Neural network field theories: non-Gaussianity, actions, and locality

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 015002 https://doi.org/10.1088/2632-2153/ad17d3

MACHINE LEARNING
Science and Technology

PAPER

Neural network field theories: non-Gaussianity, actions, and locality

Mehmet Demirtas[1,2], **James Halverson**[1,2], **Anindita Maiti**[1,2,4,5,*], **Matthew D Schwartz**[1,3] and **Keegan Stoner**[1,2]

[1] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Boston, MA, United States of America
[2] Department of Physics, Northeastern University, Boston, MA 02115 United States of America
[3] Department of Physics, Harvard University, Cambridge, MA 02138 United States of America
[4] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 United States of America
[5] Perimeter Institute for Theoretical Physics, 31 Caroline St N, Waterloo, ON N2L 2Y5, Canada
[*] Author to whom any correspondence should be addressed.

E-mail: amaiti@perimeterinstitute.ca

Keywords: neural network field theory correspondence, Feynman rules for neural network field theories, non-perturbative field theories via neural networks

OPEN ACCESS

RECEIVED
15 September 2023

REVISED
6 December 2023

ACCEPTED FOR PUBLICATION
18 December 2023

PUBLISHED
9 January 2024

Abstract

Both the path integral measure in field theory (FT) and ensembles of neural networks (NN) describe distributions over functions. When the central limit theorem can be applied in the infinite-width (infinite-N) limit, the ensemble of networks corresponds to a free FT . Although an expansion in $1/N$ corresponds to interactions in the FT, others, such as in a small breaking of the statistical independence of network parameters, can also lead to interacting theories. These other expansions can be advantageous over the $1/N$-expansion, for example by improved behavior with respect to the universal approximation theorem. Given the connected correlators of a FT, one can systematically reconstruct the action order-by-order in the expansion parameter, using a new Feynman diagram prescription whose vertices are the connected correlators. This method is motivated by the Edgeworth expansion and allows one to derive actions for NN FT. Conversely, the correspondence allows one to engineer architectures realizing a given FT by representing action deformations as deformations of NN parameter densities. As an example, ϕ^4 theory is realized as an infinite-N NN FT.

Contents

1. Introduction

The last decade has seen remarkable progress in machine learning (ML) in a wide variety of fields, including traditional ML fields such as natural language processing, image recognition, and gameplay (see [1] for reviews, and [2] for some breakthroughs in the literature), but also in the physical sciences [3], and more recently to obtain rigorous results in pure mathematics [4]. This progress has been facilitated in part by the

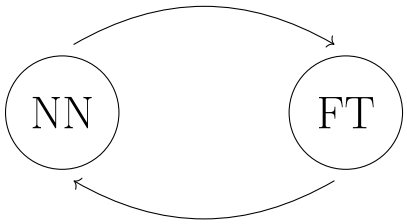


Figure 1. In a NN-FT correspondence, ideas from one may give insights into the other. In this paper we are primarily interested in understanding when NN FT exhibit physical principles such as non-Gaussianity and locality, with an eye towards applications in both ML and especially physics in the future.

increasing complexity of deep neural networks (NN), both in terms of the number of parameters appearing in them and their architecture. However, despite their empirical success, the theoretical foundations of deep NN are still not fully understood. Natural questions emerge:

- Are ideas from the sciences, such as physics, useful in NN theory?
- As it develops, does ML theory lead to progress in the sciences?

A growing literature (see below), gives an affirmative answer to the first, but the second is less clear; it is applied ML, not theoretical ML, that is primarily used in the sciences.

In this paper we explore both of these questions by further developing a correspondence between NN and field theory (FT). This connection was already implicit in Neal's PhD thesis [5] in the 1990's, where he demonstrated that an infinite width single-layer NN is (under appropriate assumptions) a draw from a Gaussian process (GP). This is the so-called NNGP correspondence, and in recent years it has been shown that most modern NN architectures [6, 7] have a parameter N such that the NN is drawn from a GP in the $N \to \infty$ limit. The NNGP correspondence is of interest from a physics perspective because GPs are generalized non-interacting (free) FT, and NN provide a novel way to realize them. Non-Gaussianities emerge at finite-N, which correspond to turning on interactions that are generally non-local, and may be captured by statistical cumulant functions, known as connected correlators in physics. As we will see, since Gaussianity in the $N \to \infty$ limit emerges by the central limit theorem (CLT), non-Gaussianities may be studied more generally by parametrically violating necessary conditions of the CLT.

These results provide a first glimpse that there is a more general NN-FT correspondence that should be developed in its own right, taking inspiration from both physics and ML. In this introduction we will review the central ideas of the correspondence and introduce principles for understanding the literature, which we review in part. Readers familiar with the background are directed to section 1.3 for a summary of our results.

1.1. NN-FT correspondence

At first glance, NN and FT seem very different from one another. However, in both cases, the central objects of study are *random functions*. The random function ϕ associated to a NN is defined by its *architecture*, which is a composition of simpler functions that involves parameters θ. At program initialization, parameters are drawn as $\theta \sim P(\theta)$, yielding a randomly initialized NN, i.e. a random function. In FT, the random functions are simply the fields themselves, typically described by specifying their probability density function directly, $P(\phi) = \exp(-S[\phi])$, via the Euclidean action functional $S[\phi]$; we work in Euclidean signature throughout.

We therefore have two different origins for the statistics of a FT, shown in figure 1. To exemplify the point, consider a FT defined by an ensemble of networks or fields $\phi : \mathbb{R} \to \mathbb{R}$,

$$\phi(x) = a\,\sigma(b\,\sigma(cx)) \qquad a \sim P(a),\ b \sim P(b),\ c \sim P(c), \tag{1.1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ acts element-wise and is generally taken to be non-linear. Here the statistics of the ensemble arise from how it is *constructed*, rather than from the density $\exp(-S[\phi])$ over functions from which it is drawn. We will refer to such a description as the *parameter space* description of a *NN FT*. The construction of ϕ defined in (1.1) has two parts, the architecture that defines its functional form, and the choice of distributions from which the parameters a, b, and c are drawn. This particular architecture is a feedforward network with depth two, width one, and activation function σ. In this description of the FT, one does not necessarily know the action $S[\phi]$, but the theory may nevertheless be studied because the architecture and parameter densities define its statistics.

For instance, the correlation functions of a NN FT can be expressed as

$$G^{(n)}(x_1,\ldots,x_n) := \mathbb{E}[\phi(x_1)\ldots\phi(x_n)] = \int \mathrm{d}\theta\, P(\theta)\,\phi(x_1)\ldots\phi(x_n), \tag{1.2}$$

where we denote the set of parameters of the NN by θ, and the network / field ϕ depends on parameters through its architecture. Alternatively, we could provide a *function space* description of the theory by specifying the action $S[\phi]$ and express the correlation functions as

$$G^{(n)}(x_1,\ldots,x_n) = \int D\phi\, e^{-S[\phi]}\,\phi(x_1)\ldots\phi(x_n), \tag{1.3}$$

as in a first course on quantum FT (QFT). These expressions may be derived from the partition function

$$Z[J] = \mathbb{E}[e^{\int \mathrm{d}^d x J(x)\phi(x)}], \tag{1.4}$$

where the parameter space and function space results arise by specifying how the expectation value is computed,

$$Z[J] = \int \mathrm{d}\theta\, P(\theta)\, e^{\int \mathrm{d}^d x J(x)\phi(x)} \tag{1.5}$$

$$Z[J] = \int D\phi\, e^{-S[\phi]+\int \mathrm{d}^d x J(x)\phi(x)}. \tag{1.6}$$

In this work, many calculations will be carried out in terms of a general expectation value $\mathbb{E}[\cdot]$ that denotes agnosticism towards the origin of the statistics; explicit calculations may be carried out by replacing $\mathbb{E}$ with one description or the other, as in passing from a general expression (1.4) to those of parameter space (1.5) and function space (1.6).

Parameter space and function space provide two different descriptions of a FT, which could be thought of as different duality frames [8]. When one defines a FT by a NN architecture, the parameter space description is readily available, but the action is not known, *a priori*. However, if the parameter distributions are easy to sample then the fields are also easy to sample: one just initializes NN on the computer. On the other hand, in FT we normally proceed by first specifying an action; in this case, the probability of a given field configuration is known because $P[\phi] = \exp(-S[\phi])$ is known, but fields are notoriously hard to sample, as evidenced by the proliferation of Monte Carlo techniques in lattice FT.

1.1.1. Example: NNGP correspondence in parameter space and function space

Let us study an example to make the abstract notions more concrete. Consider a fully-connected feedforward network $\phi : \mathbb{R}^d \to \mathbb{R}$ with depth one and width N,

$$\phi(x) = \sum_{i=1}^{N}\sum_{j=1}^{d} a_i\,\sigma\,(b_{ij}x_j), \qquad a \sim \mathcal{N}(0,\sigma^2/N),\ b \sim \mathcal{N}(0,\sigma^2/d), \tag{1.7}$$

where σ is an elementwise non-linearity such as tanh or $\mathrm{ReLU}(z) := \max(0,z)$. Here, the set of parameters θ is given by the union of the a-parameters and the b-parameters. As we will see in detail in section 2, if the parameters are drawn independently then the connected correlation functions

$$G_c^{(2k)}(x_1,\ldots,x_{2k}) \propto \frac{1}{N^{k-1}}, \tag{1.8}$$

and the odd-point correlation functions vanish due to a having zero mean. In the $N\to\infty$ limit, also known as the GP limit, then, the only non-vanishing connected correlator has two points,

$$G_c^{(2)}(x_1,x_2) \tag{1.9}$$

which demonstrates that the theory is Gaussian; this is the NNGP correspondence. Concretely, following [9], we may compute the two-point function as

$$G^{(2)}(x,y) = \mathbb{E}[\phi(x)\phi(y)] = \int \mathrm{d}a\,\mathrm{d}b\,P(a)P(b)\; a_{i_1}\sigma(b_{i_1 j_1}x_{j_1})\; a_{i_2}\sigma(b_{i_2 j_2}y_{j_2}) \tag{1.10}$$

where we have used Einstein summation and left the details of the Gaussian parameter densities $P(a)$ and $P(b)$ implicit. For a fixed choice of σ one may evaluate this integral analytically or via Monte Carlo sampling,

resulting in the two-point function; analytic integrated results for $\sigma = \tanh$ and $\sigma = \text{Erf}$ are presented in [9]. Since the parameter space calculation establishes Gaussianity of the theory, we infer the action

$$S[\phi] = \int \mathrm{d}^d x \mathrm{d}^d y \, \phi(x) \, G^{(2)}(x,y)^{-1} \, \phi(y), \tag{1.11}$$

where the inverse of the two-point function satisfies $\int \mathrm{d}^d y G^{(2)}(x,y)^{-1} G^{(2)}(y,z) = \delta^{(d)}(x-z)$.

As a concrete example, we refer the reader to section 4.2, which recalls a NN realization of free scalar FT from [10] that uses a cos activation. In that case we have

$$G^{(2)}(x,y)^{-1} = \delta(x-y)(\nabla^2 + m^2) \tag{1.12}$$

which reproduces the usual free scalar action

$$S[\phi] = \int \mathrm{d}^d x \, \phi(x) \left(\nabla^2 + m^2\right) \phi(x), \tag{1.13}$$

in this case realized via a concrete NN architecture.

Thus, in the GP limit, both the parameter space and function space descriptions of the FT are readily available. Building on [11] using the Edgeworth expansion, we will see methods for computing approximate actions at finite-N, and we will also develop techniques to engineer desired actions.

1.2. Organizing principles and related work

We have discussed a foundational principle underlying the NN-FT correspondence, that parameter space and function space provide two different descriptions of the statistics of an ensemble of NN or fields. Though we have given an example, and there are many more, we are still in very general territory and it is not clear where to go. Accordingly, we would like to provide other organizing principles:

- **NN-for-FT vs. FT-for-NN:** are we aiming to better understand physics or ML?
- **Fixed Initialization vs. Learning:** are we aiming to understand a fixed NN-FT at initialization, or a one-parameter family of NN-FTs defined by some dynamics, such as ML training dynamics or FT flows?

Much of the existing literature can be classified within each of these principles, and they also set context for discussing our results. We will first review some results for network ensembles at initialization, and then during and after training. With these ideas in place, we will turn to the idea of using NN-FT in service of FT.

For literature that is most similar in perspective to this introduction (prior to this reference section), see [10] and the works that preceded it [8, 12], by subsets of the authors.

1.2.1. Initialization

A NN with parameters θ and parameter distribution $P(\theta)$ is initialized on a computer by drawing $\theta \sim P(\theta)$ and inserting them into the architecture, generating a random function $\phi(x)$ that is sampled from a distribution $P(\phi)$ that may or may not be known. In the $N \to \infty$ NNGP limit, $P(\phi)$ is Gaussian. This was shown for feed forward networks in Neal's thesis [5], as well as more recently in [6]; was generalized to a plethora of architectures, e.g. convolutional layers [7, 13, 14], recurrent layers, graph convolutions [15], skip connections [16], attention [17], and batch /layer normalization in [18], pooling [14], and transformers [19, 20]. The generality of this result arises from the generality in which central limit theorem behavior manifests itself in NN; see [7] for a systematic treatment in the tensor programs formalism.

Since Gaussianity follows from the central limit theorem, one generally expects non-Gaussianities in the form of $1/N$-corrections. Study of these non-Gaussianities was initiated a few years ago; e.g. [21] computed leading non-Gaussianities via the connected four-point function, [22] showed for deep feedforward networks how $P(\phi)$ is perturbed by $1/N$-corrections, [12] proposed using effective FT to model non-Gaussian $P(\phi)$ for NN, and [23] developed an effective theory approach and an L/N expansion that controls feature learning in deep feedforward networks; for concreteness in our examples, we are interested in the distribution of networks at initialization and take $L = 1$. This L/N expansion allowed [23] to also study signal propagation through the network, identify universality classes, and tune hyperparameters to criticality.

Methods borrowed from FT have been useful in studying NNs at initialization. For example, perturbative methods like Feynman diagrams were employed in [12, 24, 25]. Various schemes for renormalization group flow, including non-perturbative ones, were applied to NNs in [26]. Global symmetries of NN-FTs were shown to arise from symmetry invariances of NN parameter distributions in [8]. While the results of this paper were being finalized, a recent paper [27] brought forward a different diagrammatic approach to effective FT in deep feedforward networks.

1.2.2. Learning

Although we do not study the dynamics of learning in this paper, it is a goal for future work. Therefore, we would like to review some of the literature.

