PAPER • **OPEN ACCESS**

A duality connecting neural network and cosmological dynamics

To cite this article: Sven Krippendorf and Michael Spannowsky 2022 *Mach. Learn.: Sci. Technol.* **3** 035011

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 035011 https://doi.org/10.1088/2632-2153/ac87e9

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
4 May 2022

REVISED
28 July 2022

ACCEPTED FOR PUBLICATION
8 August 2022

PUBLISHED
30 August 2022

PAPER

A duality connecting neural network and cosmological dynamics

Sven Krippendorf[1,*] **and Michael Spannowsky**[2,3]

[1] Arnold Sommerfeld Center for Theoretical Physics, Ludwig-Maximilians-Universität, Theresienstr. 37, 80333 Munich, Germany
[2] Department of Physics, Institute for Particle Physics Phenomenology, Durham University, South Road, Durham DH1 3LE, United Kingdom
[3] Department of Physics, Durham University, Durham DH1 3LE, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: sven.krippendorf@physik.uni-muenchen.de

Keywords: neural tangent kernels of neural networks, neural network dynamics, cosmological dynamics with scalar fields

Abstract

We demonstrate that the dynamics of neural networks (NNs) trained with gradient descent and the dynamics of scalar fields in a flat, vacuum energy dominated Universe are structurally profoundly related. This duality provides the framework for synergies between these systems, to understand and explain NN dynamics and new ways of simulating and describing early Universe models. Working in the continuous-time limit of NNs, we analytically match the dynamics of the mean background and the dynamics of small perturbations around the mean field, highlighting potential differences in separate limits. We perform empirical tests of this analytic description and quantitatively show the dependence of the effective field theory parameters on hyperparameters of the NN. As a result of this duality, the cosmological constant is matched inversely to the learning rate in the gradient descent update.

1. Introduction

Effective field theories (EFT) are one of the most powerful paradigms of contemporary physics, which can capture the dynamics in virtually all science areas, thereby covering the vast scales relevant in cosmology, the intermediate scales relevant for living organisms, and the very small scales of particle physics. One aspect of the astounding power of EFT is that they do not require a detailed understanding of the microphysical processes at an elementary level to describe the dynamics in question. Instead, effective degrees of freedom and interactions are being utilised to describe physical processes at the relevant scales.

Here, we propose and exemplify how to identify the EFT associated with neural network (NN) training dynamics. The network dynamics are characterised by how the NN output changes during training, i.e. how the network weights are updated. Deep neural networks (DNN) have entered a revolutionary phase, where their near-term prospect foresees the deployment of sophisticated methods to safety-critical applications, e.g. medical applications, autonomous driving, or financial investments. DNN are often described as black boxes whose features are not explainable, i.e. it is generally not understood what state NNs strive towards after and during training. Because of their intrinsic high degree of non-linearity, the ubiquitous deep learning algorithms, which are at the forefront of the machine learning (ML) developments in most areas, still lack transparency and explainability. Thus, being able to improve our understanding of the intrinsic dynamics of NNs during training is of particular importance for the future development and deployment of DNNs in addressing real challenges.

In general, a DNN starts from some initial functional representation which changes guided by the loss function. Following the gradient-based optimisation algorithm, aiming to minimise the loss function, it evolves to contain non-linear features which are well-suited for classification or regression tasks. To avoid getting stuck in local minima, gradient descent methods are often augmented. One such method is to modify the weight update by a momentum term [1], an aggregate of gradients, which becomes the exponential moving average of the history of gradients. The parameters of a network are updated regularly, after a discrete number of steps, measured in batches or epochs. When taking the continuous limit for the network's

learning-rate equation in the presence of a momentum term, one finds a linear second-order differential equation in the time component. Limiting the network output to a single value, its structure equates a scalar field, defined over the continuous, potentially multi-dimensional, input space, very similar to the equations that govern the dynamics of the early universe. Thus, it is plausible that such scalar field equations, can describe the training of a sufficiently wide NN.

It is therefore intriguing to compare the training of a NN with the evolution of a scalar field in the Friedmann–Lemaître–Robertson–Walker (FLRW) background, which is a solution of Einstein's field equations of general relativity. The relatively simple form of these scalar field equations are a direct consequence of the assumption that the Universe is homogeneous and isotropic. However, non-linear features can develop from random fluctuations around a homogeneous time-evolving background (see [2–4] for a concise overview). The time evolution of the scalar background field depends on the gravitational theory underlying the dynamics, which includes in the FLRW case of general relativity whether other energy components dominate the Universe. This is naturally analogous to choices for the optimisation algorithm, most prominently the choice of the learning rate used for the gradient-based update.

In this paper, we establish this duality between previously unrelated dynamical systems by demonstrating that the evolution of the NN follows very comparable dynamics of a scalar field in an FLRW-background, thereby resembling the dynamics in the early Universe. Relating both dynamical systems offers the potential for significant synergies drawing from knowledge on the respective process: on the one side, this includes the fact that the formation of structure in our Universe can be modelled very accurately in the EFT framework of ΛCDM and in similar ways extensions of this standard model in the EFT language are utilised to describe potential additional features which can be tested observationally, a prime example being the plethora of dark matter models which range from ultralight axion-like particles to primordial black holes [5]. On the other hand, the functional dependence of the EFT on hyperparameters of the NN and the optimiser allow the optimisation of such parameters without having to train multiple networks. This duality also relates scales of different parameters (i.e. optimisation, training set size, and NN architecture) to one dynamical/analytical framework.

This new EFT perspective, in comparison to previous work on approximate NN dynamics, in particular the neural tangent kernel approach (NTK) [6–8], allows us to understand and utilise the hierarchies in the dynamics appearing in the NTK formulation. This is to say that we can identify a mode dominating the dynamics, the mean field of the NN, and that the evolution of the other modes is heavily influenced by the evolution of the mean field. Following the EFT framework, we also identify modes which effectively are not evolving and which can be neglected without loosing accuracy in training. This EFT approach allows for a systematic investigation of whether all types of dynamics which can appear in the early Universe have already been constructed in the context of NNs.

