quad (7.7)$$

where $\eta = \text{diag}(-1, 1, \dots, 1)$. However, in general the minimal interactions cannot be ignored and then

$$g^i_{\mu\nu} \frac{\partial \bar{x}_i^{(\mu)}}{\partial \tau} \frac{\partial \bar{x}_i^{(\nu)}}{\partial \tau} = 0 \quad \text{for all } i \quad (7.8)$$

where the metric tensor $g^i_{\mu\nu}$ describes the strength of the interactions. Of course, such a description is only valid if the minimal interactions are weak which is the assumption we are going to make.

To estimate the dynamics of hidden variables \bar{x}^μ we assume that the activation function is linear $\hat{f}'_d = \hat{I}$ (with the slope set to one without loss of generality) and then from (7.1) and (7.2) we have

$$\bar{x}^{(\mu)}(t+1) \approx w_{ij} \bar{x}_j^{(\mu)} \quad (7.9)$$

and (7.5) becomes

$$\frac{\partial \bar{x}_i^{(\mu)}}{\partial \tau} \approx \alpha (w_{ij} - \delta_{ij}) \bar{x}_j^{(\mu)}. \quad (7.10)$$

According to the second law of learning it is expected that the neural network must have evolved to a network with a very low complexity such as a network whose weight matrix is a permutation matrix

$$\hat{w} = \hat{\pi}. \quad (7.11)$$

For example, consider a permutation matrix with only a single cycle which (up to permutations of elements) is given by

$$\pi_{ij} = \begin{cases} 1 & \text{if } i - 1 = j \pmod{N} \\ 0 & \text{otherwise.} \end{cases} \quad (7.12)$$

Then equation (7.10) can be rewritten as

$$\frac{\partial \bar{x}_i^{(\mu)}}{\partial \tau} = \alpha \bar{x}_{i-1 \pmod{N}}^{(\mu)}(t) - \alpha \bar{x}_i^{(\mu)}(t). \quad (7.13)$$

If we take a continuous limit by defining $\bar{x}^{(\mu)}(\tau, \sigma)$ such that

$$\frac{\partial}{\partial \sigma} \bar{x}^{(\mu)}(\tau, \sigma) = \alpha (\bar{x}_i^{(\mu)}(t) - \bar{x}_{i-1 \pmod{N}}^{(\mu)}(t)) \quad (7.14)$$

then (7.13) becomes

$$\frac{\partial \bar{x}^{(\mu)}}{\partial \tau} = -\frac{\partial \bar{x}^{(\mu)}}{\partial \sigma}. \quad (7.15)$$

This equation has a simple solution of a periodic “right-moving” wave. In the light-cone coordinates $\xi^\pm \equiv \tau \pm \sigma$, the equation of motion (7.15) is

$$\frac{\partial \bar{x}^{(\mu)}}{\partial \xi^+} = 0 \quad (7.16)$$

and the constraint equation (7.7) is

$$\eta_{\mu\nu} \frac{\partial \bar{x}^{(\mu)}}{\partial \xi^-} \frac{\partial \bar{x}^{(\nu)}}{\partial \xi^-} = 0. \quad (7.17)$$

8 Relativistic strings

In the last section we have shown that an equation for a “right-moving” wave (7.16) can emerge in a statistical description of D minimally-interacting subsystems of state vectors. A natural question arises if a “left-moving” wave can also emerge in some limit and if so can the dynamics be described in terms of relativistic strings in an emergent space-time?

To answer this question we first note that the permutation weight matrix (7.11) (with an arbitrary number of cycles) is such that,

$$\hat{\pi}^T \hat{\pi} = \hat{\pi} \hat{\pi}^T = \hat{I} \quad (8.1)$$

and thus

$$\hat{G}(\hat{\pi}) = (\hat{\pi} - \hat{I})^T (\hat{\pi} - \hat{I}) = \hat{I} - \hat{\pi} - \hat{\pi}^T + \hat{\pi}^T \hat{\pi} = \hat{G}(\hat{\pi}^T). \quad (8.2)$$

Since the free energy (6.9) depends on $\hat{\pi}$ only through \hat{G} the very same ensemble of the state vectors can equally likely evolve either towards $\hat{\pi}$ or towards $\hat{\pi}^T$. However, if the exact state of the microscopic weight matrix is unknown, then one must consider an ensemble which contains both options and then the average state vector is given by

$$\bar{x}_i^{(\mu)} = \frac{1}{2} \int d^N x^{(\mu)} p(\mathbf{x}^{(\mu)}, \hat{\pi}) x_i^{(\mu)} + \frac{1}{2} \int d^N x^{(\mu)} p(\mathbf{x}^{(\mu)}, \hat{\pi}^T) x_i^{(\mu)} = \frac{1}{2} \bar{x}_i^{(\mu-)} + \frac{1}{2} \bar{x}_i^{(\mu+)} \quad (8.3)$$

where the two terms represent statistical averages with respect to the two distributions.