NN may be trained to perform useful tasks via a variety of learning schemes, such as supervised learning or reinforcement learning, that utilizes a learning algorithm to update the system, such as stochastic gradient descent. In practice this involves training one or a handful of randomly initialized NN to convergence. However, in general there is nothing special about the initial networks that were trained; in the absence of compute limitations, one would prefer to train *all* the networks and compute an ensemble average at convergence. Theoretically, this amounts to tracking the distributional flow of the NN ensemble, and in principle it may be done in either parameter space or function space.

In the $N \to \infty$ limit, most known architectures define NN that are draws from GPs. Since the architecture defines a GP, it could be used as a prior in Bayesian inference, the learning algorithm of interest in Neal's original work [5]. On the other hand, gradient descent with continuous time is governed by the neural tangent kernel (NTK) [28], which becomes deterministic and training time t-independent in the so-called frozen-NTK limit. In this limit, $N \to \infty$ and the NN dynamics is well-approximated by that of a model that is linear in the NN parameters. This frozen behavior is a vast simplification of the dynamics and is known to exist for many architectures, such as convolutional NN [29], graph NN [30], recurrent networks [31], and attention layers [19]. For supervised learning with MSE loss, the NN ensemble trained under gradient descent remains a GP for all times t, including $t \to \infty$, with known mean and covariance; the dynamics becomes that of kernel regression, with kernel given by the frozen-NTK. How is this related to Neal's desire to relate Bayesian inference and trained NN? If all but the last layer's weights are frozen, then the NTK is the NNGP kernel and the distribution of the NN ensemble converges to the GP Bayesian posterior as $t \to \infty$.

In summary, in the $N \to \infty$ limit, the distribution of the NN ensemble is Gaussian. If it undergoes supervised training with MSE loss, it remains Gaussian at all times and converges to the Bayesian GP posterior in a particular case [32]. In general, however, gradient descent induces non-Gaussianities.

At finite-N, the NN ensemble is non-Gaussian. In the Bayesian context, this defines a non-Gaussian prior, and inference may be performed for weakly non-Gaussian priors via a $1/N$-expansion [21]. In the gradient descent context, the NTK is no longer frozen and evolves during training, significantly complicating the dynamics. Work by Roberts, Yaida, and Hanin develops a theory of an evolving NTK in [23]. They apply it in detail to fully-connected networks of depth L, demonstrate the relevance of L/N as an expansion parameter, and develop an effective model for the dynamics. Such $1/N$ corrections to dynamical NTK were previously studied by other authors in [24, 33]. Bordelon and Pehlevan have developed a systematic understanding of the evolution of NTK and parametric interpolations between rich and lazy training regimes using the framework of dynamical mean FT, see [34]. Some of these authors have studied the $O(1/N)$ suppressed corrections to training dynamics of finite width Bayesian NNs in [35]. A separate work, [36], presents close-to-Gaussian NN processes including stationary Bayesian posteriors in the joint limit of large width and large data set, using $1/N$ as an expansion parameter. Moreover, the authors of [37] explore a correspondence between learning dynamics in the continuous time limit and early Universe cosmology, and [38] analyzes connected correlation functions propagating through NN.

1.2.3. NN-for-FT

NN, including the ones we have discussed thus far, generally have $\mathbb{R}^n$ as their domain and therefore naturally live in Euclidean signature. They define statistical FT that may or may not have analytic continuations to quantum field theories in Lorentzian signature. Nevertheless, statistical FT are interesting in their own right and NN-FT provides a novel way to study them.

Using an architecture to define a FT enables a parameter space description that makes sampling, and therefore numerical simulation on a lattice, easy. If one can determine an easily sampled NN architecture that engineers standard Euclidean ϕ^4 theory, for instance, this could lead to improved results on the lattice by avoiding Monte Carlo entirely[6]. This is an engineering problem that is work-in-progress; it is not clear that the ϕ^4 NN-FT realization in this work is easily sampled. Alternatively, by simply fixing an easily sampled architecture with interesting physical properties such as symmetries and strong coupling, lattice simulation could be performed immediately.

For uses in fundamental and formal quantum physics, one might wish to know when a NN architecture defines a QFT. Since NN architectures are usually defined in Euclidean signature, we may instead ask when a Euclidean FT admits an analytic continuation to Lorentzian signature that defines a QFT. The situation is

[6] This lattice approach should be contrasted with works [39] that train a normalizing flow to give proposals for the accept/reject step of MCMC.

complicated by the fact that in general we do not know the action, but instead have access to the Euclidean correlation functions, expressed in parameter space.

Fortunately, the Osterwalder–Schrader (OS) theorem [40] of axiomatic FT gives necessary and sufficient conditions, expressed in terms of the correlators, for the existence of a QFT after continuation. The axioms include

- **Euclidean Invariance.** Correlation functions must be Euclidean invariant, which becomes Lorentz invariance after analytic continuation. See [10] for an infinite ensemble of NN architectures realizing Euclidean invariance.
- **Permutation Symmetry**. Correlation functions must be invariant under permutations of their arguments, a collection of points in Euclidean space. This is automatic in NN-FTs with scalar outputs.
- **Reflection Positivity**. Correlation functions must satisfy a positivity condition known as reflection positivity, which is necessary for unitarity and the absence of negative-norm states in the analytically continued theory.
- **Cluster Decomposition.** Correlation functions must satisfy cluster decomposition, which says that interactions must shut off at infinite distance. As a condition on connected correlators, cluster decomposition is

$$\lim_{b\to\infty} G_c^{(n)}(x_1,\ldots,x_p,x_{p+1}+b,\cdots,x_n+b)\to 0, \tag{1.14}$$

  for any value of $1<p<n$. We have assumed permutation symmetry to simplify notation, putting the shifts into x_{p+1} into x_n.

These ideas were utilized in [10] to define NN *quantum* FT: a NN-QFT is a NN architecture whose correlation functions satisfy the OS axioms, and therefore defines a QFT upon analytic continuation. To date, the only known example is a NN architecture that engineers a standard free scalar FT in d-dimensions, though we improve the situation in this work by developing techniques to engineer local Lagrangians, which automatically satisfy the OS axioms. To make further progress on NN-QFT in a general setting, one needs especially a deeper understanding of reflection positivity and cluster decomposition in interacting NN-FTs; we study the latter.

1.3. Summary of results and paper organization

Since there are a number of different themes and concepts in this paper, we would like to highlight some of the major conceptual results:

- **Parametric Non-Gaussianity: $1/N$ and Independence Breaking.**
  section 2 approaches interactions in NN-FT (non-Gaussianity) by parametrically breaking necessary conditions for the central limit theorem to hold. Violating the infinite-N limit is well studied, but we also systematically study interactions arising from the breaking of statistical independence, and apply these ideas in examples.
- **Computing Actions with Feynman Diagrams.**
  In section 3 we develop a general FT technique for computing the action diagrammatically. The coupling functions are computed with a new type of connected Feynman diagram, whose vertices are the connected correlators. This is a swapping of the normal role of couplings and connected correlators, which arises from a 'duality' that becomes apparent via the Edgeworth expansion. The technique is also applied to NN-FT, including an analysis of how actions may be computed in the two regimes of parameteric non-Gaussianity developed in section 2, $1/N$ and independence breaking.
- **Engineering Actions in NN-FT.**
  In section 4 we develop techniques for engineering actions in NN-FT. This is to be distinguished from the approach of section 3: instead of fixing an architecture, computing its correlators, and then computing its action via Feynman diagrams, in section 4 we fix a desired action and develop techniques for designing architectures that realize the action. Adding a desired term to the action manifests itself in NN-FT by deforming the parameter distribution, which breaks statistical independence if it is a non-Gaussianity. Using this technique, local actions may be engineered at infinite-N.
- **ϕ^4 as a NN-FT.**
  In section 4.2 we design an infinite width NN architecture that realizes ϕ^4 theory, using the techniques that we developed.
- **The Importance of $N\to\infty$ for Interacting Theories.** In physics, interesting theories defined by a fixed action S generally have a wide variety of finite action field configurations, which have non-zero probability

density. This is potentially at odds with the universal approximation theorem: if a single finite-action configuration cannot be realized by an architecture A, but only approximated, then any NN-FT associated to A cannot realize the FT associated to S. If the $1/N$ is an expansion parameter for both non-Gaussianities and the degree of approximation, as e.g. with single-layer width-N networks, this simple no-go theorem suggests that exact NN-FT engineering of well-studied theories in physics occurs most naturally at infinite-N, as we saw in the case of ϕ^4 theory.

These are highlights of the paper. For more detailed summaries of results, we direct you to the beginning of each section.

2. Connected correlators and the central limit theorem

Interacting FT with a Lagrangian description are defined by non-Gaussian field densities $\exp(-S[\phi])$. If the non-Gaussianities are small, the theory is close to Gaussian and weakly interacting, in which case correlation functions may be computed in perturbation theory using Feynman diagrams. The non-Gaussianities are captured by the higher connected correlation functions, which vanish in the Gaussian limit. They are known as cumulants in the statistics literature and may be obtained from a generating functions $W[J]$ as

$$G_c^{(n)}(x_1,\ldots,x_n) := \left(\frac{\delta}{\delta J(x_1)}\cdots\frac{\delta}{\delta J(x_n)}W[J]\right)\Bigg|_{J=0}, \qquad W[J] := \ln Z[J]. \tag{2.1}$$

In the absence of a known Lagrangian description, connected correlators still encode the presence of non-Gaussianities, since the theory is Gaussian if $G_c^{(n)} = 0$ for $n > 2$.

In this section we systematically study non-Gaussianities in NN-FT. Since the parameter space description exists for any NN-FT, we choose to study non-Gaussianities via connected correlators (rather than actions), which may be studied in parameter space even when the action is unknown. We are interested in non-Gaussianities in NN-FT for a number of reasons. In the NN-for-FT direction, it is important for understanding interactions in the associated FT. Conversely, in the FT-for-NN direction, understanding non-Gaussianities is important for capturing the statistics of finite networks and networks with correlations in the parameter distributions, which generally develop during training.

The essential idea in our approach is to recall the origin of Gaussianity, and then parametrically move away from it. Specifically, many FT defined by NN architectures admit an $N \to \infty$ limit in which they are Gaussian, and the Gaussianity has a statistical origin: the central limit theorem (CLT). The CLT states that the distribution of the standardized sum of N independent and identically distributed random variables approaches a Gaussian distribution in the limit $N \to \infty$. Therefore we may systematically study non-Gaussianities in NN FT by violating assumptions of the CLT, e.g. via $1/N$ corrections and breaking the independence condition, both of which affect connected correlators.

There are a number of results and themes in this section, which is organized as follows:

- **CLT.** In section 2.1 we review the CLT from the perspective of cumulant generating functionals, which will be useful in NN-FT since in general we do not have a simple expression for the action but do have access to cumulants.
- **Independence Breaking.** In section 2.2 we introduce how non-Gaussianities may also arise by violating the statistical independence assumption of the CLT. We characterize this by a family of joint densities with parameter α that factorize (become independent) when $\alpha = 0$. We study the α-dependence of cumulants via Taylor series, showing that α controls non-Gaussianities independently of those arising from $1/N$-corrections. A simple example of independence-breaking induced non-Gaussianities at $N = \infty$ is given in section 2.2.1.
- **Connected Correlators and Interactions in NN-FT.** In section 2.3 we study non-Gaussianities in NN-FT, decomposing the field $\phi(x)$ into N constituent neurons as in [10]. We study the case of independent neurons in section 2.3.1, where we present the N-scaling of connected correlators and also two examples: single-layer Cos-net, which exhibits full Euclidean symmetry in all of its correlators, and $d = 1$ ReLU-net, which we show exhibits an interesting bilocal structure in its two-point and four-point functions.

  In section 2.3.2 we turn to breaking neuron independence in NN-FT, building on the independence breaking results of [10], which gives a new source of interactions and a generalized formula for connected correlators. Specifically, we introduce a general formalism for the expansion of the cumulant generating functional in terms of independence-breaking parameters, and therefore the computation of connected correlators. As an example, we deform the Cos-net theory to have non-independent neurons via

non-independent input weights, doing the deformation in a way that preserves Euclidean invariance, and compute the independence-breaking correction to the connected four-point function.
- **Identical-ness Breaking.** Interactions may also arise from breaking the identical-ness assumption of the CLT. See appendix B for an example of a NN-FT with non-Gaussianities arising from identical-ness breaking.

Equipped with two different types of parameters that induce non-Gaussianity in connected correlators, $1/N$ and independence-breaking parameters, we will see how this may be used to approximate actions in section 3.

2.1. Review: CLT from generating functions

In order to understand non-Gaussianities in NN-FTs, it is useful to recall essential aspects of the CLT in the case of a single random variable, since they carry over to the NN-FT case. We will do so using the language of generating functions and cumulants (connected correlators), since we may use them to study Gaussianity and non-Gaussianity even if the NN-FT action is unknown.

Of course, the CLT is among the most fundamental theorems of statistics. There are many variants of it in the literature, with different sets of assumptions. Here, we will describe a particularly simple version of it and provide a proof, showing how key assumptions come into play. For a more in depth discussion of the CLT, see e.g. [41].

Consider N random variables X_i. Assume that they are *identical*, *independent*, *mean-free*, and have *finite variance*. The CLT states that the standardized sum

$$\phi = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \tag{2.2}$$

is drawn from a Gaussian distribution in the limit $N \to \infty$. In other words, even if the X_i are sampled from complicated, non-Gaussian distributions, these details wash out and their sum is drawn from a Gaussian distribution.

To see the Gaussianity in a way that may be extrapolated to NN-FT, it is useful to introduce generating functions. The moment generating function of ϕ is defined as

$$Z_\phi[J] := \mathbb{E}[e^{J\phi}] = \mathbb{E}[e^{J\sum_i X_i/\sqrt{N}}], \tag{2.3}$$

from which we can extract the moments by taking derivatives,

$$\mu_r^\phi := \mathbb{E}[\phi^r] = \left(\frac{\partial}{\partial J}\right)^r Z_\phi[J]. \tag{2.4}$$

In physics language, J is the source, $Z_\phi[J]$ is the partition function, and μ_r^ϕ is the rth correlator of ϕ. The cumulant generating functional (CGF) of ϕ is the logarithm of the moment generating functional

$$W_\phi[J] := \log \mathbb{E}[e^{J\phi}] = \log \mathbb{E}[e^{J\sum_i X_i/\sqrt{N}}], \tag{2.5}$$

and the cumulants κ_r^ϕ are computed by taking derivatives of $W_\phi[J]$,

$$\kappa_r^\phi := \left(\frac{\partial}{\partial J}\right)^r W_\phi[J]. \tag{2.6}$$

A random variable is Gaussian only if its cumulants $\kappa_{r>2}^\phi$ vanish. Fundamental properties of CGFs include

$$W_{X+c}[J] = cJ + W_X[J], \tag{2.7}$$
$$W_{cX}[J] = \log \mathbb{E}[e^{Jcx}] = W_X[cJ] \tag{2.8}$$

where $c \in \mathbb{R}$ is a constant, which imply

$$\kappa_1^{X+c} = \kappa_1^X + c \qquad \kappa_{r>1}^{X+c} = \kappa_{r>1}^X \tag{2.9}$$
$$\kappa_r^{cX} = c^r \kappa_r^X, \tag{2.10}$$

respectively.