This duality between field theory dynamics and NNs updated with gradient descent also offers a new perspective on describing the observed energy distribution in our Universe. For instance, we find a close relationship between the learning rate and the vacuum energy. Further, as training a NN with gradient descent is reasonably cheap, it is a compelling question whether our duality between these two systems leads to more efficient simulation algorithms than *standard* discretisation of continuous dynamics (e.g. in lattice simulations). Put differently, it is interesting to see which phenomenological EFT models share dynamics with NNs. In addition, this NN perspective necessitates the presence of gravitational dynamics in any field theoretical evolution given sufficient excursions in field theory space where gravitational dynamics cannot be decoupled.

To demonstrate a duality between both dynamical systems, we match the evolutionary dynamics in the continuous-time limit in section 2. We perform this matching explicitly for the simple case of a NN with a single output dimension related to a single scalar field analytically. The scalar field is exposed to a scalar potential associated with the NN architecture and the loss function. Moreover, the optimisation algorithm is related to the gravitational dynamics of the scalar field. At this stage, we require that the empirical NTK be a valid approximation of the NN dynamics during training. To make this match with cosmological dynamics, we perform a basis change that diagonalises the empirical NTK. We find that the dynamics of the modes are of the same structure with the difference that the NN modes evolution can be suppressed in the relevant evolution term by the smaller eigenvalues appearing in the NTK, which depends on the choice of the NN architecture and the training data. In section 3, we empirically test this duality by training homogeneous NNs to match the dynamics of a homogeneous Universe and then standard input dependent NNs to match both the dynamics of the homogeneous Universe and the evolution of the perturbation. We utilise standard fully-connected NNs, which are optimised using gradient descent with and without momentum subject to a polynomial loss function resembling of mean squared error and, more importantly, those of standard polynomial inflationary potentials [9]. In section 4 we connect our duality with previous directions of work in both ML and physics before concluding in section 5.

2. Analytically matching dynamics

Here we match the dynamics of a scalar field in an FLRW background and the continuous-time dynamics of NNs, which are updated via gradient descent with momentum, closely following the notation in [7][4]. To make the connection, we start with the homogeneous and isotropic evolution of a perfect fluid, i.e. a spatially homogeneous and isotropic scalar field. The most straightforward realisation is with networks that are homogenous and isotropic with respect to the input. We then relax this homogeneity and isotropy assumption and match the dynamics for small perturbations around such a homogeneous and isotropic background.

In our comparison, we must study the NN dynamics on the function space level rather than the 'microscopic' NN parameter level. This perspective is in contrast to parameter-based approaches, which share a similar aim to explain the NN dynamics and their learning behaviour [12].

2.1. Matching homogeneous and isotropic dynamics

We are interested in describing the dynamics of a NN $f(t,x,\theta)$ subject to gradient descent:

$$\begin{aligned} f\colon\ \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m &\to \mathbb{R} \\ (t,x,\theta) &\mapsto f(t,x,\theta)\,. \end{aligned} \tag{1}$$

Gradient descent with momentum yields a continuous time differential equation for the update of the NN which is of second order. The discrete update equation of weights is:

$$\theta_{i+1} = \theta_i + \beta(\theta_i - \theta_{i-1}) - \eta\,\nabla_\theta \mathcal{L}(\mathcal{D})|_{\theta=\theta_i}\,, \tag{2}$$

where β denotes the momentum parameter, and η the learning rate. $\mathcal{L}$ is the loss function which we consider to be of the following type:

$$\begin{aligned} \mathcal{L} : \mathbb{R}\times\mathbb{R} &\to \mathbb{R} \\ (y, y_{\text{target}}) &\mapsto \mathcal{L}(y, y_{\text{target}})\,, \end{aligned} \tag{3}$$

where we average this expression when it is evaluated over a (training) dataset $\mathcal{D}$ with the length of the dataset $N = |\mathcal{D}|$. Its continuous version is:

$$\ddot{\theta} = \tilde{\beta}\dot{\theta} - \nabla_\theta \mathcal{L}(\mathcal{D})\,, \tag{4}$$

where $\tilde{\beta} = (\beta - 1)/\sqrt{\eta}$ and the continuous and discrete time are related via $t = i\sqrt{\eta}$. The update equation for the NN output is obtained by utilising the chain rule $\dot{f} = \dot{\theta}\nabla_\theta f$, and becomes:

$$\ddot{f}(x) = \tilde{\beta}\dot{f}(x) - \nabla_\theta f(x)\nabla_\theta f(\mathcal{D})\nabla_f \mathcal{L}(\mathcal{D})\,, \tag{5}$$

which can be evaluated locally for each point x of the input space. The (non-local) data used for updating the weights $\mathcal{D}$ has to be applied when calculating $\dot{\theta}$.

Before focusing on specific limits in the NN dynamics, let us briefly discuss the scalar dynamics in a homogeneous and isotropic Universe, described by:

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0\,,\ \text{ where } 3H^2 = \frac{\dot{\phi}^2}{2} + V(\phi)\,, \tag{6}$$

where $\phi(x,t) = \phi(t)$ denotes the homogeneous and isotropic scalar field and $V(\phi)$ its scalar potential. H is the Hubble parameter which is the ratio of the change of the metric's scale factor and the scale factor itself. Einstein's equations set the Hubble parameter in the case of our scalar field as in (6).

To achieve a matching between the differential equations in (5) and (6), we specialise first to the homogeneous and isotropic case for the input (feature) space, i.e. $f(x,t) = f(t)$ and then discuss the general case. In this case, the prefactor is defined by $\alpha \equiv \nabla_\theta f(x,t)\nabla_\theta f(\mathcal{D},t)$. In the NTK limit, i.e. the limit where the network is infinitely wide, the prefactor is constrained to its value at the beginning of training, $\alpha(t) = \alpha(t=0)$. Thus, one obtains a constant α. In this limit the matching becomes straightforward as the NN dynamics simplify to:

$$\ddot{f} - \tilde{\beta}\dot{f} + V'_{\text{eff}}(f) = 0\,, \tag{7}$$

[4] Previous work on continuous-time updates include [10, 11].

where the potential V_{eff} is specified in due course.