Following the analysis of the previous section the dynamics of $\bar{x}_i^{(\mu-)}$ and $\bar{x}_i^{(\mu+)}$ can be obtained from (7.10) for the respective weight matrices,

$$\frac{\partial \bar{x}_i^{(\mu-)}}{\partial \tau} \approx \alpha (\pi_{ij} - \delta_{ij}) \bar{x}_j^{(\mu-)} \quad (8.4)$$

$$\frac{\partial \bar{x}_i^{(\mu+)}}{\partial \tau} \approx \alpha (\pi_{ij}^T - \delta_{ij}) \bar{x}_j^{(\mu+)}. \quad (8.5)$$

In a continuum limit the equations are given by

$$\frac{\partial \bar{x}^{(\mu-)}}{\partial \tau} = - \frac{\partial \bar{x}^{(\mu-)}}{\partial \sigma} \quad (8.6)$$

$$\frac{\partial \bar{x}^{(\mu+)}}{\partial \tau} = + \frac{\partial \bar{x}^{(\mu+)}}{\partial \sigma} \quad (8.7)$$

whose solutions represent respectively the right- and left-moving waves. Then the dynamics of the hidden variables (8.3) is indeed given by a 1 + 1 dimensional wave equation

$$\frac{\partial^2 \bar{x}^{(\mu)}}{\partial \tau^2}(\tau, \sigma) = \frac{\partial^2 \bar{x}^{(\mu)}}{\partial \sigma^2}(\tau, \sigma). \quad (8.8)$$

In the light-cone coordinates the wave equation is

$$\frac{\partial}{\partial \xi^-} \frac{\partial}{\partial \xi^+} \bar{x}^{(\mu)}(\tau, \sigma) = 0 \quad (8.9)$$

and the constraints

$$\eta_{\mu\nu} \frac{\partial \bar{x}^{(\mu)}}{\partial \xi^-} \frac{\partial \bar{x}^{(\nu)}}{\partial \xi^-} = \eta_{\mu\nu} \frac{\partial \bar{x}^{(\mu)}}{\partial \xi^+} \frac{\partial \bar{x}^{(\nu)}}{\partial \xi^+} = 0. \quad (8.10)$$

The action which gives rise to the wave equations (8.9) and constraints (8.10) is the Polyakov action which can be written in a covariant form as

$$\mathcal{A} = \int d\sigma d\tau \sqrt{-h} h^{ab} \eta_{\mu\nu} \frac{\partial \bar{x}^{(\mu)}}{\partial \xi^a} \frac{\partial \bar{x}^{(\nu)}}{\partial \xi^b} \quad (8.11)$$

where h_{ab} is the world-sheet metric and h is its determinant.

In summary, we showed that D non-interacting subsystems of the state vectors $\mathbf{x}^{(d)}$ can be described with $D + 1$ scalar fields in $1 + 1$ dimensions. Alternatively one can view the configuration space of the scalar fields as an emergent space-time and then our system can be described with a motion of relativistic strings in $D + 1$ dimensions (8.11). This is very similar to what is usually done in string theory, with one major difference. Our strings arise from the dynamics of the average state vectors $\bar{\mathbf{x}}^{(\mu)}$ and not from the dynamics of the bias vector \mathbf{b} and weight matrix \hat{w} which undergo learning. Recall that the trainable variables \mathbf{b} and \hat{w} (or equivalently \mathbf{q}) near equilibrium can be modeled by quantum mechanics (Sec. 5) and further away from the equilibrium by classical mechanics (Sec. 6). In contrast, the state vectors $\bar{\mathbf{x}}^{(\mu)}$ represent hidden variables of the quantum theory, but their dynamics (in certain limits) is conveniently described by relativistic strings.