We would like to see the Gaussianity of ϕ under CLT assumptions by computing cumulants. This is possible since $\kappa_{r>2} = 0$ is necessary for Gaussianity; conversely, we may study non-Gaussianities in terms of non-vanishing higher cumulants. Specifically, for a sum of *independent* random variables the moment generating function factorizes,

$$Z_{X_1+\cdots+X_N}[J] = \prod_i^N Z_{X_i}[J]. \tag{2.11}$$

Consequently, the CGF and the cumulants become

$$W_{X_1+\cdots+X_N}[J] = W_{X_1}[J] + \cdots + W_{X_N}[J], \tag{2.12}$$

$$\kappa_r^{X_1+\cdots+X_N} = \kappa_r^{X_1} + \cdots + \kappa_r^{X_N}. \tag{2.13}$$

Using the identities in (2.10) we can write the cumulants of ϕ as

$$\kappa_r^{\phi} = \frac{\kappa_r^{X_1} + \cdots + \kappa_r^{X_N}}{N^{r/2}}. \tag{2.14}$$

When the X_i are *identical* this simplifies to

$$\kappa_r^{\phi} = \frac{\kappa_r^{X_i}}{N^{r/2-1}}. \tag{2.15}$$

The cumulants $\kappa_{r>2}^{\phi}$ vanish in the $N \to \infty$ limit. To establish that ϕ is Gaussian, we also need to show that κ_1^{ϕ} and κ_2^{ϕ} are finite. As the X_i are mean-free, $\kappa_1^{\phi} = \kappa_1^{X_i}/\sqrt{N} = 0$, while $\kappa_2^{\phi} = \kappa_2^{X_i}$ is finite by assumption. Thus, ϕ is Gaussian distributed. This is the CLT, cast into the language of cumulants.

We emphasize that this result relies not only on the $N \to \infty$ limit, but also on the independence assumption (2.13).

2.2. Non-Gaussianity from independence breaking

We wish to study the emergence of non-Gaussianity by breaking the independence condition.

To do so, we must parameterize the breaking of statistical independence. Let $p(X;\alpha)$ be a family of joint distributions on X_i parameterized by a hyperparameter α that must be chosen in order to define the problem. We choose the family of joint distributions to be of the form

$$p(X;\alpha = 0) = \prod_i p(X_i), \tag{2.16}$$

i.e. $p(X)$ is independent in the $\alpha \to 0$ limit, but $\alpha \neq 0$ in general controls the breaking of independence. Then we obtain

$$W_{\phi}[J] = \log \mathbb{E}[e^{J\sum_i X_i/\sqrt{N}}] = \log \int \prod_j \mathrm{d}X_j\, p(X;\alpha)\, e^{J\sum_i X_i/\sqrt{N}} \tag{2.17}$$

which when expanded around $\alpha = 0$ yields

$$W_{\phi}[J] = \log \left[\prod_j \mathbb{E}_{p(X,\alpha=0)}[e^{JX_j/\sqrt{N}}] + \sum_{k=1}^{\infty} \frac{\alpha^k}{k!} \int \prod_j \mathrm{d}X_j\, e^{J\sum_i X_i/\sqrt{N}}\, \partial_{\alpha}^k p(X;\alpha)|_{\alpha=0} \right], \tag{2.18}$$

where the first term of the log uses independence of $p(X;\alpha = 0)$.

To deal with the α-dependent terms, we generalize a trick appearing regularly in ML, e.g. in the policy gradient theorem in reinforcement learning. There, the fact that $p\,\partial_{\alpha} \log p = \partial_{\alpha} p$ allows us to write

$$\partial_{\alpha} \mathbb{E}[\mathcal{O}] = \mathbb{E}[\mathcal{O}\,\partial_{\alpha} \log p] \tag{2.19}$$

for any α-independent operator $\mathcal{O}$. Generalizing, we define

$$\mathcal{P}_k := \frac{1}{p} \partial_{\alpha}^k p, \tag{2.20}$$

and note that it satisfies the recursion relation

$$\mathcal{P}_{k+1} = \mathcal{P}_1 \mathcal{P}_k + \partial_{\alpha} \mathcal{P}_k, \tag{2.21}$$

which allows for efficient computation. We can then write (2.18) as

$$W_\phi[J] = \log\left[\prod_j \mathbb{E}_{p(X,\alpha=0)}\left[e^{JX_j/\sqrt{N}}\right] + \sum_{k=1}^{\infty}\frac{\alpha^k}{k!}\mathbb{E}_{p(X,\alpha=0)}\left[e^{J\sum_i X_i/\sqrt{N}}\mathcal{P}_k|_{\alpha=0}\right]\right]. \tag{2.22}$$

In the limit $\alpha \to 0$, the X_j become independent, and we have

$$\lim_{\alpha\to 0} W_\phi[J] = \sum_j \log\mathbb{E}_{p(X_j)}\left[e^{JX_j/\sqrt{N}}\right] = \sum_j \lim_{\alpha\to 0} W_{X_j/\sqrt{N}}[J], \tag{2.23}$$

where ϕ is now a sum of N independent variables X_j, and its CGF is the sum of CGFs of $X_j/\sqrt{N}$, as expected; details of the calculations are in appendix C.

We have now discussed two mechanisms that result in non-Gaussianities: $1/N$ corrections and independence breaking. While one can use either or both of these mechanisms to generate and control non-Gaussianities, more caution is required to use independence breaking alone, at infinite N. This is because the non-Gaussianities that are generated by independence breaking might depend on N as well as α. For example, if the leading corrections to higher cumulants κ_r^ϕ scale as αN^{a_r} with $a_r < 0$ for all $r > 2$, ϕ will be Gaussian regardless of independence breaking. While if $a_r > 0$, κ_r^ϕ will diverge, which is undesirable. In the following, we will present an example where $a_r = 0$ for all r and the non-Gaussianities are generated by independence breaking alone.

2.2.1. Example: independence breaking at infinite N

Let us provide an example of independence breaking non-Gaussianities that persist in the $N \to \infty$ limit, showing how one can control higher cumulants by adjusting the correlations between random variables. Consider the normalized sum of N random variables,

$$\phi = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i, \tag{2.24}$$

where X_i is the product of two random variables a_i and h_i,

$$X_i = a_i h_i. \tag{2.25}$$

This architecture can be interpreted as the last layer of a fully connected NN, where h_i are the outputs of the neurons in the previous layer, $a_i/\sqrt{N}$ are the weights, and ϕ is the output. First, let us consider the simple case where a_i and h_i are independent, Gaussian random variables[7],

$$P(\vec{a},\vec{h}) = P_{\text{ind}}(\vec{a},\vec{h}) = (2\pi\sigma_a\sigma_h)^{-N}\exp\left(-\frac{1}{2\sigma_a^2}\sum_{i=1}^{N}a_i^2 - \frac{1}{2\sigma_h^2}\sum_{i=1}^{N}h_i^2\right),$$

where σ_a and σ_h are positive and finite. Since a_i and h_i are independent, so are X_i. The CLT applies and ϕ is Gaussian.

Next, we will perturb $P(\vec{a},\vec{h})$ to break independence. To that end, we introduce an auxiliary random variable H and define,

$$\begin{aligned} P(\vec{a},\vec{h},H) &= P_{\text{ind}}(\vec{a},\vec{h},H) \\ &= (2\pi\sigma_a\sigma_h)^{-N}(\sqrt{2\pi}\sigma_h)^{-1}\exp\left(-\frac{1}{2\sigma_a^2}\sum_{i=1}^{N}a_i^2 - \frac{1}{2\sigma_h^2}\sum_{i=1}^{N}h_i^2 - \frac{1}{2\sigma_h^2}H^2\right), \end{aligned} \tag{2.26}$$

where we set the standard deviation of H to σ_h, for simplicity. We then define a correction term,

$$P_{\text{corr}}(\vec{a},\vec{h},H) = P_{\text{ind}}(\vec{a},\vec{h},H)\cdot\exp\left(-\frac{1}{2\sigma_h^2}\sum_{i=1}^{N}(h_i - H)^2\right). \tag{2.27}$$

Finally, putting these together we define,

$$P(\vec{a},\vec{h},H;\alpha) = (1-\alpha)P_{\text{ind}}(\vec{a},\vec{h}) + \alpha P_{\text{corr}}(\vec{a},\vec{h}). \tag{2.28}$$

[7] The word 'Gaussian' happens to appear many times in this example. To clarify: though a and h are both Gaussian by construction, ah is not, and ϕ is Gaussian in the CLT limit.

When $\alpha = 0$, the second term vanishes and both a_i and h_i are independent. As we turn on $\alpha > 0$, the a_i remain independent, but correlations are induced between the h_i through a direct coupling to H in $P_{\text{corr}}(\vec{a}, \vec{h})$.

To quantify the non-Gaussianity of ϕ as a function of α, we compute the CGF,

$$\begin{aligned} W_\phi[J] &= \log \mathbb{E}[e^{J\sum_i X_i/\sqrt{N}}] \\ &= \log \int \prod_{i=1}^{N} \mathrm{d}a_i \, \mathrm{d}x_i \, P(\vec{a}, \vec{h}, H; \alpha) e^{J\sum_i a_i h_i/\sqrt{N}}. \end{aligned} \tag{2.29}$$

As $P(\vec{a}, \vec{h}, H; \alpha)$ is Gaussian, (2.29) can be evaluated analytically to give

$$W_\phi[J] = \log \left[(1-\alpha) \left(\frac{N}{N - J^2 \sigma_a^2 \sigma_h^2} \right)^{N/2} + \alpha \left(\frac{N^{N/2} (N - J^2 \sigma_a^2 \sigma_h^2)^{\frac{1-N}{2}}}{\sqrt{N - (N+1) J^2 \sigma_a^2 \sigma_h^2}} \right) \right]. \tag{2.30}$$

The odd cumulants vanish, as the ϕ ensemble has a $\mathbb{Z}_2$ symmetry $\phi \to -\phi$ (due to evenness of $P(\mathrm{a})$), while the even cumulants κ_r^ϕ can be computed by taking derivatives of $W_\phi[J]$. For example, the second and the fourth cumulants are

$$\kappa_2^\phi = \sigma_a^2 \sigma_h^2 (1+\alpha), \tag{2.31}$$

$$\kappa_4^\phi = \sigma_a^4 \sigma_h^4 \left(9\alpha - 3\alpha^2 + \frac{6 + 12\alpha}{N} \right). \tag{2.32}$$

In the limit $N \to \infty, \alpha \to 0$, the second cumulant is finite while all higher cumulants vanish, and ϕ is Gaussian as expected. At finite $\alpha > 0$, all even cumulants are finite and in general nonzero. The ability to tune α thus allows one to control the degree of non-Gaussianity of ϕ. Note that breaking independence in the large N limit is not a particularly efficient way to sample from a non-Gaussian distribution of a single variable.

2.3. Connected correlators in NN-FT

We wish to establish that the ideas exemplified above—that non-Gaussianities may arise via finite-N corrections or independence breaking—generalize to continuum NN-FT.

In outline, one may think of this conceptually as passing from a single random variable ϕ (0d FT) to a discrete number of random variables ϕ_i (lattice FT), and finally to a continuous number of random variables $\phi(x)$ (continuum FT), where $x \in \mathbb{R}^d$. This is a textbook procedure in the context of the function-space path integral. Here we wish to instead emphasize the general procedure and parameter space perspective.

Consider the case that the continuum field $\phi(x)$ is built out of neurons $h_i(x)$ [10] as

$$\phi(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_i(x). \tag{2.33}$$

If the $h_i(x)$ are independent, the CLT states that $\phi(x)$ is Gaussian in the limit $N \to \infty$. This is the essence of the NNGP correspondence.

Motivated by the single variable case, we will study non-Gaussianities arising from both finite-N corrections and breaking of the independence condition. The CGF of $\phi(x)$ is

$$W_\phi[J] = \log Z_\phi[J] = \sum_{r=1}^{\infty} \int \prod_{i=1}^{r} \mathrm{d}^d x_i \, \frac{J(x_1) \ldots J(x_r)}{r!} \, G_c^{(r)}(x_1, \ldots, x_r), \tag{2.34}$$

where we have performed a series expansion in terms of the cumulants, a.k.a. the connected correlation functions $G_c^{(r)}$ of ϕ. This is a straightforward generalization of (2.5) to the continuum. When the odd-point functions vanish the connected four-point function is

$$G_c^{(4)}(x_1, \ldots, x_4) = G^{(4)}(x_1, \ldots, x_4) - (G^{(2)}(x_1, x_2) G^{(2)}(x_3, x_4) + 2 \text{ perms}), \tag{2.35}$$

which will capture leading-order non-Gaussianities in many of our examples.

In the following, we will quantify non-Gaussianities in terms of non-vanishing cumulants, as well as directly in the action via an Edgeworth expansion.

2.3.1. Finite-N corrections with independent neurons

We first study non-Gaussianities in the case where the neurons $h_i(x)$ are i.i.d. but N is finite, e.g. single hidden layer networks, shown in [42]. We can express the CGF (2.34) in terms of the connected correlation functions of the neurons,

$$W_{\phi(x)}[J] = \log \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int \mathrm{d}^d x J(x) h_i(x)\right)\right] = \sum_{r=1}^{\infty}\int\prod_{i=1}^{r}\mathrm{d}^d x_i \, \frac{J(x_1)\cdots J(x_r)}{r!}\,\frac{G^{(r)}_{c,h_i}(x_1,\cdots,x_r)}{N^{r/2-1}}. \tag{2.36}$$

This result relies on the fact that for independent h_i, the expectation of the product is the product of the expectations, which turns the first expression into a sum on neuron CGFs. For identically distributed neurons the sum gives a factor of N, and the normalization $1/\sqrt{N}$ gives the r-dependent N-scaling. This result lets us express the connected correlators of $\phi(x)$ in terms of the connected correlators of $h_i(x)$,

$$G^{(r)}_{c}(x_1,\cdots,x_r) = \frac{G^{(r)}_{c,h_i}(x_1,\cdots,x_r)}{N^{r/2-1}}. \tag{2.37}$$

This N-scaling implies

$$\lim_{N\to\infty} G^{(r>2)}_{c}(x_1,\cdots,x_r) = 0, \tag{2.38}$$

establishing Gaussianity in the limit.