The second approximation which we make – this time on the FLRW side of the duality – is that we assume H to be dominated by the vacuum energy, a parameter which is eliminated from $V'(\phi)$. Given these approximations, an effective potential of the following form matches the dynamics of scalar fields in a homogeneous and isotropic Universe with those of a NN:

$$V_{\text{eff}} = \alpha\mathcal{L} + V_0 \,, \text{ where } V_0 = \frac{\tilde{\beta}^2}{3} \,, \tag{8}$$

where the vacuum energy V_0 is added to ensure a matching of the two respective friction terms. Thus, we find that the network dynamics resemble the dynamics of a vacuum energy dominated Universe, i.e. $3H^2 \approx V_0$.

It is rather remarkable that the vacuum energy has the interpretation of the combination of momentum parameter and learning rate in the NN dynamics. To our knowledge, this connection is new. It is also remarkable that in this picture, it is clear that a change in the NN architecture can change α, which in turn determines the training dynamics, e.g. whether the training dynamics are vacuum energy dominated or not, and this direct matching applies.

We also note that a scalar field in our Universe can change the optimiser parameters dynamically. A systematic investigation of the exact matching between optimiser algorithms and standard phases of cosmological evolution, e.g. domination of the kinetic terms, is left for future work. We note that a parametrically small cosmological constant in this dual picture corresponds to a parametrically large learning rate in the optimisation algorithm. This matching of optimiser parameters and the vacuum energy matches respective 'global', i.e. non-local, quantities.

Finally, let us comment on the scales which have been set in this duality. In particular, the Hubble scale is relevant for the time evolution and the appropriate scale in the spacetime metric. The duality relates to the previously unrelated scales associated with the input data and the time steps. In particular, the Hubble time is $1/H$, and it corresponds to the only time-scale in the NN. This is in contrast to our Universe where additional dynamics are taking place beyond the ones related to *training*.

2.2. Matching perturbations

In early Universe physics, quantum fluctuations around the classical background can grow large and are responsible, in the standard cosmological picture, for structure formation. The equations for small perturbations $\delta\phi \ll 1$ can be obtained by considering the entire equations of motion with the ansatz $\phi + \delta\phi$. In NNs, inhomogeneities are essential for successfully addressing tasks, e.g. the classification of images. To establish the similarities between both dynamics, we focus on NNs which have small perturbations where we specify the appropriate size in due course.

2.2.1. NN perturbations

We start with the continuous time version of the dynamics:

$$\ddot{f}(t,x) - \tilde{\beta}\dot{f}(t,x) + \Theta(t,x,X)\nabla_{f(X)}\mathcal{L} = 0 \,, \tag{9}$$

where $\Theta(t,x,X)$ denotes the empirical neural tangent kernel at a given time t during the learning process[5]. In the classical limit we focused on a constant spatially homogeneous and isotropic value for the empirical neural tangent kernel $\Theta(X,X) \approx \alpha$. Now, we are interested in taking into account the deviation from this approximation $f(t,x) = \bar{f}(t) + \delta f(t,x)$.

To derive the equations, let us write the last term in (9) more explicitly:

$$\begin{aligned}\Theta(t,x,X)\nabla_{f(X)}\mathcal{L} &= \sum_{y\in X}\Theta(t,x,y)\nabla_{f(y)}\mathcal{L} \\ &= \sum_{y\in X}\Theta(t,x,y)\mathcal{L}'(f(y)) \\ &= \sum_{y\in X}\Theta(t,x,y)\left[\mathcal{L}'(\bar{f}) + \mathcal{L}''(\bar{f})\,\delta f(y)\right] \,,\end{aligned} \tag{10}$$

where we have assumed in the first line that our loss function can be expressed as a sum over individual data points, i.e. point-wise local contributions over the input space. The second line is a simple rewriting and the third line is a linear expansion in δf, which is exact for a quadratic loss function. In general, the

[5] The linearisation in the NTK limit takes this to be evaluated at the beginning of training.

approximation in (10) remains valid for small perturbations δf where the deviation from the exact NTK result provides a first criterion on when perturbations can be considered small.

To make contact with the previous mean-field evolution, we perform a basis transformation to diagonalise the neural tangent kernel, decoupling the modes' evolution associated with the respective eigenvalues. In particular we find in our experiments the largest eigenvalue corresponds to the mean field (i.e. the uniform evolution) α from before:

$$\begin{aligned}\Theta(t,x,X)\nabla_{f(X)}\mathcal{L} &= \Theta_{ij}\mathcal{L}'_j \\ &\to A_{ik}\Theta_{kl}(B^T)_{lj}A_{jm}\mathcal{L}'_m\,,\end{aligned} \tag{11}$$

where A and B are matrices which can be used to diagonalise $\Theta(t=0,X,X)$, $B^TA=\mathbf{1}$, and $\Theta(x_i,x_j)=\Theta_{ij}$. We construct A and B from the normalised eigenvectors v_i of Θ_{ij} as:

$$A=\begin{pmatrix} v_1^T \\ v_2^T \\ \dots \\ v_N^T \end{pmatrix}\,,\; B=\begin{pmatrix} v_1^T \\ v_2^T \\ \dots \\ v_N^T \end{pmatrix}\,,\; v_i^T\cdot v_j=\delta_{ij}\,, \tag{12}$$

$$A\,\Theta\,B^T=\mathrm{diag}\,(\lambda_1,\dots,\lambda_N)\,. \tag{13}$$

Using this basis transformation we can now define what is meant with perturbations around the mean field, namely we utilise:

$$f_i=\sqrt{N}v_1\bar{f}+\delta f_i\,,\text{and}\;\mathcal{L}_i=\frac{m^2}{N}(f_i-f_0)\,, \tag{14}$$

where we used a quadratic loss function with the minimum at f_0 for illustrative purposes.

This basis change and this definition of perturbations not only decouples the dynamics of modes but also singles out the mean field evolution. Introducing $\delta\tilde{f}_i=v_i^T\cdot\delta\mathbf{f}$, this gives the following equations for the mean and the perturbation evolution:

$$0=\ddot{\bar{f}}+\tilde{\beta}\,\dot{\bar{f}}+\frac{\lambda_1}{N}\,\mathcal{L}'(\bar{f})\,, \tag{15}$$

$$0=\delta\ddot{\tilde{f}}_i+\tilde{\beta}\,\delta\dot{\tilde{f}}_i+\frac{\lambda_i}{N}\,\delta\tilde{f}_i\,\mathcal{L}''(\bar{f})\,. \tag{16}$$

More details on the derivation can be found in appendix A.