9 Emergent gravity

Consider a discrete action for the hidden variables (or state vectors),

$$\mathcal{A} = g_{\mu\nu}^i \left(\alpha^2 \left\langle x_i^{(\mu)} G_{ij} x_j^{(\nu)} \right\rangle_x - \frac{d\bar{x}_i^{(\mu)}}{d\tau} \frac{d\bar{x}_i^{(\nu)}}{d\tau} \right). \quad (9.1)$$

where $g_{\mu\nu}^i$ describes interactions between the subsystems (7.8). This action is a lot more general than (8.11), but it can be approximated by the string action for a flat target space, $g_{\mu\nu}^i = \eta_{\mu\nu}$, for a permutation weight matrix, $\hat{w} = \hat{\pi}$, and for a linear activation function $\hat{f}'_d = \hat{I}$. To study the dynamics in the emergent space-time it is convenient to rewrite (9.1) as

$$\mathcal{A} = \int d^D X \sqrt{-g} g_{\mu\nu} T^{\mu\nu} \quad (9.2)$$

where g is the determinant of $g_{\mu\nu}$ and

$$\sqrt{-g} T^{\mu\nu} \equiv \left(\alpha^2 \left\langle x_i^{(\mu)} G_{ij} x_j^{(\nu)} \right\rangle_x - \frac{d\bar{x}_i^{(\mu)}}{d\tau} \frac{d\bar{x}_i^{(\nu)}}{d\tau} \right) \prod_{\alpha} \delta(X^{\alpha} - \bar{x}_i^{\alpha}) \quad (9.3)$$

is the energy-momentum tensor density.

The equilibrium dynamics of neural networks was first modeled using the principle of maximum entropy with a constraint imposed on the loss function [16], but to study a non-equilibrium dynamics of the trainable variables the principle of the stationary entropy production had to be used with a constraint was imposed on the dynamics of free energy (4.8). In this section we study a non-equilibrium dynamics of the hidden variables, and so the constraint should be imposed on the action which describes the dynamics of the state vectors (9.2). Then, according to the principle of stationary entropy production, the quantity which must be extremized is

$$\mathcal{S}_x[g] = \int d^{D+1} X \sqrt{-g} \mathcal{R}(g) + \kappa \left(A - \int d^{D+1} X \sqrt{-g} g_{\mu\nu} T^{\mu\nu} \right) \quad (9.4)$$

where $\sqrt{-g} \mathcal{R}(g)$ is the local entropy production density, κ is a Lagrange multiplier and A is a constant which represents average \mathcal{A} . Note that the energy momentum tensor density (9.3)

does not depend on the metric and so varying the corresponding term in (9.4) with respect to the metric produces the desired result

$$\frac{\delta}{\delta g_{\alpha\beta}} \left(\int d^{D+1}X \sqrt{-g} g_{\mu\nu} T^{\mu\nu} \right) = \sqrt{-g} T^{\alpha\beta}. \quad (9.5)$$

However, if we are not following the microscopic dynamics of all of the elements of the bias vector and weight matrix, then it is more useful to define

$$\mathcal{L}_M(g, Q) \equiv -g_{\mu\nu} \langle T^{\mu\nu} \rangle_Q \quad (9.6)$$

where Q represents the trainable variables in \mathbf{q} (or equivalently in \mathbf{b} and \hat{w}) which were not averaged over. Then the action (9.4) can be written as

$$\mathcal{S}_x[g, Q] = \int d^{D+1}X \sqrt{-g} (\mathcal{R}(g) + \kappa \mathcal{L}_M(g, Q)) + \kappa A \quad (9.7)$$

where $\mathcal{L}_M(g, Q)$ plays the role of the ‘‘matter’’ Lagrangian and then the energy momentum tensor should be defined as

$$\sqrt{-g} \mathcal{T}^{\alpha\beta} \equiv -\frac{\delta}{\delta g_{\alpha\beta}} \left(\int d^{D+1}X \sqrt{-g} \mathcal{L}_M(g, Q) \right). \quad (9.8)$$

The parameter κ is a Lagrange multiplier which imposes a ‘‘global’’ constraint

$$\frac{\delta \mathcal{S}_x[g]}{\delta \kappa} = A + \int d^{D+1}X \sqrt{-g} \mathcal{L}_M(g, Q) = 0 \quad (9.9)$$

but one can also impose the constraint ‘‘locally’’ by demanding that

$$A = \frac{2}{\kappa} \int d^{D+1}X \sqrt{-g} \Lambda \quad (9.10)$$

and then the total action becomes

$$\mathcal{S}_x[g, Q] = \int d^{D+1}X \sqrt{-g} (\mathcal{R}(g) - 2\Lambda + \kappa \mathcal{L}_M(g, Q)) \quad (9.11)$$

where Λ is the ‘‘cosmological constant’’.