2.3.1.1. Examples: single layer Cos-net and ReLU-net

We will now consider two single hidden layer architectures with finite N and i.i.d. parameters. While the methods we describe in this section can be employed to study NN with arbitrary depth $L > 1$, inducing statistical correlations among neurons [42], single hidden layer architectures suffice to demonstrate their utility.

2.3.1.2. ReLU-net

First, we will consider an architecture with a single hidden layer and ReLU activation functions. As ReLU activations are ubiquitous in ML applications, this is a natural example to study. Consider

$$\phi(x) = W^1_i R(W^0_{ij}x_j) \;\text{ where } R(z) = \begin{cases} z, \text{ for } z \geqslant 0 \\ 0, \text{ otherwise} \end{cases}, \tag{2.39}$$

with $d = d_{\text{out}} = 1$, $W^0 \sim \mathcal{N}(0, \frac{\sigma^2_{W_0}}{d})$, $W^1 \sim \mathcal{N}(0, \frac{\sigma^2_{W_1}}{N})$.

We compute the two-point function in the parameter space description (1.2) to obtain

$$G^{(2)}_{c,\text{ReLU}}(x,y) = \frac{\sigma^2_{W_0}\sigma^2_{W_1}}{2}\left(R(x)R(y) + R(-x)R(-y)\right), \tag{2.40}$$

which has a factorized structure in the terms that one might call bi-local: the function depends independently on x and y, regardless of any relation between them. This result is exact and does not receive $1/N$ corrections. Non-Gaussianities induced by $1/N$ corrections manifest as a nonzero 4-pt connected correlation function,

$$G^{(4)}_{c,\text{ReLU}}(x_1,x_2,x_3,x_4) = \frac{1}{N}\left(\frac{15\sigma^4_{W_0}\sigma^4_{W_1}}{4d^2}\left(\sum_{j=\pm1} R(jx_1)R(jx_2)R(jx_3)R(jx_4)\right) - \frac{\sigma^4_{W_0}\sigma^4_{W_1}}{4d^2}\left(\sum_{\mathcal{P}(abcd)}\sum_{j=\pm1} R(jx_a)R(jx_b)R(-jx_c)R(-jx_d)\right)\right). \tag{2.41}$$

As expected, $G^{(4)}_{c,\text{ReLU}}(x_1,x_2,x_3,x_4)$ scales as $1/N$.

2.3.1.3. Cos-net

Next, let us study a single hidden layer network with cosine activation functions. The NN-FT associated to Cos-net (and its generalizations) is Euclidean invariant [10], which is interesting on physical grounds, e.g. to satisfy one of the OS axioms to establish an NN-QFT. Euclidean invariance may be established using the mechanism of [8] for determining symmetries from parameter space correlators, which absorbs symmetry transformations into parameter redefinitions, yielding invariant correlators when the relevant parameter distributions are invariant under the symmetry.

Cos-net was defined in [10], where its 2-point function and connected 4-point function were also computed. The architecture is

$$\phi(x) = W_i^1 \cos(W_{ij}^0 x_j + b_i^0) \tag{2.42}$$

where $W^1 \sim \mathcal{N}(0, \sigma_{W_1}^2/N)$, $W^0 \sim \mathcal{N}(0, \sigma_{W_0}^2/d)$, and $b^0 \sim \text{Unif}[-\pi, \pi]$. As before, the correlation functions are computed in parameter space (1.2). The 2-pt function

$$G_{c,\text{Cos}}^{(2)}(x_1, x_2) = \frac{\sigma_{W_1}^2}{2} e^{-\frac{1}{2d}\sigma_{W_0}^2 (\Delta x_{12})^2} \tag{2.43}$$

is manifestly translation invariant, with $\Delta x_{12} = x_1 - x_2$. The 4-pt correlation function is

$$G_{c,\text{Cos}}^{(4)}(x_1, x_2, x_3, x_4) = \frac{\sigma_{W_1}^4}{8N} \sum_{\mathcal{P}(abcd)} \left(3e^{-\frac{\sigma_{W_0}^2 (\Delta x_{ab} + \Delta x_{cd})^2}{2d}} - 2e^{-\frac{\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2 \right)}{2d}} \right), \tag{2.44}$$

where $\Delta x_{ij} := x_i - x_j$ and $\mathcal{P}(abcd)$ denotes the three independent ways of drawing pairs $(x_a, x_b), (x_c, x_d)$ from the list of external vertices (x_1, x_2, x_3, x_4).

We see the manifest Euclidean invariance of these correlators, and that non-Gaussianities are encoded in $G_{c,\text{Cos}}^{(4)}$ as a $1/N$ corrections.

2.3.2. Generalized connected correlators from independence breaking

We now wish to generalize our theories and connected correlators to including the possibility that non-Gaussianities arise not only from $1/N$-corrections, but also from independence breaking, e.g. by developing correlations between the neurons $h_i(x)$. Previously, [10, 42] studied mixed non-Gaussianities at finite N and statistical correlations among neurons.

Generalizing our approach from section 2.2, we parameterize breaking of statistical independence by promoting the distribution of neurons $P(h)$ to depend on a vector of hyperparameters $\vec{\alpha} \in \mathbb{R}^q$, $P(h; \vec{\alpha})$.

Since independence is necessary for Gaussianity via the CLT, and we will sometimes wish to perturb around the Gaussian fixed point, we require

$$P(h; \vec{\alpha} = \vec{0}) = \prod_i P(h_i), \tag{2.45}$$

where the hyperparameter vector $\vec{\alpha}$ must be chosen as part of the architecture definition. From this expression, the neurons become independent when $\vec{\alpha} = 0$.

For a general $P(h; \vec{\alpha})$, the CGF is

$$W_\phi[J] = \log \left[\int \left(\prod_{i=1}^N Dh_i \right) P(h; \vec{\alpha}) e^{\frac{1}{\sqrt{N}} \sum_{i=1}^N \int \mathrm{d}x h_i(x) J(x)} \right]. \tag{2.46}$$

For small values of α, we can expand $P(h; \vec{\alpha})$,

$$P(h; \vec{\alpha}) = P(h; \vec{\alpha} = 0) + \sum_{r=1}^{\infty} \sum_{s_1, \cdots, s_r = 1}^{q} \frac{\alpha_{s_1} \cdots \alpha_{s_r}}{r!} \partial_{\alpha_{s_1}} \cdots \partial_{\alpha_{s_r}} P(h; \vec{\alpha}) \Big|_{\vec{\alpha}=0}. \tag{2.47}$$

Analogous to the single variable case, we define

$$\mathcal{P}_{r, \{s_1, \cdots, s_r\}} := \frac{1}{P(h; \vec{\alpha})} \partial_{\alpha_{s_1}} \cdots \partial_{\alpha_{s_r}} P(h; \vec{\alpha}) \tag{2.48}$$

satisfying the recursion relation

$$\mathcal{P}_{r+1, \{s_1, \cdots, s_{r+1}\}} = \frac{1}{r+1} \sum_{\gamma=1}^{r+1} (\mathcal{P}_{1, s_\gamma} + \partial_{\alpha_{s_\gamma}}) \mathcal{P}_{r, \{s_1, \cdots, s_{r+1}\} \setminus s_\gamma}. \tag{2.49}$$

Finally, we can expand (2.34) in $\vec{\alpha}$,

$$W_\phi[J] = \log\left[e^{W_{\phi,\vec{\alpha}=0}[J]} + \sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q}\frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^dx h_i(x)J(x)}\cdot\mathcal{P}_{r,\{s_1,\cdots,s_r\}}\big|_{\vec{\alpha}=0}\right]\right], \tag{2.50}$$

where $W_{\phi,\vec{\alpha}=0}[J]$ is given in (2.36). This form of $W_\phi[J]$ makes it clear how one can tune N and $\vec{\alpha}$ to generate and manipulate non-Gaussianities; for details see appendix C.

For appropriately small independence breaking hyperparameter $\vec{\alpha}$, and other attributes of the architecture, the ratio of second term to first term in the logarithm of (2.50) is small. In such cases, one can approximate (2.50) using Taylor series expansion $\log(1+x)\approx x$ around $x=0$. The CGF becomes

$$W_\phi[J] = W_{\phi,\vec{\alpha}=0}[J] + \sum_{s=1}^{q}\frac{\alpha_s}{e^{W_{\phi,\vec{\alpha}=0}[J]}}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^dx h_i(x)J(x)}\cdot\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right], \tag{2.51}$$

and the cumulants

$$\begin{aligned} G_c^{(r)}(x_1,\cdots,x_r) &= \frac{\partial^r W_\phi[J]}{\partial J(x_1)\cdots\partial J(x_r)}\Big|_{J=0}, \\ &= G_c^{(r),\text{i.i.d.}} + \vec{\alpha}\cdot\Delta G_c^{(r)} + O(\vec{\alpha}^2). \end{aligned} \tag{2.52}$$

are proportional to $\vec{\alpha}$ at the leading order. The leading order expression in $\vec{\alpha}$ is evaluated in (C.21).

2.3.2.1. Example: single layer Cos-net

Let us exemplify the non-Gaussianities generated by statistical independence breaking of a single layer Cos-net architecture given in (2.42). We can break this independence by modifying the distribution from which the weights W_{ij}^0 (an $N\times d$ matrix) are sampled

$$P(W^0) = c\exp\left[-\sum_{i,j}\left(\frac{d}{2\sigma_{W_0}^2}(W_{ij}^0)^2 + \frac{\alpha_{\text{IB}}}{N^2}\sum_{i_1,j_1,i_2,j_2}(W_{i_1j_1}^0)^2(W_{i_2j_2}^0)^2\right)\right], \tag{2.53}$$

where c is a normalization constant. The rotational invariance preserving term $\frac{\alpha_{\text{IB}}(\text{Tr}(W^{0T}W^0))^2}{N^2}$ introduces mixing between the weights W_{ij}^0 and parametric independence is explicitly broken. The degree of independence breaking can be controlled by tuning α_{IB}.

We wish to compute the connected correlation functions to quantify the non-Gaussianities generated by independence breaking. In general, this is a difficult problem. However, when $\alpha_{\text{IB}}\ll 1$, we can perform a perturbative expansion in α_{IB}. Setting $d=1$ for simplicity, we obtain

$$G_{c,\text{Cos}}^{(2)}(x_1,x_2) = \frac{\alpha_{\text{IB}}\sigma_{W_0}^4\sigma_{W_1}^2 e^{-\frac{\sigma_{W_0}^2(\Delta x_{12})^2}{2}}}{2N}\left[-\frac{\left(1-5\sigma_{W_0}^2(\Delta x_{12})^2+\sigma_{W_0}^4(\Delta x_{12})^4\right)}{N}+\sigma_{W_0}^2(\Delta x_{12})^2\right], \tag{2.54}$$

$$\begin{aligned} G_{c,\text{Cos}}^{(4)}(x_1,\cdots,x_4) = G_{\substack{c,\text{Cos}\\ \text{i.i.d.}}}^{(4)}(x_1,\cdots,x_4) + \frac{\alpha_{\text{IB}}\sigma_{W^0}^4\sigma_{W^1}^4}{8N^2}\sum_{\mathcal{P}(abcd)}\Big[6-\big(2\sigma_{W^0}^2(\Delta x_{ab}^2+\Delta x_{cd}^2) \\ +2\sigma_{W^0}^4\Delta x_{ab}^2\Delta x_{cd}^2\big)e^{-\frac{\sigma_{W_0}^2}{2}(\Delta x_{ab}^2+\Delta x_{cd}^2)}+\big(3+3\sigma_{W^0}^2(\Delta x_{ab}+\Delta x_{cd})^2\big)e^{-\frac{\sigma_{W_0}^2}{2}(\Delta x_{ab}+\Delta x_{cd})^2}\Big], \end{aligned} \tag{2.55}$$

to leading order in α_{IB}, where $G_{\substack{c,\text{Cos}\\ \text{i.i.d.}}}^{(4)}(x_1,\cdots,x_4)$ is obtained at $d=1$ from (2.44). Non-Gaussianities at finite N, and $\alpha_{\text{IB}}\neq 0$ still preserve the translation invariance of the 2nd and 4th cumulants of Cos-net architecture.

We refer the reader to appendix B.2 for details, where we also compute leading order non-Gaussian corrections to first two cumulants in a single hidden layer Gauss-net at $\alpha_{\text{IB}}\neq 0$, finite N, for $d=1$.

3. Computing actions from connected correlators

In section 2 we systematically studied non-Gaussianities in NN FT by parametrically violating two assumptions of the CLT: infinite-N and independence. The study was performed at the level of connected correlators, rather than actions, because every NN-FT admits a parameter space description of connected correlators, even if an action is not known.

In this section we will develop these techniques for calculating actions from connected correlators, including in terms of Feynman diagrams in which the connected correlators are vertices. More specifically:

- **Field Density from Connected Correlators: Edgeworth Expansion.** In section 3.1 we review how knowledge of the cumulants of a single random variable may be used to approximate its probability density, and then we generalize to the FT case, which has a continuum of random variables. This gives an expression for $P[\phi] = \exp(-S[\phi])$ in terms of connected correlation functions. We present an explicit example in the case of a single variable.
- **Computing the Action with Feynman Diagrams.** Given the Edgeworth expansion, we develop a method to compute the action perturbatively via Feynman diagrams, which becomes clear due to a formal similarity between the Edgeworth expansion and the partition function of a FT. This is a result that is applicable to general FT.
- **NN FT Actions.** In section 3.3 we specify the analysis of section 3.2 to the case of NN FT. We derive the leading order form of the action for the case of non-Gaussianities induced either by $1/N$-corrections or independence breaking.
- **NN FT Examples.** In section 3.4 we derive the leading-order action in $1/N$ for concrete NN architectures.

3.1. Field density from connected correlators: Edgeworth expansion

The Edgeworth expansion from statistics (see e.g. [43] for a textbook statistics description and [11] for an ML study) can be used to construct the probability density from the cumulants. The key observation which allows the Edgeworth expansion to be applied in a FT is that the normal relation for the generating function in terms of the action

$$e^{W[J]} = \int \mathrm{d}\phi \, e^{-S[\phi]+J\phi} \tag{3.1}$$

can be inverted to express the action in terms of the generating functional. Adding a source term in the exponent, mapping $J \to iJ$ and integrating over J, we have

$$\int \mathrm{d}J e^{W[iJ]-iJ\phi} = \int \mathrm{d}J e^{-iJ\phi} \int \mathrm{d}\phi' e^{-S[\phi']+iJ\phi'} = e^{-S[\phi]} \tag{3.2}$$

where

$$\int \mathrm{d}J e^{iJ(\phi'-\phi)} = \delta[\phi'-\phi] \tag{3.3}$$

has been used. Deforming the J integration contour back to real J then results in

$$P[\phi] = e^{-S[\phi]} = \int \mathrm{d}J e^{W[J]-J\phi}, \tag{3.4}$$

This gives the probability density and action in terms of $W[J]$. This result can also be thought of as arising from an inverse Fourier transform of the characteristic function.