2.2.2. Scalar field perturbations

Let us now compare the above equations with the general equations of motion for a scalar field in a FLRW background which read:

$$\ddot{\phi}+\nabla^2\phi+3H\dot{\phi}+V'(\phi)=0\,. \tag{17}$$

In the mentioned scenario, where the scalar field can be expanded into a universal background and a small perturbation $\phi(t,x)=\phi(t)+\varphi(t,x)$ equation (17) can be separated into:

$$\ddot{\phi}+3H\dot{\phi}+V'(\phi)=0\,, \tag{18}$$

$$\ddot{\varphi}+\nabla^2\varphi+3H\dot{\varphi}+V''(\phi)\varphi=0\,, \tag{19}$$

where in the last equation we are working to linear order in the perturbations ($V'(\phi+\varphi)\approx V'(\phi)+\varphi V''(\phi)$). Equation (18) corresponds to the homogeneous and isotropic background field equation whose dynamics we have already matched. Instead, (19) is identical to (16) when neglecting the spatial gradient term which is a valid approximation of long wavelength/small momentum perturbations and assuming $\lambda_i=\lambda_1$.

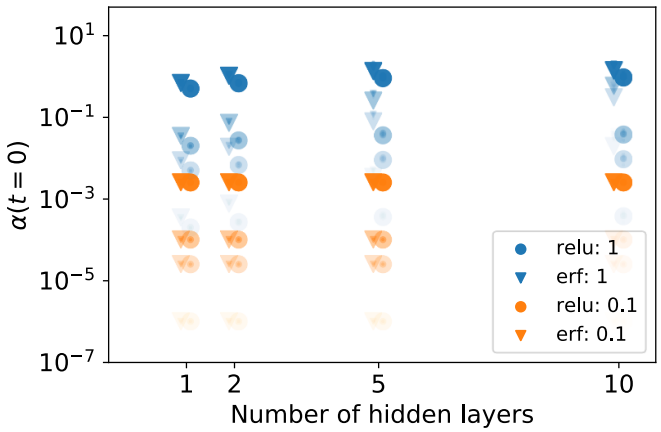

Figure 1. The values of α at initialisation in our hyperparameter scan. The weight initialisations are 1 (blue) and 0.1 (orange). We have varied the width of the hidden dense layers among the values $\{100, 1000, 10000\}$ which is shown with increasing sizes of the marker. Different bias initialisations are taken from $\{0.001, 0.005, 0.01, 0.05, 0.5\}$ with increased visibility for larger initialisation. As indicated, we utilised *erf* and *relu* activation functions and the number of dense layers.

3. Quantitatively matching dynamics

We now turn to experimentally test our analytic framework from the previous section by confronting the actual NN dynamics with our predicted dynamics. In this context, the aim of our experiments is two-fold:

(a) We would like to assess how accurate our approximations for the mean field and small perturbations around it are, i.e. when are equations (7), (15), and (16) capturing the NN dynamics.
(b) We are interested in quantitatively determining the dependence of the effective field theory potential on the NN parameters, i.e. determining α and $\tilde{\alpha}$ empirically.

Our experiments are greatly facilitated by the empirical implementation of the NTK [8].

For our experiments we choose a simple polynomial loss function which features a minimum away from the initial values of the NN,

$$\mathcal{L}(f) = \frac{m^2}{2}(f - f_0)^n\,, \tag{20}$$

where n is even and we focus on the cases $n = \{2, 4\}$, vary $f_0 = \{1, 10\}$ and scale the potential as $m = \{0.01, 0.1, 0.2\}$. Besides the simplicity on the NN side, this is to match with standard field theory potentials, where $n = 2$ corresponds to the case of chaotic inflation [9]. For our NN, we use fully connected NNs with different widths of the hidden units, several hidden units, activation functions (*relu* and *erf*), and the standard deviation of the normal distributions used for the weight initialisations of our dense layers. To study homogeneous and isotropic NNs, i.e. $f(x, t) = f(t)$, we simply evaluate our networks at a single point. In this case the empirical neural tangent kernel is $\Theta(t, x, y) = \alpha$.

Figure 1 shows the dependence of α as a function of network parameters. The largest effect results from changing the initialisation distributions, followed by smaller differences for the activation function and the number of hidden layers. We find that *erf* for activation function leads to larger values than *relu*. In contrast, the width has only a smaller effect.

To match the dynamics throughout the entire training, we need to satisfy $\alpha\mathcal{L}(f) \ll V_0$. Given a choice for the NN and the loss function parameters m and f_0, this condition implies a maximally allowed learning rate which is reduced when adding non-vanishing momentum. Figure 2 shows this dependence and indicates that for an extensive range of parameters, generic values of the learning rates are allowed. This condition relates the learning rate with the scale of the effective potential. In turn, this implies, for instance, that an architecture choice with a larger α requires a lower learning rate to keep this approximation of vacuum energy domination valid.

Having chosen hyperparameters, we train our networks explicitly and compare them with the predicted dynamics. Figure 3 shows the deviation of the predicted dynamics to the actual dynamics at the end of training (i.e. after 1000 epochs) in comparison to the difference of the NN output at the beginning and the

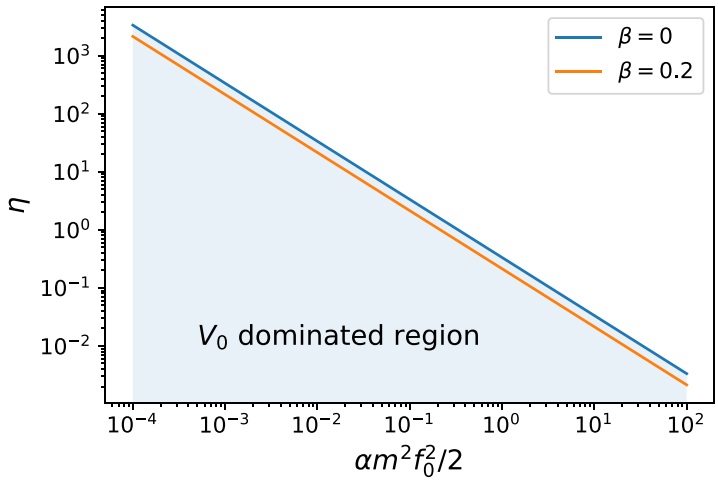

Figure 2. Region of η where V_0 is larger than the contribution from the other potential terms. When including momentum the allowed region shrinks. The line is set by $V_0 = \alpha\mathcal{L}$.