Recall that the deviations of the metric $g_{\mu\nu}(\mathbf{X})$ (or $g_{\mu\nu}^i$) from the flat metric $\eta_{\mu\nu}$ represent local interactions between subsystems (7.8). Therefore, if our system is in the process of equilibration, then the entropy production should be a local function of the metric tensor. Using a phenomenological approach due to Onsager [21] we can expand the entropy production around equilibrium [41],

$$\sqrt{-g} \mathcal{R} = \sqrt{-g} L^{\mu\nu \alpha\beta \gamma\delta} g_{\alpha\beta, \mu} g_{\gamma\delta, \nu}. \quad (9.12)$$

where

$$g_{\alpha\beta, \mu} \equiv \frac{\partial g_{\alpha\beta}}{\partial X^\mu} \quad (9.13)$$

and $\sqrt{-g} L^{\mu\nu \alpha\beta \gamma\delta}$ is the Onsager tensor density. The overall space of such tensors is pretty large, but it turns out that a very simple and highly symmetric choice leads to general relativity:

$$\sqrt{-g} L^{\mu\nu \alpha\beta \gamma\delta} = \frac{1}{4} \sqrt{-g} \left(2g^{\alpha\gamma} g^{\beta\nu} g^{\mu\delta} - g^{\alpha\gamma} g^{\beta\delta} g^{\mu\nu} - g^{\alpha\beta} g^{\gamma\delta} g^{\mu\nu} \right). \quad (9.14)$$

After integrating by parts, neglecting boundary terms and collecting all other terms we get

$$\begin{aligned} \int d^{D+1}X \sqrt{-g} \mathcal{R} &= \int d^{D+1}X \sqrt{-g} g^{\mu\nu} 2 \left(\Gamma^\alpha_{\nu[\mu,\alpha]} + \Gamma^\beta_{\nu[\mu} \Gamma^\alpha_{\alpha]\beta} \right) = \\ &= \int d^{D+1}X \sqrt{-g} \frac{1}{4} \left(2g^{\alpha\gamma} g^{\beta\nu} g^{\mu\delta} - g^{\alpha\gamma} g^{\beta\delta} g^{\mu\nu} - g^{\alpha\beta} g^{\gamma\delta} g^{\mu\nu} \right) g_{\alpha\beta,\mu} g_{\gamma\delta,\nu} \end{aligned} \quad (9.15)$$

where

$$\Gamma^\mu_{\gamma\delta} \equiv \frac{1}{2} g^{\mu\nu} (g_{\nu\gamma,\delta} + g_{\nu\delta,\gamma} - g_{\gamma\delta,\nu}) \quad (9.16)$$

and

$$\Gamma^\alpha_{\mu\nu,\beta} \equiv \frac{\partial}{\partial X^\beta} \Gamma^\alpha_{\mu\nu}. \quad (9.17)$$

Thus, upon varying (9.4) with respect to the metric we get the Einstein equations

$$\mathcal{R}_{\mu\nu} - \frac{1}{2} \mathcal{R} g_{\mu\nu} + \Lambda g_{\mu\nu} = \kappa \mathcal{T}_{\mu\nu} \quad (9.18)$$

where the Ricci tensor is defined as usual

$$\mathcal{R}_{\mu\nu} \equiv 2 \left(\Gamma^\alpha_{\nu[\mu,\alpha]} + \Gamma^\beta_{\nu[\mu} \Gamma^\alpha_{\alpha]\beta} \right). \quad (9.19)$$

Note that according to definition (9.14) the Onsager tensor need not be positive definite which would be inconsistent with the second law of thermodynamics, but is permitted by the second law of learning.

10 Holography

In the preceding sections we applied the principle of the stationary entropy production to study the dynamics of the neural networks in two different limits. In the first limit the trainable variables \mathbf{q} were treated stochastically, but their dynamics was constrained by the hidden variables \mathbf{x} through the free energy, F . The resulting dynamics of the system was shown to exhibit quantum and classical behaviors described by the functional $\mathcal{S}_q[p, F]$ (see (5.1)). In the second limit the hidden variables \mathbf{x} were treated stochastically, but their dynamics was constrained by the trainable variables \mathbf{q} through the action, \mathcal{A} . The resulting dynamics of the system was shown to exhibit a behavior described by the action of a gravitational metric theory, such as general relativity, $\mathcal{S}_x[g, Q]$ (see (9.11)). The two limits are certainly very different: the “gravitational” theory describes very sparse and deep neural networks and in the “quantum” theory the network can be very dense and shallow. However, one might wonder if it may possible to map the sparse and deep neural network to the dense and shallow neural network without losing the ability of the neural network to learn. If the answer is affirmative, then this would imply that the two descriptions - quantum and gravitational (or dense and sparse, or shallow and deep) - are dual and either one can be used to describe the learning dynamics.