Then to apply the Edgeworth expansion for a single random variable ϕ, we write $W[J]$ in terms of cumulants

$$W[J] = \sum_{r=1}^{\infty} \frac{\kappa_r}{r!} J^r, \tag{3.5}$$

which lets us write

$$\begin{aligned} P[\phi] &= \exp\left[\sum_{r=3}^{\infty} \frac{\kappa_r}{r!}(-\partial_\phi)^r\right] \int \mathrm{d}J e^{\kappa_2 \frac{J^2}{2} + \kappa_1 J - J\phi}, \\ &= \exp\left[\sum_{r=3}^{\infty} \frac{\kappa_r}{r!}(-\partial_\phi)^r\right] e^{-\frac{(\phi-\kappa_1)^2}{2\kappa_2}}, \end{aligned} \tag{3.6}$$

where the Gaussian integral has been performed by mapping $J \to iJ$ (alternatively, working with the characteristic function the whole time) and we have neglected the normalization factor. We have an expression for the density P_ϕ as an expansion around the Gaussian with mean κ_1 and variance κ_2.

The result may be extended to the FT case, where ϕ is replaced by $\phi(x)$, a continuum of mean free random variables. Then the relation is

$$e^{-S[\phi]} = \frac{1}{Z}\exp\left(\sum_{r=3}^{\infty}\frac{(-1)^r}{r!}\int\prod_{i=1}^{r}\mathrm{d}^d x_i\, G_c^{(r)}(x_1,\ldots,x_r)\frac{\delta}{\delta\phi(x_1)}\cdots\frac{\delta}{\delta\phi(x_r)}\right)e^{-S_G[\phi]}, \tag{3.7}$$

where the GP action S_G is defined as

$$S_G[\phi] = \frac{1}{2}\int \mathrm{d}^d x_1 \mathrm{d}^d x_2\, \phi(x_1) G_c^{(2)}(x_1,x_2)^{-1}\phi(x_2), \tag{3.8}$$

To the extent that there is a perturbative ordering to the correlators through some expansion parameter (such as $\frac{1}{N}$ or independence breaking), this expression can be evaluated perturbatively to systematically construct an action from the cumulants[8].

3.1.1. 1D example: sum of N uniform random variables

Let us demonstrate the Edgeworth expansion in a simple example. Consider the standardized sum of N i.i.d. random variables sampled from a uniform distribution

$$\phi = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i, \quad X_i \sim \mathrm{Unif}(-1/2,1/2)\ \forall i. \tag{3.10}$$

The cumulants of X_i are

$$\kappa_1^{X_i} = 0, \tag{3.11}$$

$$\kappa_r^{X_i} = \frac{B_r}{r} \text{ for } r \geqslant 2, \tag{3.12}$$

where B_r is the rth Bernoulli number[9]. Plugging this into (2.15), the cumulants of ϕ are

$$\kappa_r^{\phi} = \frac{B_r}{rN^{r/2-1}}. \tag{3.13}$$

At finite N, the cumulants $\kappa_{r>2}^{\phi}$ are nonzero and ϕ is non-Gaussian. Using these cumulants, we can write down the probability distribution function of ϕ via an Edgeworth expansion,

$$\begin{aligned} P_\phi &= \frac{1}{Z}\exp\left[\sum_{r=3}^{\infty}\frac{\kappa_r^{\phi}}{r!}(-\partial_\phi)^r\right]e^{-\phi^2/2\kappa_2^{\phi}} \\ &= \frac{1}{Z}\exp\left[\sum_{r=3}^{\infty}\frac{B_r}{r!rN^{r/2-1}}(-\partial_\phi)^r\right]e^{-\phi^2/B_2} \end{aligned} \tag{3.14}$$

Truncating the sum at $r=4$, expanding the exponential, and keeping terms up to $O(1/N)$ we get

$$\begin{aligned} P_\phi &= \frac{1}{Z}\left[1+\kappa_4^{\phi}\left(\frac{1}{8(\kappa_2^{\phi})^2}-\frac{1}{4(\kappa_2^{\phi})^3}\phi^2+\frac{1}{24(\kappa_2^{\phi})^4}\phi^4\right)+O(1/N^{3/2})\right]e^{-\phi^2/2\kappa_2^{\phi}}, \\ &= \frac{1}{Z'}\exp\left[-\left(\frac{1}{2\kappa_2^{\phi}}+\frac{\kappa_4^{\phi}}{4(\kappa_2^{\phi})^3}\right)\phi^2+\frac{\kappa_4^{\phi}}{24(\kappa_2^{\phi})^4}\phi^4+O(1/N^{3/2})\right], \\ &= \frac{1}{Z'}\exp\left[\left(-6+\frac{18}{5N}\right)\phi^2-\frac{36}{5N}\phi^4+O(1/N^{3/2})\right], \end{aligned} \tag{3.15}$$

where on the second line we absorbed the constant term into the normalization constant Z'. At order $O(N^0)$, the exponent in (3.15) is quadratic and ϕ is Gaussian distributed. Gaussianity is then broken by a quartic interaction at order $O(1/N)$.

[8] In the finite-dimensional version of the Edgeworth expansion, it is sometimes convenient to further express the powers of derivatives in terms of probabilist's Hermite polynomials using

$$(-\partial_x)^r e^{-\frac{x^2}{2}} =: H_r(x)\, e^{-\frac{x^2}{2}}, \tag{3.9}$$

In the FT case, using Hermite polynomials provides no obvious advantage.

[9] B_r vanishes for odd $r \geqslant 2$. First few nonzero Bernoulli numbers are: $B_2 = \frac{1}{6}, B_4 = -\frac{1}{30}, B_6 = \frac{1}{42}$.

Table 1. The Edgeworth expansion for $P[\phi]$ and the interaction expansion of $Z[J]$ are formally related by a change of variables, given here up to constant factors. Due to this relationship, non-local couplings and connected correlators may both be computed by appropriate connected Feynman diagrams.

	Field picture	Source picture
Field	$\phi(x)$	$J(x)$
CGF	$W[J] = \log(Z[J])$	$S[\phi] = -\log(P[\phi])$
Cumulant	$G_c^{(r)}(x_1, \ldots, x_r)$	$g_r(x_1, \ldots, x_r)$

It is worth noting that the cumulants of ϕ are given by simple closed form expressions, see equation (3.13), while P_ϕ involves a perturbative expansion in $1/N$. This is in contrast to weakly coupled FT, where we often start from a simple action expressed in closed form and calculate the connected correlation functions via a perturbative expansion in the coefficients of interaction terms.

3.2. Computing the action with Feynman diagrams

In a FT a powerful tool for organizing a perturbation expansion is with Feynman diagrams. Just as Feynman diagrams can be used to compute the cumulants perturbatively in an expansion parameter from an action, they can also be used to compute the action perturbatively from the cumulants. To understand the derivation, recall the expression for the partition function

$$e^{W[J]} = Z[J] = c' \exp\left(\sum_{r=3}^{\infty} \int \prod_{i=1}^{r} \mathrm{d}^d x_i \, g_r(x_1, \cdots, x_r) \frac{\delta}{\delta J(x_1)} \cdots \frac{\delta}{\delta J(x_r)}\right) e^{-S_0[J]}, \tag{3.16}$$

where we have introduced couplings g_r instead of $g_r/r!$,

$$S_0[J] = \int \mathrm{d}x_1 \mathrm{d}x_2 J(x_1) \Delta(x_1, x_2) J(x_2), \tag{3.17}$$

and $\Delta(x_1, x_2)$ is the free propagator. The expression (3.16) arises by taking the usual expression for the partition function

$$Z[J] = \int D\phi \, e^{-S_{\text{free}}[\phi] - S_{\text{int}} + \int \mathrm{d}^d x J(x)\phi(x)} \tag{3.18}$$

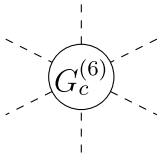
and replacing the ϕ's in the interaction terms

$$S_{\text{int}} = \sum_{r=3}^{\infty} \int \prod_{i=1}^{r} \mathrm{d}^d x_i \, g_r(x_1, \ldots, x_r) \, \phi(x_1) \ldots \phi(x_r) \tag{3.19}$$

by $\delta/\delta J$'s. Pulling the J-derivatives outside of the $\int D\phi$ in (3.18) and performing the Gaussian integral yields (3.16). These manipulations closely mirror the Edgeworth expansion.

The Edgeworth expansion (3.7) is related to the partition function (3.16) by a simple change of variables, given in table 1, which one might think of as a duality map between a field picture and a source picture. This relationship between the Edgeworth expansion and the partition function immediately tells us that the analog of $g_r(x_1, \ldots, x_n)$ are the connected correlation functions $G_c^{(r)}(x_1, \cdots, x_r)$ in (3.7).

We may therefore compute the couplings $g_r(x_1, \ldots, x_n)$ in the same way that we compute the connected correlators $G_c^{(r)}(x_1, \cdots, x_r)$. In a weakly coupled FT, one can compute the connected correlation functions $G_c^{(r)}(x_1, \cdots, x_r)$ in terms of the couplings $g_r(x_1, \cdots, x_r)$ perturbatively via Feynman diagrams. An Edgeworth expansion allows us to do the converse and compute the couplings $g_r(x_1, \cdots, x_r)$ in terms of the connected correlation functions $G_c^{(r)}(x_1, \cdots, x_r)$. The similarity between (3.7) and (3.16) suggests that the terms in the expansion for $g_r(x_1, \cdots, x_r)$ can be represented by Feynman diagrams, whose vertices are connected correlators, e.g.



(3.20)

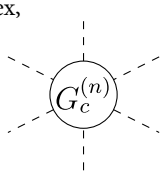
Table 2. Feynman rules for computing g_r from each connected diagram with $G_c^{(n)}$ vertices.

Feynman Rules for $g_r(x_1,\ldots,x_r)$.

1. Internal points associated to vertices are unlabeled, for diagrammatic simplicity. Propagators therefore connect to internal points in all possible ways.
2. For each propagator between z_i and z_j,

$$Z_i ---- Z_j = G_c^{(2)}(z_i,z_j)^{-1}. \tag{3.22}$$

3. For each vertex,

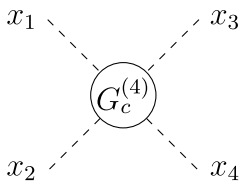


$$= (-1)^n \int d^d y_1 \cdots d^d y_n \, G_c^{(n)}(y_1,\cdots,y_n). \tag{3.23}$$

4. Divide by symmetry factor and insert overall $(-)$.

in the case of a six-point vertex. Notably, the vertex is itself a function and lines enter the n-point vertex at n locations.

To compute the coupling $g_r(x_1,\ldots,x_r)$ in terms of Feynman diagrams, one sums over all connected r-point Feynman diagrams made out of $G_c^{(n)}$ vertices. By convention, we do not label internal points on the vertices, in order to simplify the combinatorics. For instance, the four-point coupling $g_4(x_1,\ldots,x_4)$ has a diagram

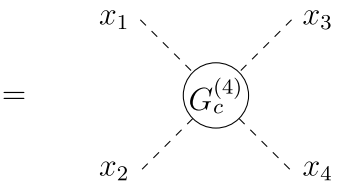


(3.21)

where it is to be understood that connections to internal points in a vertex appear in all possible combinations. Analytic expressions may be obtained from the diagrams via the Feynman rules given in table 2. If $G_c^{(2)}(x_i,y_j)^{-1} = \frac{\delta^2 S_G}{\delta\phi(x_i)\delta\phi(y_j)}$ involves differential operators, it can be evaluated by Fourier transformation, see appendix D

As an example, let us compute a contribution to the quartic coupling $g_4(x_1,x_2,x_3,x_4)$ from a $G_c^{(4)}$ vertex

$$\begin{aligned} g_4(x_1,\ldots,x_4) = -\frac{1}{4!}\Big[\int \mathrm{d}y_1\mathrm{d}y_2\mathrm{d}y_3\mathrm{d}y_4\, G_c^{(4)}(y_1,y_2,y_3,y_4)\, G_c^{(2)}(y_1,x_1)^{-1} G_c^{(2)}(y_2,x_2)^{-1} \\ \times G_c^{(2)}(y_3,x_3)^{-1} G_c^{(2)}(y_4,x_4)^{-1} + \text{perms}\Big] + \ldots \end{aligned} \tag{3.24}$$



(3.25)

where the dots represent contributions from other diagrams, and 'perms' represents other diagrams from permutations over internal points. A combinatoric factor of 4! from summing over internal points cancels out the prefactor $1/4!$ from Edgeworth expansion.

The Edgeworth expansion (3.7) involves an infinite sum. Correspondingly, computing $g_r(x_1,\cdots,x_r)$ requires summing over infinitely many Feynman diagrams. When all but finitely many terms in the expansion are parametrically suppressed, the expansion can be truncated at finite order to provide an approximation of $g_r(x_1,\cdots,x_r)$. We will apply these rules to concrete examples later in this section and demonstrate how approximations to $g_r(x_1,\cdots,x_r)$ can be obtained systematically.

While our focus is on NN FT, we emphasize that Edgeworth expansions can be utilized in any FT where the connected correlation functions are known, and the expansion in (3.7) is not divergent.