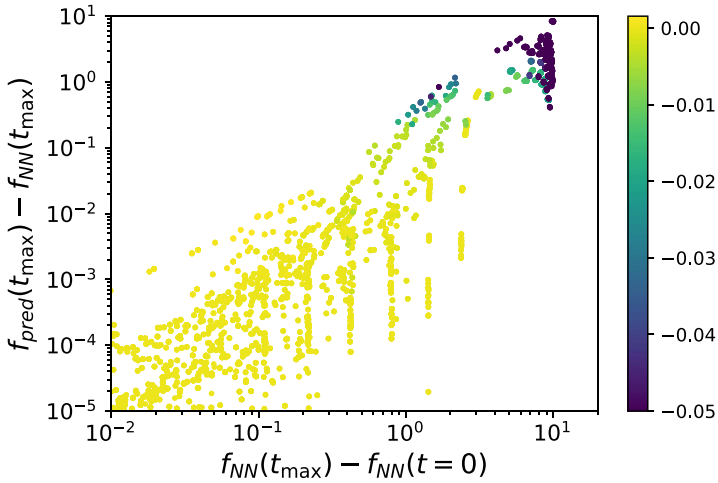

Figure 3. Comparison of the NN evolution and the predicted evolution according to equation (15). The observed discrepancy is related to the actual evolution in function space during training and found to be significantly smaller. The colour corresponds to the difference of α at the beginning and the end of training $\alpha(t_{\max}) - \alpha(t=0)$.

end of training[6]. We find excellent agreement across our hyperparameter search, and the most significant discrepancy is observed for large differences in α, which corresponds to the colour coding of the individual points $(\alpha(t_{\max}) - \alpha(t=0))$.

For networks with input dependence, we choose a one-dimensional input and data from a finite size interval around zero for simplicity, particularly because it allows for visualisation. The interval is taken as $0.1, 1$ times the respective Hubble length, set by $3\sqrt{\eta}/(1-\beta)$ and we use 100 equally spaced points from this interval. At initialisation, the network varies across the training interval, where the variation depends on the hyperparameter choices and the input interval. The distribution of the largest eigenvalue is shown in figure 4 which is very similar to the distribution of α observed in the uniform networks. We note a large hierarchy between the first and the following eigenvalues (in particular the higher eigenvalues), which is in agreement with previous observations in [6]. It is this suppression that leads to a freezing of the perturbations as predicted subsequently from our evolution of the perturbations in (16). In contrast, the mean evolution, corresponding to the largest eigenvalue, is observed quantitatively as the previous homogeneous and isotropic evolution. The most significant discrepancies are related to hyperparameters with the most considerable deviations in α during training. Figure 5 compares the predicted mean evolution with the

[6] We show results for the following hyperparameter choices: $\{1, 2, 5\}$ hidden layers with widths $\{100, 1000\}$, $\{erf, relu\}$ activations, and initialisations $\{(1, 0.05), (0.1, 0.05)\}$ learning and momentum rate combinations $\{(0.1111, 0.00), (0.4444, 0.00), (0.0011, 0.5), (0.0278, 0.5), (0.0278, 0.5), (0.1111, 0.5)\}$. The loss function choices were as mentioned before.

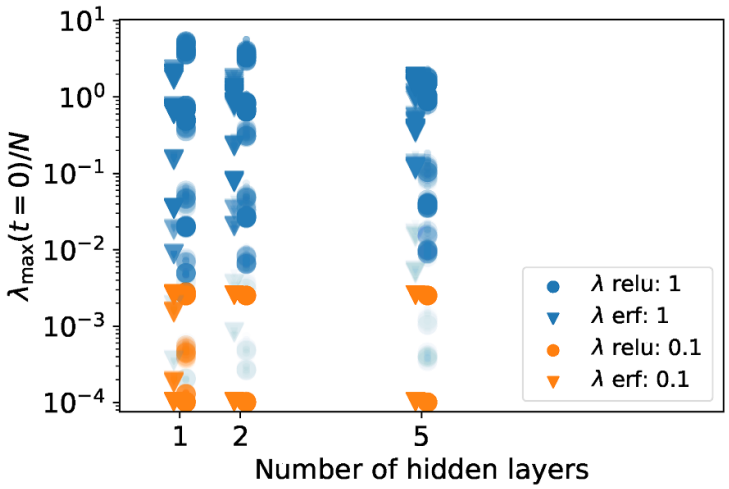

Figure 4. The values of the largest eigenvalues of the NTK-kernel divided by the number of points in the interval in our hyperparameter scan. The marker coding corresponds to the same coding as in figure 1.

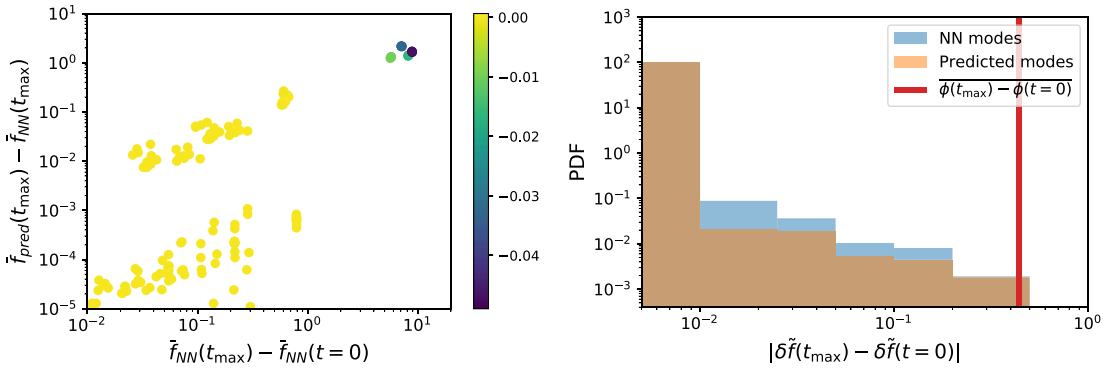

Figure 5. Left: the difference between the predicted and observed mean evolution using equation (15) in comparison to the actual mean evolution. Right: histogram of the modes at the beginning and the end of the training interval for the actual NN modes and the predicted modes. The red line shows the hierarchically larger mean value of the respective mean evolutions.

actual mean evolution on the left side[7]. On the right side, we show the distribution of the observed discrepancies at the end of training between the observed and the predicted mode evolution. We observe the hierarchical separation between the evolution of the perturbation and the mean evolution. All networks have been trained for 1000 epochs.