In this section we shall explore an idea that the duality not only exists, but is also holographic in a sense that the degrees of freedom of the gravitational theory, i.e. \mathbf{x} , \mathbf{b} and \hat{w} , can be mapped to only boundary degrees of freedom of the quantum theory, i.e. \mathbf{x}^∂ , \mathbf{b}^∂ and \hat{w}^∂ . The non-equilibrium dynamics of both systems is governed by the principle of stationary entropy production and to justify such a mapping the entropy production of

the gravitational system ΔS_x should correspond to the entropy production of the quantum system ΔS_q^∂ . Roughly speaking, this means that the uncertainty in the position of neurons in the bulk, \mathbf{x} , should correspond to the uncertainty in the values of quantum variables on the boundary, i.e. \mathbf{b}^∂ and \hat{w}^∂ . For example, consider a mapping defined by

$$\mathbf{x}^\partial = \hat{P}_\partial \mathbf{x} \quad (10.1)$$

$$\mathbf{b}^\partial = \hat{P}_\partial \mathbf{b} \quad (10.2)$$

$$\hat{w}^\partial(\epsilon) = \hat{P}_\partial \frac{\epsilon \hat{w}}{\hat{I} - \epsilon \hat{w}} \hat{P}_\partial^T \quad (10.3)$$

In a microscopic picture the gravitational system consists of long chains of neurons (see Sec. 7) connecting different pairs of the boundary neurons, i and j , but the length of these chains is encoded in the elements of the boundary weight matrix,

$$d(i, j) = \log_\epsilon \left(w_{ij}^\partial(\epsilon) \right) - 1. \quad (10.4)$$

The smaller the element w_{ij}^∂ , the larger the number of intermediate bulk neurons connecting i to j . Whenever any two chains of neurons i - j and k - l have a chance of intersecting and forming two other chains of neurons i - l and k - j , the entropy of the bulk theory changes. On the other side of the duality, the same event can lead to the corresponding elements w_{ij}^∂ , w_{kl}^∂ , w_{kj}^∂ and w_{il}^∂ to change or, in other words, to the entropy production in the boundary theory. Thus, it is not too unreasonable to expect that the entropy production in both system are related.

The holographic duality can be formulated more precisely by considering the action functionals which determine the dynamics in both theories. In the boundary theory the action $\mathcal{S}_q [p(\mathbf{q}^\partial), F(\mathbf{q}^\partial)]$ is given by equation (5.1) and in the bulk theory the action $\mathcal{S}_x [g(\mathbf{X}), Q(\mathbf{X})]$ is given by equation (9.11). For the two systems to be dual the two actions must be proportional

$$\mathcal{S}_x [g(\mathbf{X}), Q(\mathbf{X})] \sim \mathcal{S}_q [p(\mathbf{q}^\partial), F(\mathbf{q}^\partial)], \quad (10.5)$$

or, using (5.1) and (9.11),

$$\int d^{D+1} X \sqrt{-g} \quad (\mathcal{R}(g) - 2\Lambda + \kappa \mathcal{L}_M(g, Q)) \quad (10.6)$$

$$\sim \int dt d^{K^\partial} q^\partial \sqrt{p} \left(-4D \frac{\partial^2}{\partial q_k^2} + \gamma \left(\frac{\partial}{\partial q_k^\partial} \right)^2 F + \mu \frac{\partial F}{\partial t} + \mu \gamma \left(\frac{\partial}{\partial q_k^\partial} \right)^2 F + \mu V \right) \sqrt{p}.$$

The left hand side describes the bulk gravitational theory, the right hand side describes the boundary theory and the duality transformation is nothing but changes of variables between (g, Q) and (p, F) . Note, however, that the boundary theory can be approximated by quantum mechanics only in the limit when the entropy production due to learning (i.e. the quantity in the box in (10.6)) is subdominant. Therefore the holography described by (10.5) should be considered as more general than the holography discussed, for example, in context of the AdS/CFT correspondence where the CFT side is quantum and the AdS side is gravitational.