3.2.1. Example: non-local ϕ^4 theory

Aside from any application in NN-FT, it is interesting to study the self-consistency of the Edgeworth expansion. We do so in a famous case, ϕ^4 theory, generalized to the case of non-local quartic interactions, in order to demonstrate the ability of the Edgeworth method to handle non-locality. Consider the action

$$S[\phi] = \int \mathrm{d}^d x_1 \mathrm{d}^d x_2 \, \frac{1}{2}\phi(x_1) G^{(2)}_{G,\phi}(x_1,x_2)^{-1}\phi(x_2) + \frac{1}{4!}\int \mathrm{d}^d x_1 \mathrm{d}^d x_2 \mathrm{d}^d x_3 \mathrm{d}^d x_4 \lambda(x_1,x_2,x_3,x_4)\phi(x_1)\phi(x_2)\phi(x_3)\phi(x_4), \tag{3.26}$$

where $G^{(2)}_{G,\phi}(x_1,x_2)^{-1}$ and $\lambda(x_1,x_2,x_3,x_4)$ are both totally symmetric, and $G^{(2)}_{G,\phi}(x_1,x_2)^{-1}$ is the operator in the free action $S_G[\phi]$. We denote the free propagator $D(x_1,x_2)$ so that $\int d^d x' \, G^{(2)}_{G,\phi}(x_1,x_2)^{-1} D(x',x_2) = \delta^d(x_1 - x_2)$. We can then expand $G^{(2)}_c(x_1,x_2)$ in $\lambda(x_1,x_2,x_3,x_4)$, and at leading order,

$$G^{(2)}_c(x_1,x_2) = D(x_1,x_2) + \frac{1}{2}\int \mathrm{d}^d y_1 \cdots \mathrm{d}^d y_4 \lambda(y_1,y_2,y_3,y_4) D(x_1,y_1) D(y_2,y_3) D(y_4,x_2), \tag{3.27}$$

where the $\frac{1}{2}$ is a symmetry factor. Similarly,

$$G^{(4)}_c(x_1,\cdots,x_4) = \int \mathrm{d}^d x_1' \cdots \mathrm{d}^d x_4' \, \lambda(x_1',x_2',x_3',x_4') D(x_1,x_1') D(x_2,x_2') D(x_3,x_3') D(x_4,x_4') + O(\lambda^2). \tag{3.28}$$

There are no other connected correlators that have contributions at $O(\lambda)$. To perform an Edgeworth expansion, we first need to write down the inverse propagator,

$$G^{(2)}_c(x_1,x_2)^{-1} = G^{(2)}_{G,\phi}(x_1,x_2)^{-1} - \frac{1}{2}\int \mathrm{d}^d x_3 \mathrm{d}^d x_4 \lambda(x_1,x_2,x_3,x_4) D(x_3,x_4) + O(\lambda^2). \tag{3.29}$$

Given (3.29), it is easy to verify that

$$\int \mathrm{d}x' G^{(2)}_c(x_1,x')^{-1} G^{(2)}_c(x',x_2) = \delta(x_1 - x_2) + O(\lambda^2). \tag{3.30}$$

At this point, let us introduce a shorthand notation to improve readability, rewriting $\int \mathrm{d}^d x_1 \mathrm{d}^d x_2 \, G^{(2)}_c(x_1,x_2)$, $\int \mathrm{d}^d x_1 \mathrm{d}^d x_2 \, G^{(2)}_c(x_1,x_2)^{-1}$ and $\int \mathrm{d}^d x_1 \cdots \mathrm{d}^d x_4 \, G^{(4)}_c(x_1,\cdots,x_4)$ as,

$$G_{xy} = D_{xy} + \frac{1}{2}\lambda_{1234} D_{1x} D_{23} D_{4y} + O(\lambda^2), \tag{3.31}$$

$$G^{-1}_{xy} = G^{(2)}_{G,\phi}(x,y)^{-1} - \frac{1}{2}\lambda_{xy12} D_{12} + O(\lambda^2), \tag{3.32}$$

$$G_{1234} = \lambda_{1'2'3'4'} D_{1'1} D_{2'2} D_{3'3} D_{4'4} + O(\lambda^2), \tag{3.33}$$

respectively. Finally, we obtain the Edgeworth expansion at $O(\lambda)$ by plugging in (3.29) and (3.28) into (3.7),

$$P[\phi] = \frac{1}{Z}\exp\left(\frac{1}{4!}G_{1234}\delta_1\delta_2\delta_3\delta_4\right)\exp\left(-\frac{1}{2}\phi_x G^{-1}_{xy}\phi_y\right) + O(\lambda^2), \tag{3.34}$$

where $\delta_1 := \delta/\delta\phi(x_1)$. Expanding the first exponential and performing the derivatives we obtain

$$P[\phi] = \frac{1}{Z}\left[1 - \frac{\lambda_{1234}}{8} D_{12} D_{34} - \frac{\lambda_{1234}}{4!}\phi_1\phi_2\phi_3\phi_4\right]\exp\left(-\frac{1}{2}\phi_x G^{(2)}_{G,\phi}(x,y)^{-1}\phi_y\right) + O(\lambda^2), \tag{3.35}$$

with $\lambda_{1234} := \int \mathrm{d}^d x_1 \cdots \mathrm{d}^d x_4 \, \lambda(x_1,\cdots,x_4)$, and $\phi_x := \phi(x)$. The second term does not depend on ϕ and can be absorbed into the normalization factor, resulting in

$$P[\phi] = \frac{1}{Z'}\exp\left(-\frac{1}{2}\phi_x G^{(2)}_{G,\phi}(x_1,x_2)^{-1}\phi_y - \frac{\lambda_{1234}}{4!}\phi_1\phi_2\phi_3\phi_4\right) + O(\lambda^2). \tag{3.36}$$

We have recovered the ϕ^4 action at $O(\lambda)$, as expected.

3.3. General interacting actions in NN-FT

We now study the Edgeworth expansion in NN FT. We will modify the general analysis of the previous section to the case where non-Gaussianities are generated by the two mechanisms we described in section 2, namely, by violating assumptions of the CLT by finite N corrections and independence breaking.

3.3.1. Interactions from $1/N$-corrections

As we discussed in section 2.3.1, non-Gaussianities arising due to $1/N$ corrections result in connected correlation functions that scale as

$$G_c^{(r)}(x_1,\cdots,x_r) \propto \frac{1}{N^{r/2-1}}, \tag{3.37}$$

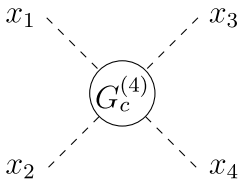
for a single hidden layer network. At large N, the action can be approximated systematically by organizing the Edgeworth expansion in powers of $1/N$, calculating the couplings via Feynman diagrams, and truncating at a fixed order in $1/N$.

To do so, we need to know how the couplings scale with N. We have studied a case in (3.25) where only the even-point correlators are non-zero, and clearly there is a $1/N$ contribution to g_4 from a single $G_c^{(4)}$ vertex; any higher order correlator $G_c^{(r>4)}$ contributes at $1/N^{r/2-1}$ and higher. Consider now contributions to the couplings $g_{r>4}$. There is a tree-level $1/N^{r/2-1}$ contribution from a single $G_c^{(r)}$ vertex and there are $1/N^{n/2-1}$ contributions from a $G_c^{(n>r)}$ vertex with an appropriate number of loops; both are more suppressed than the $1/N$ contribution to g_4. Finally, consider contributions from V number of $G_c^{(n<r)}$ vertices. Forming a connected diagram requires $nV > r$, which implies $V \geqslant 2$ and therefore the contribution is of order $1/N^{\geqslant n-1}$, which is more suppressed than $1/N$ since n begins at 3 in the Edgeworth expansion. Therefore, the single-vertex tree-level contribution to g_4 is the leading contribution in $1/N$.

The quartic coupling $g_4(x_1,x_2,x_3,x_4)$, at leading order in $G_c^{(4)} \propto 1/N$, is

$$\begin{aligned} g_4(x_1,\ldots,x_4) = -\frac{1}{4!}\Bigg[\int \mathrm{d}y_1\mathrm{d}y_2\mathrm{d}y_3\mathrm{d}y_4\, G_c^{(4)}(y_1,y_2,y_3,y_4)\, G_c^{(2)}(y_1,x_1)^{-1} G_c^{(2)}(y_2,x_2)^{-1} \\ \times G_c^{(2)}(y_3,x_3)^{-1} G_c^{(2)}(y_4,x_4)^{-1} + \text{perms}\Bigg] + O\left(\frac{1}{N^2}\right), \end{aligned} \tag{3.38}$$

$$= \tag{3.39}$$



$$+ O\left(\frac{1}{N^2}\right).$$

We may compute this coupling in a NN-FT by first computing $G_c^{(4)}$ in parameter space.

In summary, the leading-order in $1/N$ action for a single layer NN-FT is

$$S = S_G + \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_4\; g_4(x_1,\ldots,x_4)\,\phi(x_1)\ldots\phi(x_4) + O\left(\frac{1}{N^2}\right), \tag{3.40}$$

where g_4 at $O(1/N)$ is given in (3.39), under the assumption that the odd-point functions are zero, as in the architectures of section 3.4.

3.3.2. Interactions from independence breaking

Non-Gaussianities generated via independence breaking alone are qualitatively different than those from $1/N$ corrections.

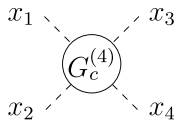
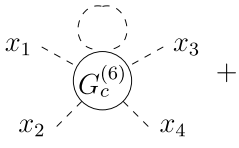
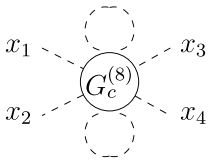
We wish to determine the leading-order action due to independence breaking. Focusing on the case where independence breaking is controlled by a single parameter α for simplicity, it follows from (2.52), that the connected correlation functions scale as

$$G_c^{(r)}(x_1,\cdots,x_r) \propto \alpha \quad \forall r > 2 \tag{3.41}$$

at $N \to \infty$ limit, since the connected correlators $G_c^{(r),\text{free}}|_{r>2}$ of the free theory vanish.

As a result, each coupling $g_r(x_1,\cdots,x_r)$ receives contributions from tree-level diagrams of all connected correlators, at leading order in α. More generally, at any given order in α, there are infinitely many diagrams from all connected correlators to $g_r(x_1,\cdots,x_r)$. For example, the expansion for $g_4(x_1,x_2,x_3,x_4)$ at $O(\alpha)$ includes terms proportional to $G_c^{(2n)}(x_1,\cdots,x_{2n})$ for all $n>1$,

$$g_4(x_1,x_2,x_3,x_4) = -\sum_{n=2}^{\infty}\frac{(-1)^{2n}}{(2n)!}\left[\int \mathrm{d}y_1\cdots\mathrm{d}y_{2n}\, G_c^{(2n)}(y_1,\ldots,y_{2n})\, G_c^{(2)}(y_1,x_1)^{-1}\right.$$
$$\left.\times\, G_c^{(2)}(y_2,x_2)^{-1}\, G_c^{(2)}(y_3,x_3)^{-1}\, G_c^{(2)}(y_4,x_4)^{-1}\prod_{m=5}^{2n-1} G_c^{(2)}(y_m,y_{m+1})^{-1}+\text{perms}\right]+O(\alpha^2), \tag{3.42}$$

$$=(-1)^{2n}\Bigg($$  $$+$$  $$+$$  $$+\cdots\Bigg)+O(\alpha^2), \tag{3.43}$$

where summing over internal points y_i cancels out $\frac{1}{2n!}$ prefactor from each $G_c^{(2n)}$. The terms in the parenthesis constitute an infinite sum.

This structure makes it impossible to systematically approximate $g_r(x_1,\cdots,x_r)$ with a finite number of terms via a perturbative expansion in α, unless some other structure correlates with it. Note that this is a feature of NN FT where non-Gaussianities are generated *only* by independence breaking. Approximation via a finite number of terms would be possible in cases where connected correlation functions scale with both α *and* $1/N$. In the limit of $N\to\infty$, the leading-order in α action for a NN-FT is

$$S = S_G + \sum_{r=4}^{\infty}\int \mathrm{d}^d x_1\ldots\mathrm{d}^d x_r\, g_r(x_1,\ldots,x_r)\,\phi(x_1)\ldots\phi(x_r)+O(\alpha^2), \tag{3.44}$$

where $g_{r>4}$'s are computed similar to (3.43). Such an action can not be approximated by a finite truncation, unless the theory exhibits additional structure.

3.4. Example actions in NN-FT

Next, we exemplify the Feynman rules from section 3.2 in a few single layer NN architecture examples at finite width and i.i.d. parameters, and evaluate the leading order in $1/N$ quartic coupling and NN-FT action. The quartic coupling is

$$g_4(x_1,\cdots,x_4) = -\frac{1}{4!}\left[\int \mathrm{d}^d y_1\cdots\mathrm{d}^d y_4\, G_c^{(4)}(y_1,\cdots,y_4) G_c^{(2)}(y_1,x_1)^{-1}\cdots G_c^{(2)}(y_4,x_4)^{-1}+\text{perms}\right], \tag{3.45}$$

at $O(1/N)$. When $G_c^{(2)}(x_1,y_1)^{-1}$ involves differential operators, we use the methods from appendix D to evaluate g_4.

3.4.1. Single layer Cos-net

Recall the Cos-net architecture introduced earlier, $\phi(x)=W_i^1\cos(W_{ij}^0x_j+b_i^0)$, for $W^1\sim\mathcal{N}(0,\sigma_{W_1}^2/N)$, $W^0\sim\mathcal{N}(0,\sigma_{W_0}^2/d)$, and $b^0\sim\mathrm{Unif}[-\pi,\pi]$. We will consider the case where all parameters are independent and non-Gaussianities arise due to finite N corrections. To evaluate the leading order quartic coupling for this NNFT, let us first compute the inverse propagator $G_{c,\mathrm{Cos}}^{(2)}(x_1,x_2)^{-1}$, starting from the 2-pt function

$$G_{c,\mathrm{Cos}}^{(2)}(x_1,x_2) = \frac{\sigma_{W_1}^2}{2}e^{-\frac{\sigma_{W_0}^2(x_1-x_2)^2}{2d}}, \tag{3.46}$$

and inversion relation $\int \mathrm{d}^d y\, G_{c,\mathrm{Cos}}^{(2)}(x,y)^{-1}\, G_{c,\mathrm{Cos}}^{(2)}(y,z)=\delta^d(x-z)$. Translation invariance of the 2-pt function and delta function constraints $G_{c,\mathrm{Cos}}^{(2)}(x,y)^{-1}$ as a translation invariant operator. Then, performing a Fourier transformation of the 2-pt function and its inverse operator, followed by an inverse Fourier transformation, we obtain

$$G_{c,\mathrm{Cos}}^{(2)}(x,y)^{-1} = \frac{2\sigma_{W_0}^2}{\sigma_{W_1}^2 d}\,e^{-\frac{\sigma_{W_0}^2\nabla_x^2}{2d}}\delta^d(x-y), \tag{3.47}$$

where $\nabla_x^2 := \partial^2/\partial x^2$. Here, we use (D.3) to evaluate the quartic coupling as,

$$g_4^{\text{Cos}}(x_1,\cdots,x_4) = -\int \mathrm{d}^d p_1 \cdots \mathrm{d}^d p_4\, \tilde{G}^{(4)}_{c,\text{Cos}}(p_1,\ldots,p_4)\, \tilde{G}^{(2)}_{c,\text{Cos}}(-p_1)^{-1}\cdots \tilde{G}^{(2)}_{c,\text{Cos}}(-p_4)^{-1} \times e^{-ip_1x_1\cdots - ip_4x_4}, \tag{3.48}$$

where $\tilde{G}^{(4)}_{c,\text{Cos}}(p_1,\cdots,p_4)$ is from (B.8), and $\tilde{G}^{(2)}_{c,\text{Cos}}(-p)^{-1} = \frac{2\sigma_{W_0}}{\sqrt{d}\sigma^2_{W_1}} e^{\frac{dp^2}{2\sigma^2_{W_0}}}$. Using this,

$$g_4^{\text{Cos}}(x_1,x_2,x_3,x_4) = -\frac{4\sqrt{6}\pi^{3/2}\sigma^4_{W_0}}{Nd^2\sigma^4_{W_1}} \sum_{\mathcal{P}(abcd)} e^{-\frac{\sigma^2_{W_0}\nabla^2_{r_{abcd}}}{6d}} + \frac{8\pi\,\sigma^4_{W_0}}{Nd^2\sigma^4_{W_1}} \sum_{\mathcal{P}(ab,cd)} e^{-\frac{\sigma^2_{W_0}(\nabla^2_{r_{ab}}+\nabla^2_{r_{cd}})}{2d}}. \tag{3.49}$$

We introduce the abbreviation $r_{abcd} := x_a + x_b - x_c - x_d$, and $\mathcal{P}(abcd) = 12$ refers to the number of ways ordered list of indices $a,c,b,d \in \{1,2,3,4\}$ can be chosen. Similarly, $r_{ab} := x_a - x_b$, and $\mathcal{P}(ab,cd) = 12$ is the number of ways ordered pairs $(a,c),(b,d) \in \{1,2,3,4\}$ can be drawn.