4. Related work

Various related approaches have been pursued on the physics and the NN side of the duality in recent years.

Undeniably our work builds upon the description of NN dynamics proposed in the context of the NTK, which was proposed in [6] and refined in [7]. Explicit realisations of the empirical NTK which we utilise can be found here [8]. Our work connects the NTK framework into the physical framework of EFT in the early Universe, i.e. scalar fields in an expanding Universe governed by general relativity.

The NN output averaged over different samples from the underlying weight distribution can be studied. In appropriate limits, these networks become Gaussian processes [13–17]. Such Gaussian processes are closely related to the situation when one wants to understand the ensemble of quantum fluctuations of scalar field theories as discussed in [18–22]. Related to our work, we utilise the same interpretation of a NN as a scalar field. Their approach paves the way for universal statements about a particular architecture, i.e. it can characterise the size of perturbations. In [23, 24], the loss function and gradient descent updates are included following the NTK procedure. In our work, we link the loss function and the gradient descent update to

[7] This scan involved NNs with one and two hidden layers, $\{relu, erf\}$ activations, 1000 hidden units, and activations $\{(1, 0.05), (0.1, 0.01)\}$. We utilised a learning rate of 0.1111 and no momentum.

physical processes in the dual description for the first time. In particular, we propose the distinction between mean-field and perturbation evolution.

In contrast to our function-space perspective, [12] develops a framework to understand the NN dynamics by focusing on the dynamics of the individual NN parameters. This allows to identify scaling such as with the depth and the width of NNs, i.e. in our language a microscopic analytic framework to calculate coefficients such as α. Such calculation of EFT coefficients from microscopic descriptions is a standard procedure in EFT. Here, we opt for the 'cheap' option of determining the coefficients experimentally. Along similar lines [25] presents a framework for understanding NN dynamics from a parameter space perspective.

Feature learning on the ML side can also be studied using the tensor approach put forward in [26], following earlier work in this direction [27–30]. Finally, on the cosmology side, feature learning corresponds to understanding the dynamics of the perturbations (see [31] for an early overview).

On the optimiser side, various other algorithms improve performance (e.g. using Adam optimiser) where the changed update equation changes the differential equation. However, a detailed comparison of different optimisation algorithms, notably including recent physics-inspired ones [32] is beyond the scope of our paper.

Emergent field theory dynamics, including the dynamics of gravity, have been studied from various perspectives in the past. This includes the work of analogue gravity (see [33] for a review), the emergent cosmological dynamics in group field theory [34], and in holographic setups [35, 36]. Our work adds a new dual system to gravitational dynamics. Similarly to other 'dual approaches,' it is fascinating to analyse and numerically understand gravitational systems' features.

Although the presence of dark energy in our Universe is by now well-established, its theoretical explanations via field theories often suffer from the hierarchically small positive vacuum energy required for consistency with observation in comparison to other natural energy scales of fundamental physics (see [37] for an overview of the problem and current approaches). Here the scale of the vacuum energy is related to the parameters of the optimiser. In particular, it is inversely related to the learning rate, which opens a new direction for model building. This connects with work on phenomenological models of extensions of ΛCDM such as for instance in [38–42].

Extensive cosmological simulations evolve a large number of particles subject to gravitational interactions (e.g. [43]), which build the basis for simulating the currently observed matter distribution in our Universe. Cosmological models at earlier times, in particular dynamics after inflation or during periods of matter domination, often require lattice simulations, utilising software such as LatticeEasy [44]. However, both types of simulations are resource-intensive, and this limits the capability of comparing differences in structure formation due to different microscopic models of gravity and particle physics (e.g. other dark matter models).

5. Conclusions and outlook

It seems rather remarkable that the dynamics of NNs trained with gradient descent are tightly connected with a field theory of gravitational dynamics. It is the interplay of all components in the ML setup, the training data, the NN architecture, the loss, and the optimiser which are required for this duality. As all of these components are naturally required in the NN learning setup, it seems intriguing that the physical system matching such dynamics requires gravitational dynamics, and a purely non-gravitational system (i.e. where gravitational dynamics decouple) does not capture the entire space of dynamics. We find that it is the microphysical model of the optimiser which sets interesting gravitational parameters, particularly the vacuum energy.

This raises several intriguing questions for future investigations:

- Which early Universe models can be described by NNs? How can we resemble the existing classes by adapting the optimiser, and which other dynamics – not yet studied in the context of early Universe model building – can be obtained? On the physics side, this raises another question: how unique are the gravitational dynamics of general relativity, i.e. to which extent can this duality hold in modified theories of gravity. Finally, on the ML side, to which empirical modifications of the NN training process do such changes in the physics model lead?
- On the ML side, the physics perspective of the duality allows us to design novel physics-inspired optimisers in the future. They will be inspired by the 'optimisation' observed in physical systems.
- How can we extend the current duality beyond the case of a Universe with vacuum energy domination? At this stage, our duality is shown for a small subset of cosmological dynamics, i.e. the case of vacuum energy

domination. It would be exciting to extend this duality in the first step to the standard instances of matter and radiation dominated Universes. However, this requires a modification of the optimisers used for the gradient descent update, which we leave for future work.

- Neural network dynamics show behaviour similar to that of phase transitions, e.g. completely different features at initialisation and the end of training, or, put in the language of EFT different collective modes dominate the estimation of NNs. Such phase transitions have happened throughout the evolution in the early Universe wherein different phases of other EFT can be used to capture the physics. To trace the cosmological evolution throughout such phase transitions is intrinsically tricky, and lattice simulations can become expensive when high resolutions are required. It will be extremely interesting to trace such phase transitions through our duality. The NN perspective offers a flexible approach that allows capturing both features through the capabilities of the NN.
- The spectrum of the NTK is dependent on the initialisation and the choice of NN architecture. It is intriguing to investigate how to match architectures to the scope of perturbations used in cosmology. This offers a new perspective on the initial conditions in the early Universe, i.e. how are features after training dominated by the choice of initial conditions. In addition it is very interesting to identify situations where more than one eigenvalue remains large, corresponding to more than one background field involved in the dynamics.
- This duality provides analytic control on why different architectures can lead to different perturbation growth and how feature learning depends on hyperparameter choices. This is essentially captured by the second derivative of the loss function which sees the mean field act like an external force on a harmonic oscillator (e.g. allowing for parametric resonance).