11 Discussion

In this paper we discussed a possibility that the entire universe on its most fundamental level is a neural network. This is a very bold claim. We are not just saying that the artificial neural

networks can be useful for analyzing physical systems [22] or for discovering physical laws [23], we are saying that this is how the world around us actually works. With this respect it could be considered as a proposal for the theory of everything, and as such it should be easy to prove it wrong. All that is needed is to find a physical phenomenon which cannot be described by neural networks. Unfortunately (or fortunately) it is easier said than done. It turns out that the dynamics of neural networks is so complex that one can only understand it in very specific limits. The main objective of this paper was to describe the behavior of the neural networks in the limits when the relevant degrees of freedom (such as bias vector, weight matrix, state vector of neurons) can be modeled as stochastic variables which undergo a learning evolution. In this section we shall briefly discuss the main results and implications of the results for a possible emergence of quantum mechanics, general relativity and macroscopic observers from a microscopic neural network.

Emergent quantum mechanics is a relatively new [24, 25], but rapidly evolving field [26–31] which is based on a set of very old ideas, dating back to the works of de Broglie and Bohm. The de Broglie-Bohm theory (also known as pilot wave theory or Bohmian mechanics) was originally formulated in terms of non-local hidden variables [12] which makes it an easy target. The main new insight is that quantum mechanics may not be a fundamental theory, but only a mathematical tool which allows us to carry out statistical calculations in certain dynamical systems. If correct, then one should be able to derive all of the essential ingredients (complex wave-function, Schrödinger equation, etc.) from first principle. In this paper we did exactly that for a dynamical system of a neural network which contains two different types of degrees of freedom: trainable (e.g. bias vector and weight matrix) and hidden (e.g. state vector of neurons). What we showed is that the dynamics of the trainable variables near equilibrium is described by Madelung (or equivalently Schrödinger) equations with free energy (for a canonical ensemble of hidden variables) representing the quantum phase (see Sec. 5), and further away from the equilibrium their dynamics is described by Hamilton-Jacobi equations with free energy representing the Hamilton’s principal function (see Sec. 6). This demonstrates that the neural networks can indeed exhibit emergent quantum and also classical behaviors. It is important to emphasize that the learning dynamics was essential and the stochastic dynamics alone would not have produced the desired result.

Emergent (or entropic) gravity is also a relatively new field [7–9], but it is far less clear if or when progress is being made. The main problem is that emergent gravity is not just about gravity, but is also about emergent space [32–35], emergent Lorentz invariance [36–38], emergent general relativity [39–41] etc. Quite remarkably, neural networks open up a new avenue to address all these problems in context of the learning dynamics. It turns out that a dynamical space-time can indeed emerge from a non-equilibrium evolution of the hidden variables (i.e. state vector of neurons) in a manner very similar to string theory. In particular, if one considers D minimally-interacting (through bias vector and weight matrix) subsystems with average state vectors, $\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^D$ (and the total average state vector $\bar{\mathbf{x}}^0$) then the dynamics of $\bar{\mathbf{x}}^\mu$ can be modeled with relativistic strings in an emergent $D + 1$ dimensional space-time (see Secs. 7 and 8) and if the interactions are described by a metric tensor, then the dynamics can be modeled with Einstein equations (see Sec. 9). Once again, not only stochastic, but also learning dynamics was essential for the equilibration of the emergent space-time to exhibit behavior of a gravitational theory such as general relativity. This demonstrates that the dynamics of a neural network in the appropriate limits can be approximated by both emergent quantum mechanics and emergent general relativity, but the two limits are very different. The gravitational theory describes very sparse and deep neural

networks and in the quantum theory the neural network can be very dense and shallow. However, it is possible that there exists a holographic duality map between the bulk neurons of the deep and sparse network to the boundary neurons of the shallow and dense network (see Sec. 10).