With this, Cos-net field theory action at $O(1/N)$ is

$$S_{\text{Cos}}[\phi] = \frac{2\sigma^2_{W_0}}{\sigma^2_{W_1} d}\int \mathrm{d}^d x\,\phi(x)\, e^{-\frac{\sigma^2_{W_0}\nabla^2_x}{2d}}\phi(x) - \int \mathrm{d}^d x_1\cdots \mathrm{d}^d x_4 \left[\frac{4\sqrt{6}\pi^{3/2}\sigma^4_{W_0}}{Nd^2\sigma^4_{W_1}} \sum_{\mathcal{P}(abcd)} e^{-\frac{\sigma^2_{W_0}\nabla^2_{r_{abcd}}}{6d}} - \frac{8\pi\,\sigma^4_{W_0}}{Nd^2\sigma^4_{W_1}} \sum_{\mathcal{P}(ab,cd)} e^{-\frac{\sigma^2_{W_0}(\nabla^2_{r_{ab}}+\nabla^2_{r_{cd}})}{2d}}\right]\phi(x_1)\cdots\phi(x_4) + O(1/N^2). \tag{3.50}$$

The NNGP action is local, but the leading order quartic interaction is non-local.

3.4.2. Single layer Gauss-net

As our next example, consider the output of a single-layer Gauss-net

$$\phi(x) = \frac{W^1_i \exp(W^0_{ij}x_j + b^0_i)}{\sqrt{\exp[2(\sigma^2_{b_0} + \frac{\sigma^2_{W_0}}{d}x^2)]}}, \tag{3.51}$$

for parameters drawn i.i.d. from $W^0 \sim \mathcal{N}(0, \frac{\sigma^2_{W_0}}{d})$, $W^1 \sim \mathcal{N}(0, \frac{\sigma^2_{W_1}}{N})$, and $b^0 \sim \mathcal{N}(0, \sigma^2_{b_0})$. The propagator is identical to Cos-net FT, and so is $G^{(2)}_{c,\text{Gauss}}(x_1,x_2)^{-1}$. We evaluate Gauss-net quartic coupling g_4, using (D.3), and (B.12) for $\tilde{G}^{(4)}_{c,\text{Gauss}}$, as

$$\begin{aligned} g_4^{\text{Gauss}}(x_1,\cdots,x_4) = &-\frac{4\sqrt{2}\,\pi^{3/2}\sigma^4_{W_0}}{\sqrt{3}N^2d^4\sigma^4_{W_1}} \sum_{\mathcal{P}(abcd)} \left[d^2N + 2\sigma^4_{W_0} - \frac{\sigma^5_{W_0}(d-\sigma^2_{W_0}\nabla^2_{r_{abcd}})}{d^{3/2}}\right] e^{-\frac{\sigma^2_{W_0}\nabla^2_{r_{abcd}}}{6d}} \\ &+ \frac{8\pi\,\sigma^4_{W_0}}{N^2d^4\sigma^4_{W_1}} \sum_{\mathcal{P}(ab,cd)} \left[d^2N + 6\sigma^4_{W_0} - 4d^3\sigma^5_{W_0} + \frac{\sigma^6_{W_0}}{d} + \left(\frac{2\sigma^7_{W_0}}{d^{3/2}} - \frac{\sigma^8_{W_0}}{d^2}\right)(\nabla^2_{r_{ab}} + \nabla^2_{r_{cd}}) \right. \\ &\left. + \frac{\sigma^{10}_{W_0}}{d^3}\nabla^2_{r_{ab}}\nabla^2_{r_{cd}}\right] e^{-\frac{\sigma^2_{W_0}(\nabla^2_{r_{ab}}+\nabla^2_{r_{cd}})}{2d}}, \end{aligned} \tag{3.52}$$

where $\mathcal{P}(ab,cd)$ and $\mathcal{P}(abcd)$ are defined as before.

Thus, Gauss-net FT action at $O(1/N)$,

$$S_{\text{Gauss}}[\phi] = \frac{2\sigma^2_{W_0}}{\sigma^2_{W_1} d}\int \mathrm{d}^d x\,\phi(x)\, e^{-\frac{\sigma^2_{W_0}\nabla^2_x}{2d}}\phi(x) + \int \mathrm{d}^d x_1\cdots \mathrm{d}^d x_4\, g_4^{\text{Gauss}}\,\phi(x_1)\cdots\phi(x_4), \tag{3.53}$$

differs from Cos-net FT at the level of quartic interaction.

4. Engineering actions: generalities, locality, and ϕ^4 theory

In section 3 we used the Edgeworth expansion and a 'duality' between fields and sources to compute couplings (including non-local ones) in the action as connected Feynman diagrams whose vertices are given by the usual connected correlators $G_c^{(n)}(x_1,\ldots,x_n)$. This general FT result is applicable in NN-FT of fixed architectures, but it does not answer the question of how to engineer an architecture that realizes a given action.

In this section we study how to design actions of a given type by deforming a Gaussian theory by an arbitrary operator. The result is simple and exploits the duality between the parameter-space and function-space descriptions of a FT. The main results are:

- **Action Deformations.** We develop a mechanism for expressing an arbitrary deformation of a Gaussian action as a deformation of the parameter density of a NN-FT.
- **Local Lagrangians.** We utilize the mechanism to engineer local interactions.
- **ϕ^4 Theory as a NN-FT.** Using a previous result that achieves free scalar FT as a NN-FT, we engineer local ϕ^4 theory as an NN-FT.
- **Cluster Decomposition.** We develop an approach to cluster decomposition, another notion of locality that is weaker than local interactions.

We also discuss why it might have been expected that ϕ^4 theory (and other well-studied FT) arises naturally at infinite-N.

To begin our analysis, consider the partition function of a Gaussian theory

$$Z_G[J] = \mathbb{E}_G[e^{\int \mathrm{d}^d x J(x)\phi(x)}], \tag{4.1}$$

where we have labelled both the partition function and the expectation with a G subscript to emphasize Gaussianity.

Now we wish to define a deformed theory that differs from the original only by an operator insertion, treating it in both function space and parameter space. The deformed partition function is given by

$$Z[J] = \mathbb{E}_G[e^{-\lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r \mathcal{O}_\phi(x_1,\ldots,x_r)} e^{\int \mathrm{d}^d x J(x)\phi(x)}], \tag{4.2}$$

where $\mathcal{O}_\phi$ is a non-local operator (though it may be chosen to be local) that has a subscript ϕ, denoting that it may depend on ϕ and its derivatives. In the function space, the partition function of the Gaussian theory is

$$Z_G[J] = \int D\phi\, e^{-S_G[\phi] + \int \mathrm{d}^d x J(x)\phi(x)}, \tag{4.3}$$

and the operator insertion corresponds to a deformation of the partition function to

$$Z[J] = \int D\phi\, e^{-S[\phi] + \int \mathrm{d}^d x J(x)\phi(x)} \tag{4.4}$$

where the action has been deformed

$$S_G[\phi] \to S[\phi] = S_G[\phi] + \lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r \mathcal{O}_\phi(x_1,\ldots,x_r). \tag{4.5}$$

We may treat this theory in perturbation theory in the usual way: correlators in the non-Gaussian theory are expanded perturbatively in λ and evaluated using the Gaussian expectation $\mathbb{E}_G$, which utilizes the Gaussian action when expressed in function-space.

How is this deformation expressed in parameter space, i.e. how do we think of this deformation from a NN perspective? In parameter space, the Gaussian partition function is

$$Z_G[J] = \int \mathrm{d}\theta P_G(\theta)\, e^{\int \mathrm{d}^d x J(x)\phi_\theta(x)}, \tag{4.6}$$

We remind the reader that in such a case Gaussianity is not obvious, but requires a judicious choice of parameter density $P(\theta)$ and architecture $\phi_\theta(x)$ such that we have a NN GP via the CLT. In parameter space, the deformation yields

$$Z[J] = \int \mathrm{d}\theta P_G(\theta)\, e^{-\lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r \mathcal{O}_{\phi_\theta}(x_1,\ldots,x_r)} e^{\int \mathrm{d}^d x J(x)\phi_\theta(x)}, \tag{4.7}$$

where we assume that where the operator $\mathcal{O}_{\phi_\theta}$ does not involve an explicit $\phi(x)$, but instead its parameter space representation; we will exemplify this momentarily. Again, correlators may be computed in perturbation theory in λ by expanding and evaluating in the Gaussian expectation, this time in the parameter space formulation.

We emphasize that if the function space and parameter space descriptions (4.3) and (4.6) represent the same partition function, then the deformed theories (4.4) and (4.7) are the same theory. That is, we see how

an arbitrary deformation of the action induces an associated deformation of the parameter space description. We will use this in section 4.2 to engineer ϕ^4 theory as a NN FT, and in 4.1 we will more explicitly deform a NN GP.

We end our general discussion with some theoretical considerations in NN FT, interpreting a non-Gaussian deformation $\mathcal{O}_{\phi_\theta}$ in terms of the framework of section 2, and also taking into account the universal approximation theorem.

A non-Gaussian deformation $\mathcal{O}_{\phi_\theta}$ must violate an assumption of the CLT. The architecture itself is still the same $\phi_\theta(x)$ as in the Gaussian theory. Instead, in (4.7) we may interpret the operator insertion as

$$P(\theta) := P_G(\theta)\, e^{-\lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r \mathcal{O}_{\phi_\theta}(x_1,\ldots,x_r)}, \tag{4.8}$$

i.e. same architecture, but with a deformed parameter distribution. This makes it clear that our non-Gaussian theory is still at infinite-N and therefore cannot receive non-Gaussianities in $1/N$-corrections. Instead, it receives non-Gaussianities because the deformed parameter distribution has independence breaking via the non-trivial relationship amongst the parameters in the deformation. There may also exist schemes for controlling non-Gaussian deformations in $1/N$, instead of via independence breaking, but it is beyond our scope.

Was it inevitable that systematic control over non-Gaussianities arises most naturally via independence breaking rather than $1/N$-corrections? The general answer is not clear, but we may use the control over non-Gaussianities to yield common theories, such as ϕ^4 theory in the next section. In that context we may ask a related question: was it inevitable that we obtain common interacting theories via independence breaking rather than $1/N$ corrections? This question has a better answer. Finite action configurations of a common theory, say ϕ^4 theory

$$S[\phi] = \int \mathrm{d}^d x \left[\phi(x)(\nabla^2 + m^2)\phi(x) + \frac{\lambda}{4!}\phi(x)^4\right], \tag{4.9}$$

are not *arbitrary* functions, since there may be some functions $\phi(x)$ that have infinite action. However, finite action configurations are still fairly general functions, and since they have finite action they occur with non-zero probability in the ensemble.

On the other hand, there are universal approximation theorems for NN, where the error in the approximation to a target function may decrease with increasing N. In such a case this theorem that is usually cited as a feature in ML may actually be a bug: at finite-N there exist functions that cannot be explicitly realized by a fixed architecture, but only approximated. We therefore find it reasonable to expect that there is at least one finite-action configuration $\phi(x)$ in ϕ^4 theory that cannot be realized by a finite-N NN of fixed architecture; in such a case, a NN-FT realization of ϕ^4 theory must be at infinite-N. This comment only scratches the surface, but we find the interplay between universal approximation theorems and realizable FT at finite-N to be worthy of further study.

4.1. Non-Gaussian deformation of a NN GP

To make the general picture more concrete, we would like to consider non-Gaussian deformations of any NN GP. The main result is that we may deform any NNGP by any operator we like, which breaks independence by deforming the parameter density, explaining the origin of non-Gaussianities by violating the independence.

As before, we consider a field built out of neurons,

$$\phi_\theta(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i\, h_i(x) \tag{4.10}$$

where the full set of parameters θ is realized by the set of parameters a_i and the set of parameters θ_h of the post-activations or neurons h. This equation forms the field out of a linear output layer with weights a_i acting on the post-activations, which could themselves be considered as the N-dimensional output of literally any NN. If the reader wishes, one may take ϕ to be a single-layer network by further choosing

$$h_i(x) = \sigma(b_{ij}x_j + c_i) \tag{4.11}$$

with $\sigma : \mathbb{R} \to \mathbb{R}$ a non-linear activation function such as ReLU or tanh; with this additional choice we now have θ_h comprised of b-parameters and c-parameters. Taking the parameter densities $P_G(a)$ and $P_G(\theta_h)$ to be independent and $N \to \infty$, $\phi(x) = \phi_\theta(x)$ is drawn from a GP; we have again used a subscript G to emphasize that these are the parameter densities of the Gaussian theory.

Deforming the Gaussian theory by an operator insertion, which in general is non-Gaussian, we have

$$Z[J] = \int \mathrm{d}a\, \mathrm{d}\theta_h\, P_G(a) P_G(\theta_h)\, e^{-\lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r\, \mathcal{O}_{\phi_{a,\theta_h}}(x_1,\ldots,x_r)} e^{\int \mathrm{d}^d x J(x)\phi_\theta(x)}. \tag{4.12}$$

We may interpret the operator insertion as deforming the independent Gaussian parameter density $P_G(a)P_G(\theta_h)$ to a non-trivial joint density

$$P(a,\theta_h) = P_G(a)P_G(\theta_h)\, e^{-\lambda \int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_r\, \mathcal{O}_{\phi_{a,\theta_h}}(x_1,\ldots,x_r)}. \tag{4.13}$$

The partition function is then

$$Z[J] = \int \mathrm{d}a\, \mathrm{d}\theta_h\, P(a,\theta_h)\, e^{\int \mathrm{d}^d x J(x)\phi_\theta(x)}, \tag{4.14}$$

an infinite-N non-Gaussian NN-FT where the operator insertion deforms the parameter density. At initialization, if one draws the parameters θ_h first, one may think of this as affecting the density from which the a-parameters are drawn; the draws of a-parameters are no longer independent.