We hope to return to some of these questions in the not so distant future, shedding more light on the tantalising question: *Is our Universe a neural network trained with gradient descent?*

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Derivation of perturbation equation

Here we provide some helpful identities and a detailed derivation of equations (15) and (16). Using the decomposition of our NN into mean field and perturbations the derivatives of the loss function can be decomposed as follows:

$$f_i = \sqrt{N} v_1 \bar{f} + \delta f_i \,, \tag{21}$$

$$\begin{aligned} \mathcal{L}_i &= \frac{m^2}{N}(f_i - f_0) = \frac{m^2}{N}(\bar{f} - f_0) + \frac{m^2}{N}\delta f_i = \frac{1}{N}\mathcal{L}'(\bar{f}) + \frac{1}{N}\mathcal{L}''(\bar{f})\,\delta f_i \\ &= \frac{(v_1)_i}{\sqrt{N}}\mathcal{L}'(\bar{f}) + \frac{1}{N}\mathcal{L}''(\bar{f})\,\delta f_i \,, \end{aligned} \tag{22}$$

$$\mathcal{L}_i'' = \frac{m^2}{N} = \frac{1}{N}\mathcal{L}''(\bar{f}) \,, \tag{23}$$

Applying the basis transformations on the individual terms leads to:

$$A_{kl}\ddot{f}_l = \begin{pmatrix} \sqrt{N}\ddot{\bar{f}} + \left(v_1^T \cdot \delta\ddot{\mathbf{f}}\right) \\ \left(v_\mu^T \cdot \delta\ddot{\mathbf{f}}\right) \end{pmatrix} , \tag{24}$$

$$A_{kl}\mathcal{L}_l' = \begin{pmatrix} \frac{1}{\sqrt{N}}\mathcal{L}'(\bar{f}) + \frac{1}{N}\left(v_1^T \cdot \delta\mathbf{f}\right)\mathcal{L}''(\bar{f}) \\ \frac{1}{N}\left(v_\mu^T \cdot \delta\mathbf{f}\right)\mathcal{L}''(\bar{f}) \end{pmatrix} . \tag{25}$$

Combining these results, we have:

$$0 = A_{kl}\ddot{f}_l + \tilde{\beta} A_{kl}\dot{f}_l + A_{kl}\Theta_{lm}B^T_{mn}A_{no}\mathcal{L}_o' \tag{26}$$

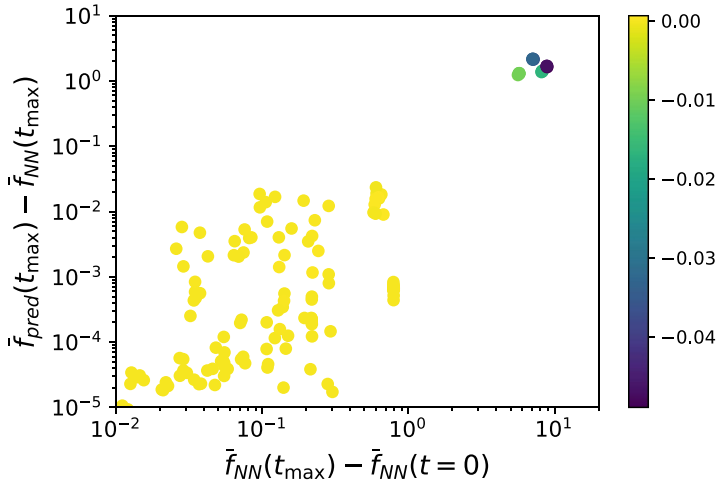

Figure 6. The difference between the predicted and observed mean evolution using equation (7) as the evolution for the inhomogeneous NNs used for figure 5 in comparison to the actual mean evolution.

$$= \begin{pmatrix} \sqrt{N}\ddot{\bar{f}} + \left(v_1^T \cdot \delta\ddot{\mathbf{f}}\right) \\ \left(v_\mu^T \cdot \delta\ddot{\mathbf{f}}\right) \end{pmatrix} + \tilde{\beta} \begin{pmatrix} \sqrt{N}\dot{\bar{f}} + \left(v_1^T \cdot \delta\dot{\mathbf{f}}\right) \\ \left(v_\mu^T \cdot \delta\dot{\mathbf{f}}\right) \end{pmatrix} + \begin{pmatrix} \frac{\lambda_1}{\sqrt{N}}\mathcal{L}'(\bar{f}) + \frac{\lambda_1}{N}\left(v_1^T \cdot \delta\mathbf{f}\right)\mathcal{L}''(\bar{f}) \\ \frac{\lambda_\mu}{N}\left(v_\mu^T \cdot \delta\mathbf{f}\right)\mathcal{L}''(\bar{f}) \end{pmatrix}. \tag{27}$$

This leads directly to the equations (15) and (16).

Appendix B. Mean field predictions for inhomogeneous networks

We empirically found agreement with the dynamics predicted in equation (15) as was shown in figure 5. To ensure consistency with the evolution in the homogeneous and isotropic case we also check the predictions for the mean field using the mean value α of the NTK instead of using λ_1/N. The resulting discrepancy is shown in figure 6. We find no significant deviations between the two predictions across our network sample.