We now come to one of the most controversial questions: how can macroscopic observers emerge in a physical system? The question is extremely important not only for settling some philosophical debates, but for understanding the results of real physical experiments [12] and cosmological observations [13]. As was already mentioned, our current understanding of fundamental physics does not allow us to formulate a self-consistent and paradoxes-free definition of observers and a possibility that observers is an emergent phenomenon is certainly worth considering. Indeed, if both quantum mechanics and general relativity are not fundamental, but emergent phenomena, then why cannot macroscopic observers also emerge in some way from a microscopic neural network. Of course this is a lot more difficult task and we are not going to resolve it completely, but we shall mention an old idea that might be relevant here. It is the principle of natural selection. We are not talking about cosmological natural selection [42], but about the good old biological natural selection [43], although the two might actually be related. Indeed, if the entire universe is a neural network, then something like natural selection might be happening on all scales from cosmological ($> 10^{+15}$ m) and biological ($10^{+2} - 10^{-6}$ m) all the way to subatomic ($< 10^{-15}$ m) scales. The main idea is that some local structures (or architectures) of neural networks are more stable against external perturbations (i.e. interactions with the rest of the network) than other local structures. As a result the more stable structures are more likely to survive and the less stable structures are more likely to be exterminated. There is no reason to expect that this process might stop at a fixed time or might be confined to a fixed scale and so the evolution must continue indefinitely and on all scales. We have already seen that on the smallest scales the learning evolution is likely to produce structures of a very low complexity (i.e. second law of learning) such as one dimensional chains of neurons, but this might just be the beginning. As the learning progresses these chains can chop off loops, form junctions and according to natural selection the more stable structures would survive. If correct, then what we now call atoms and particles might actually be the outcomes of a long evolution starting from some very low complexity structures and what we now call macroscopic observers and biological cells might be the outcome of an even longer evolution. Of course, at present the claim that natural selection may be relevant on all scales is very speculative, but it seems that neural networks do offer an interesting new perspective on the problem of observers.

Acknowledgments. This work was supported in part by the Foundational Questions Institute (FQXi).

References

- [1] E. Witten, “Anti-de Sitter space and holography,” *Adv. Theor. Math. Phys.* **2**, 253 (1998)
- [2] L. Susskind, “The World as a hologram,” *J. Math. Phys.* **36**, 6377 (1995)
- [3] J. M. Maldacena, “The Large N limit of superconformal field theories and supergravity,” *Int. J. Theor. Phys.* **38**, 1113 (1999)
- [4] A. Ashtekar, “New Variables for Classical and Quantum Gravity,” *Phys. Rev. Lett.* **57** (1986) 2244-2247.
- [5] C. Rovelli and L. Smolin, “Loop Space Representation of Quantum General Relativity,” *Nucl. Phys. B* **331** (1990) 80.

- [6] A. Ashtekar, M. Bojowald and J. Lewandowski, “Mathematical structure of loop quantum cosmology,” *Adv. Theor. Math. Phys.* **7**, no.2, 233-268 (2003)
- [7] T. Jacobson, “Thermodynamics of space-time: The Einstein equation of state,” *Phys. Rev. Lett.* **75**, 1260 (1995)
- [8] Padmanabhan T. “Thermodynamical Aspects of Gravity: New insights,” *Reports on Progress in Physics.* 73 (4): 046901 (2010)
- [9] E. P. Verlinde, “On the Origin of Gravity and the Laws of Newton,” *JHEP* **1104**, 029 (2011)
- [10] H. Everett, “Relative State Formulation of Quantum Mechanics,” *Reviews of Modern Physics.* 29 (3): 454-462, (1957)
- [11] D. Bohm, “A Suggested Interpretation of the Quantum Theory in Terms of ‘Hidden Variables’ I,” *Physical Review.* 85 (2): 166-179, (1952)
- [12] J. Bell, “On the Einstein Podolsky Rosen Paradox,” *Physics.* 1 (3): 195-200, (1964)
- [13] V. Vanchurin, A. Vilenkin and S. Winitzki, “Predictability crisis in inflationary cosmology and its resolution,” *Phys. Rev. D* **61**, 083507 (2000)
- [14] G. Dvali, “Black Holes as Brains: Neural Networks with Area Law Entropy,” *Fortsch. Phys.* **66**, no. 4, 1800007 (2018)
- [15] K. Hashimoto, S. Sugishita, A. Tanaka and A. Tomiya, “Deep learning and the AdS/CFT correspondence,” *Phys. Rev. D* **98**, no.4, 046019 (2018)
- [16] V. Vanchurin, “Towards a theory of machine learning,” [arXiv:2004.09280 [cs.LG]]
- [17] E. T. Jaynes, “Information Theory and Statistical Mechanics,” *Physical Review. Series II.* 106 (4): 620-630, (1957)
- [18] E. T. Jaynes, “Information Theory and Statistical Mechanics II,” *Physical Review. Series II.* 108 (2): 171-190, (1957)
- [19] Prigogine, I. “Etude Thermodynamique des phénomènes irréversibles”. Desoer, Liège, (1947)
- [20] M. J. Klein, P. H. E. Meijer, “Principle of minimum entropy production.” *Phys. Rev.* 96: 250-255, (1954)
- [21] Onsager, L. “Reciprocal relations in irreversible processes, I”. *Physical Review.* 37 (4) 405-426 (1931)
- [22] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborov, “Machine learning and the physical sciences,” *Rev. Mod. Phys.* **91**, no.4, 045002 (2019)
- [23] Tailin Wu and Max Tegmark, “Toward an artificial intelligence physicist for unsupervised learning,” *Physical Review E*, 100(3):033311, (2019)
- [24] S. Adler, “Quantum Theory as an Emergent Phenomenon” (Cambridge UP, Cambridge, 2004)
- [25] G. 't Hooft, “Emergent Quantum Mechanics and Emergent Symmetries,” *AIP Conf. Proc.* **957**, no.1, 154-163 (2007)
- [26] M. Blasone, P. Jizba and F. Scardigli, “Can quantum mechanics be an emergent phenomenon?,” *J. Phys. Conf. Ser.* **174**, 012034 (2009)
- [27] Grossing, G.; Fussy, S.; Mesa Pascasio, J.; Schwabl, H. “The Quantum as an Emergent System,” *Journal of Physics: Conference Series*, Volume 361, Issue 1, article id. 012008, 15 pp. (2012).
- [28] D. Acosta, P. F. de Cordoba, J. M. Isidro and J. L. G. Santander, “Emergent quantum mechanics as a classical, irreversible thermodynamics,” *Int. J. Geom. Meth. Mod. Phys.* **10**, no.4, 1350007 (2013)