ORCID iDs

Sven Krippendorf https://orcid.org/0000-0001-6374-6828
Michael Spannowsky https://orcid.org/0000-0002-8362-0576

References

[1] Polyak B 1964 Some methods of speeding up the convergence of iteration methods *USSR Comput. Math. Math. Phys.* **4** 1–17
[2] Mukhanov V 2005 *Physical Foundations of Cosmology* (Oxford: Cambridge University Press)
[3] Weinberg S 2008 *Cosmology* (Oxford: Oxford University Press)
[4] Baumann D 2009 Inflation *Theoretical Advanced Study Institute in Elementary Particle Physics: Physics of the Large and the Small*
[5] Zyla P A *et al* Particle Data Group 2020 Review of particle physics *Prog. Theor. Exp. Phys.* **2020** 083C01
[6] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *CoRR* (arXiv:abs/1806.07572 [cs.LG])
[7] Lee J, Xiao L, Schoenholz S S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2020 Wide neural networks of any depth evolve as linear models under gradient descent *J. Stat. Mech.: Theory Exp.* **2020** 124002
[8] Novak R, Xiao L, Hron J, Lee J, Alemi A A, Sohl-Dickstein J and Schoenholz S S 2019 Neural tangents: fast and easy infinite neural networks in Python (arXiv:1912.02803)
[9] Linde A D 1983 Chaotic inflation *Phys. Lett.* B **129** 177–81
[10] Qian N 1999 On the momentum term in gradient descent learning algorithms *Neural Netw.* **12** 145–51
[11] Su W, Boyd S and Candes E J 2016 A differential equation for modeling nesterov's accelerated gradient method: theory and insights *J. Mach. Learn. Res.* **17** 5312–54
[12] Roberts D A, Yaida S and Hanin B 2021 The principles of deep learning theory (arXiv:2106.10165 [cs.LG])
[13] Neal R M 2012 *Bayesian Learning for Neural Networks* vol 118 (Springer Science & Business Media)
[14] Lee J, Bahri Y, Novak R, Schoenholz S S, Pennington J and Sohl-Dickstein J 2017 Deep neural networks as Gaussian processes (arXiv:1711.00165 [stat.ML])

[15] Matthews A G D G, Rowland M, Hron J, Turner R E and Ghahramani Z 2018 Gaussian process behaviour in wide deep neural networks (arXiv:1804.11271 [stat.ML])
[16] Novak R, Xiao L, Lee J, Bahri Y, Yang G, Hron J, Abolafia D A, Pennington J and Sohl-Dickstein J 2018 Bayesian deep convolutional networks with many channels are Gaussian processes (arXiv:1810.05148 [stat.ML])
[17] Garriga-Alonso A, Rasmussen C E and Aitchison L 2018 Deep convolutional networks as shallow Gaussian processes (arXiv:1808.05587 [stat.ML])
[18] Halverson J, Maiti A and Stoner K 2021 Neural networks and quantum field theory *Mach. Learn.: Sci. Technol.* **2** 035002
[19] Maiti A, Stoner K and Halverson J 2021 Symmetry-via-duality: invariant neural network densities from parameter-space correlators (arXiv:2106.00694 [cs.LG])
[20] Erbin H, Lahoche V and Samary D O 2021 Nonperturbative renormalization for the neural network-QFT correspondence (arXiv:2108.01403 [hep-th])
[21] Halverson J 2021 Building quantum field theories out of neurons (arXiv:2112.04527 [hep-th])
[22] Grosvenor K T and Jefferson R 2022 The edge of chaos: quantum field theory and deep neural networks *SciPost Phys.* **12** 081
[23] Liu J, Tacchino F, Glick J R, Jiang L and Mezzacapo A 2021 Representation learning via quantum neural tangent kernels (arXiv:2111.04225 [quant-ph])
[24] Luo D and Halverson J 2021 Infinite neural network quantum states (arXiv:2112.00723 [quant-ph])
[25] Dyer E and Gur-Ari G 2019 Asymptotics of wide networks from Feynman diagrams (arXiv:1909.11304 [cs.LG])
[26] Yang G and Hu E J 2021 Feature learning in infinite-width neural networks (arXiv:2011.14522 [cs.LG])
[27] Yang G 2021 Tensor programs I: wide feedforward or recurrent neural networks of any architecture are Gaussian processes (arXiv:1910.12478 [cs.NE])
[28] Yang G 2020 Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation (arXiv:1902.04760 [cs.NE])
[29] Yang G 2020 Tensor programs II: neural tangent kernel for any architecture (arXiv:2006.14548 [stat.ML])
[30] Yang G 2020 Tensor programs III: neural matrix laws (arXiv:2009.10685 [cs.NE])
[31] Mukhanov V F, Feldman H A and Brandenberger R H 1992 Theory of cosmological perturbations. Part 1. Classical perturbations. Part 2. Quantum theory of perturbations. Part 3. Extensions *Phys. Rep.* **215** 203–333
[32] De Luca G B and Silverstein E 2022 Born-Infeld (BI) for AI: energy-conserving descent (ECD) for optimization (arXiv:2201.11137 [cs.LG])
[33] Barcelo C, Liberati S and Visser M 2005 Analogue gravity *Living Rev. Relativ.* **8** 12
[34] Gielen S, Oriti D and Sindoni L 2013 Cosmology from group field theory formalism for quantum gravity *Phys. Rev. Lett.* **111** 031301
[35] Maldacena J M 1998 The large *N* limit of superconformal field theories and supergravity *Adv. Theor. Math. Phys.* **2** 231–52
[36] Verlinde E P 2011 On the origin of gravity and the laws of newton *J. High Energy Phys.* **2011** 29
[37] Burgess C P 2015 The cosmological constant problem: why it's hard to get dark energy from micro-physics *100e Ecole d'Ete de Physique: Post-Planck Cosmology* pp 149–97
[38] Salti M, Kangal E and Aydogdu O 2019 Variable polytropic gas cosmology *Ann. Phys., NY* **407** 166–78
[39] Kangal E, Salti M and Aydogdu O 2019 Machine learning algorithm in a caloric view point of cosmology *Phys. Dark Universe* **26** 100369
[40] Tilaver H, Salti M, Aydogdu O and Kangal E 2021 Deep learning approach to Hubble parameter *Comput. Phys. Commun.* **261** 107809
[41] Salti M, Kangal E and Aydogdu O 2021 Evolution of CMB temperature in a Chaplygin gas model from deep learning perspective *Astron. Comput.* **37** 100504
[42] Salti M and Kangal E E 2022 Deep learning of CMB radiation temperature *Ann. Phys., NY* **439** 168799
[43] Springel V, Yoshida N and White S D M GADGET: a code for collisionless and gasdynamical cosmological simulations (arXiv:astro-ph/0003162 [astro-ph])
[44] Felder G N and Tkachev I 2008 LATTICEEASY: a program for lattice simulations of scalar fields in an expanding universe *Comput. Phys. Commun.* **178** 929–32