- [29] P. Fernández De Córdoba, J. M. Isidro and M. H. Perea, “Emergent quantum mechanics as a thermal ensemble,” *Int. J. Geom. Meth. Mod. Phys.* **11**, no.08, 1450068 (2014)
- [30] A. Caticha, “Entropic Dynamics: Quantum Mechanics from Entropy and Information Geometry,” *Annalen Phys.* **531**, no.3, 1700408 (2019)
- [31] V. Vanchurin, “Entropic Mechanics: towards a stochastic description of quantum mechanics,” *Found. Phys.* **50**, no. 1, 40 (2019)
- [32] B. Swingle, “Entanglement Renormalization and Holography,” *Phys. Rev. D* **86**, 065007 (2012)
- [33] A. Almheiri, X. Dong and D. Harlow, “Bulk Locality and Quantum Error Correction in AdS/CFT,” *JHEP* **1504**, 163 (2015)
- [34] C. Cao, S. M. Carroll and S. Michalakis, “Space from Hilbert Space: Recovering Geometry from Bulk Entanglement?,” *Phys. Rev. D* **95**, 024031 (2017);
- [35] V. Vanchurin, “Information Graph Flow: a geometric approximation of quantum and statistical systems,” *Found. Phys.* **48**, no. 6, 636 (2018)
- [36] R. B. Laughlin, “Emergent relativity,” *Int. J. Mod. Phys. A* **18**, 831-854 (2003)
- [37] G. Bednik, O. Pujols and S. Sibiryakov, “Emergent Lorentz invariance from Strong Dynamics: Holographic examples,” *JHEP* **11**, 064 (2013)
- [38] V. Vanchurin, “A quantum-classical duality and emergent space-time,” 10th Mathematical Physics Meeting, 347-366, [arXiv:1903.06083 [hep-th]]; V. Vanchurin, “Differential equation for partition functions and a duality pseudo-forest,” [arXiv:1910.11268 [hep-th]]; V. Vanchurin, “Dual Path Integral: a non-perturbative approach to strong coupling,” [arXiv:1912.09265 [hep-th]].
- [39] C. Barcelo, M. Visser and S. Liberati, “Einstein gravity as an emergent phenomenon?,” *Int. J. Mod. Phys. D* **10**, 799-806 (2001)
- [40] C. Cao and S. M. Carroll, “Bulk entanglement gravity without a boundary: Towards finding Einstein’s equation in Hilbert space,” *Phys. Rev. D* **97**, no.8, 086003 (2018)
- [41] V. Vanchurin, “Covariant Information Theory and Emergent Gravity,” *Int. J. Mod. Phys. A* **33**, no. 34, 1845019 (2018)
- [42] L. Smolin, “Did the Universe Evolve?” *Classical and Quantum Gravity* **9**:173-191 (1992)
- [43] C. Darwin, “On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life.” (1859)