For the sake of concreteness, consider the case of the single-layer network and take a general non-local quartic deformation. Then the operator insertion is

$$e^{-\int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_4\, g_4(x_1,\ldots x_4)\, \phi_{a,b,c}(x_1)\ldots\phi_{a,b,c}(x_4)}, \tag{4.15}$$

where Einstein summation is implied and we have absorbed the overall λ into the definition of the non-local coupling $g_4(x_1,\ldots,x_4)$. Inserting the equation for the NN

$$\phi_{a,b,c}(x) = \frac{1}{\sqrt{N}} a_i\, \sigma(b_{ij}x_j + c_i), \tag{4.16}$$

into the deformation, we obtain

$$e^{-\int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_4\, g_4(x_1,\ldots x_4)\, a_{i_1}\ldots a_{i_4}\, \sigma(b_{i_1 j_1}x_{j_1}+c_{i_1})\ldots\sigma(b_{i_4 j_4}x_{j_4}+c_{i_4})/N^2}, \tag{4.17}$$

which defines a deformed parameter density

$$P(a,b,c) = P_G(a)P_G(b)P_G(c)\, e^{-\int \mathrm{d}^d x_1 \ldots \mathrm{d}^d x_4\, g_4(x_1,\ldots x_4)\, a_{i_1}\ldots a_{i_4}\, \sigma(b_{i_1 j_1}x_{j_1}+c_{i_1})\ldots\sigma(b_{i_4 j_4}x_{j_4}+c_{i_4})/N^2}. \tag{4.18}$$

Then

$$Z[J] = \int \mathrm{d}a\, \mathrm{d}b\, \mathrm{d}c\; P(a,b,c)\, e^{\int \mathrm{d}^d x J(x)\, a_i\, \sigma(b_{ij}x_j+c_i)/\sqrt{N}} \tag{4.19}$$

is the partition function of a infinite-N NN-FT, as we impose $\lim N \to \infty$, with quartic non-Gaussianity induced by the breaking of independence in the joint parameter density $P(a,b,c)$.

4.2. ϕ^4 theory as a NN FT

To end this section and demonstrate the power of this technique, we would like to engineer the first interacting theory that any student learns: local ϕ^4 theory. The action is

$$S[\phi] = \int \mathrm{d}^d x \left[\phi(x)(\nabla^2 + m^2)\phi(x) + \frac{\lambda}{4!}\phi(x)^4\right]. \tag{4.20}$$

Following our prescription, we

- **Engineer the NNGP**. Using the result of [10], we take

$$\phi_{a,b,c}(x) = \sum_i \frac{a_i\, \cos(b_{ij}x_j + c_i)}{\sqrt{\mathbf{b}_i^2 + m^2}}, \tag{4.21}$$

  where the sum runs from 1 to $N = \infty$, $\mathbf{b}_i$ is the vector that is the ith row of the matrix b_{ij}, and the parameter densities of the Gaussian theory are

$$P_G(a) = \prod_i e^{-\frac{N}{2\sigma_a^2} a_i a_i} \tag{4.22}$$

$$P_G(b) = \prod_i P_G(\mathbf{b}_i) \text{ with } P_G(\mathbf{b}_i) = \text{Unif}(B_\Lambda^d) \tag{4.23}$$

$$P_G(c) = \prod_i P_G(c_i) \text{ with } P_G(c_i) = \text{Unif}([-\pi, \pi]), \tag{4.24}$$

where B_Λ^d is a d-sphere of radius Λ. The density $P_G(\mathbf{b}_i)$ is not independent in the vector index j, but all that is needed for Gaussianity is independence in the i index, which is clear due to the product nature of $P_G(b)$. The power spectrum (Fourier-transform of the two-point function) is

$$G^{(2)}(p) = \frac{\sigma_a^2 (2\pi)^d}{2\,\text{vol}(B_\Lambda^d)} \frac{1}{p^2 + m^2}, \tag{4.25}$$

which becomes the standard free scalar result $1/(p^2 + m^2)$ by a trivial rescaling

$$\phi_{a,b,c}(x) = \sqrt{\frac{2\,\text{vol}(B_\Lambda^d)}{\sigma_a^2 (2\pi)^d}} \sum_{i,j} \frac{a_i \cos(b_{ij} x_j + c_i)}{\sqrt{\mathbf{b}_i^2 + m^2}}. \tag{4.26}$$

This NN GP is equivalent to the free scalar theory of mass m in d Euclidean dimensions, with

$$G^{(2)}(p) = \frac{1}{p^2 + m^2}, \tag{4.27}$$

where Λ plays the role of a hard UV cutoff on the momentum.

- **Introduce the Operator Insertion**. Given the NNGP above, or any other NNGP realizing the free scalar FT, we wish to insert the operator

$$e^{-\frac{\lambda}{4!}\int \mathrm{d}^d x\, \phi_{a,b,c}(x)^4}, \tag{4.28}$$

associated to a local ϕ^4 interaction.

- **Absorb the Operator into a Parameter Density Deformation**. The non-Gaussian operator insertion deforms the parameter density to

$$P(a,b,c) = P_G(a) P_G(b) P_G(c)\, e^{-\frac{\lambda}{4!}\int \mathrm{d}^d x\, \phi_{a,b,c}(x)^4}, \tag{4.29}$$

where for $\phi_{a,b,c}(x)$ it is to be understood that the RHS of (4.26) is inserted, yielding an expression that is only a function of a's, b's, and c's.

- **Write the Partition Function**. We then have a partition function for the deformed theory, given by

$$Z[J] = \int \mathrm{d}a\, \mathrm{d}b\, \mathrm{d}c\; P(a,b,c)\, e^{\int \mathrm{d}^d x J(x)\, \phi_{a,b,c}(x)}, \tag{4.30}$$

where again it is to be understood that we insert the RHS of (4.26) for $\phi_{a,b,c}$ and (4.29) for $P(a,b,c)$; there are no explicit fields in the expression, it depends only on the architecture (which includes parameters a,b,c) and the joint parameter density.

Thus, the architecture (4.26) and parameter density (4.29) realize local ϕ^4 theory via the partition function (4.30). We discuss the connections between GPs, locality, and translation invariance in appendix E.

Let us briefly address RG flows. The definition of a fixed non-Gaussian theory here involves the choice of a fixed value of λ, in addition to the choice of a fixed value of Λ that was implicit in the fixing of the GP. From that starting point, decreasing Λ while keeping the correlators fixed induces an RG flow for λ governed by the usual Callan–Symanzik equation. In the language of the NN architecture, this is interpreted as a flow in the parameter density that is necessary to fix the correlators as Λ is decreased.

4.3. Cluster decomposition and space independence breaking

We now turn to a weaker notion of locality: cluster decomposition. Given a field $\phi(x)$ (or NN in our context) we say that it satisfies cluster decomposition if all connected correlation functions $G_c^{(r)}(x_1, \ldots, x_r)$ asymptote to zero in the limit where the separation between any two space points $x_i, x_j, i \neq j$ is taken to ∞,

$$\lim_{|x_i - x_j| \to \infty} G_c^{(r)}(x_1, \ldots, x_r) = 0. \tag{4.31}$$

If the probability density function of ϕ has the form

$$P(\phi) = \frac{1}{Z}\exp\left(-\int \mathrm{d}x\,\mathcal{L}\left(x, \phi(x), \frac{\partial\phi}{\partial x}, \ldots, \frac{\partial^n\phi}{\partial x^n}\right)\right) \tag{4.32}$$

where Z is a normalization constant and n is finite, we say that $\phi(x)$ has a local Lagrangian density. This is a stronger notion of locality compared to cluster decomposition, as any theory with a local Lagrangian density satisfies cluster decomposition, but the converse is not true [44].

Checking whether a theory satisfies cluster decomposition requires knowledge of the asymptotic behavior of correlation functions, but not the probability density function. As calculating the probability density function of an NN-FT is more challenging than computing the correlation functions, checking cluster decomposition is easier than determining whether there exists a local Lagrangian density that describes the system.

The main result we describe in this section is a framework that enables engineering NN architectures that satisfy cluster decomposition.

4.4. Space independent FT

We will perform our analysis by studying, and then moving away from, a case with a very strong assumption: FT that are defined by fields that have independent statistics at different space (or space) points x_I. We call these FT *space independent* (SI) FT. While one can still view such fields as random functions defined on a continuously differentiable space, in general the field configurations are discontinuous; avoiding this would require statistical correlations between nearest neighbors, violating the assumption. This 'd-dimensional' FT is really a collection of uncountably many independent 0-d theories. This means that the partition function factorizes

$$Z_{\phi_{\mathrm{SI}}}[J] = \mathbb{E}[e^{\int \mathrm{d}^d x_I J(x_I)\phi(x_I)}] = \prod_I \mathbb{E}_{\phi_{\mathrm{SI}}}[e^{J(x_I)\phi(x_I)}], \tag{4.33}$$

where the product runs over all space points x_I. This form is agnostic about the origin of the statistics and may be specified in either the function space or parameter space description. In parameter space, independent statistics at different space points x_I means that the SI theory has partition function

$$Z_{\phi_{\mathrm{SI}}}[J] = \prod_i \int \mathrm{d}\theta_I P_I(\theta_I)\, e^{J(x_I)\,\phi_{\theta_I}(x_I)}, \tag{4.34}$$

i.e. each space point x_I has its own ensemble of NN $\phi_{\theta_I}(x_I)$ with its own set of parameters θ_I that is independent of θ_J for $I \neq J$. In function space, independence means that

$$Z_{\phi_{\mathrm{SI}}}[J] = \prod_I \int D\phi_I\, e^{-S[\phi(X_I)]+J(x_I)\phi(x_I)}, \tag{4.35}$$

i.e. the action is such that the path integral factorizes. An immediate consequence of this factorization is that the action cannot contain derivatives of $\phi(x_I)$, as these would depend on the value of ϕ not only at point x_I, but a local neighborhood around it. Then, the action is of the form,

$$S(\phi(x_I)) = V[\phi(x_I)], \tag{4.36}$$

which, turning the product into a sum in the exponent, gives the more canonical form

$$Z\phi_{\mathrm{SI}}[J] = \int \left(\prod_I D\phi_I\right) e^{-\int \mathrm{d}^d x_I\,(V[\phi(X_I)]-J(x_I)\phi(x_I))}. \tag{4.37}$$

This is a FT with a potential, but no derivatives. The field values at different points of space are independent random variables. If they are identically distributed $V[\phi(x_I)]$ is fixed $\forall I$ and the different factors in $Z_{\mathrm{SI}}[J]$ enjoy an S_L permutation symmetry, where the number of space points L is infinite in the continuum limit.

Before introducing correlations between the field values at different space points, let us first study the statistics of the SI theory. Denote the cumulants of ϕ at a given point x_I as $\kappa_r^{\phi}(x_I)$. For simplicity, we will assume that the field values at different space points are identically distributed, i.e. $\kappa_r^{\phi}(x_I) = \kappa_r^{\phi}$ is fixed for all I, which will also be important for translation invariance. We also assume that they are mean free, $\kappa_1^{\phi} = 0$. Next, we consider the CGF, which takes the form

$$W_{\phi_{SI}}[J] = \log\Big(Z_{\phi_{SI}}[J]\Big) = \log\Big(\prod_I \mathbb{E}_{\phi_{\rm SI}}[e^{J(x_I)\phi(x_I)}]\Big),$$
$$= \int \mathrm{d}x \log\Big(\mathbb{E}_{\phi_{\rm SI}}[e^{J(x)\phi(x)}]\Big),$$
$$= \int \mathrm{d}x\, W[J;x], \tag{4.38}$$

where $W[J;x]$ is the CGF of ϕ at space point x. Just as the partition function $Z[J]$ factorizes into a product of partition functions associated to individual space points, the CGF $W[J] = \log Z[J]$ becomes a sum (or integral, in this case). The connected correlators are easily computed by taking derivatives[10] , where $\partial J(x_I)/\partial J(x_J) = \delta(x_I - x_J)$,

$$G_c^{(n)}(x_1,\ldots,x_n) = \left(\prod_{I=1}^n \frac{\partial}{\partial J(x_I)}\right) W_{\phi_{SI}}[J],$$
$$= \int \mathrm{d}x \left(\prod_{I=1}^n \frac{\partial}{\partial J(x_I)}\right) W[J;x]. \tag{4.39}$$

and the connected correlation functions of SI networks $\phi_{\rm SI}$ simplifies to

$$G_c^{(n)}(x_1,\ldots,x_n) = \int \mathrm{d}x \Big(\frac{\partial}{\partial J(x)}\Big)^n W[J;x] \prod_{I=1}^n \delta(x - x_I),$$
$$= \int \mathrm{d}x\, \kappa_n^\phi \prod_{I=1}^n \delta(x - x_I), \tag{4.40}$$

with n delta functions. The n-point connected correlator is nonzero only when $x_1 = x_2 = \cdots = x_n$, and its magnitude is determined by κ_n^ϕ.

The correlation functions can be written in terms of the connected correlators. For example, the two point function of ϕ is

$$G^{(2)}(x_1,x_2) = \mathbb{E}_\phi[\phi(x_1)\phi(x_2)],$$
$$= \kappa_2^\phi\, \delta(x_1 - x_2) + (\kappa_1^\phi)^2$$
$$= \kappa_2^\phi\, \delta(x_1 - x_2). \tag{4.41}$$

As $\phi(x_1)$ and $\phi(x_2)$ are independent and mean free, $G^{(2)}_{\phi_{\rm SI}}(x_1,x_2)$ is nonzero only when $x_1 = x_2$. Similarly, the four point function is

$$G^{(4)}(x_1,x_2,x_3,x_4) = \kappa_4^\phi\, \delta(x_1 - x_2)\delta(x_1 - x_3)\delta(x_1 - x_4) + (\kappa_2^\phi)^2 \left(\delta(x_1 - x_2)\delta(x_3 - x_4)\right.$$
$$\left. + \delta(x_1 - x_3)\delta(x_2 - x_4) + \delta(x_1 - x_4)\delta(x_2 - x_3)\right). \tag{4.42}$$

The statistics of the theory is completely determined by the space independence assumption and the cumulants κ_r^ϕ. The general n-point function can be expressed as

$$G^{(n)}(x_1,\ldots,x_n) = \sum_{\alpha \in S_n} \prod_{r\in\alpha} G^{(r)}_{\phi_{\rm SI},c}(x_1,\ldots,x_n), \tag{4.43}$$

where S_n denotes partitions of the set $\{1,\ldots,n\}$.

4.5. Space-time independence breaking

Clearly we do not want to stop with SI theories. We will now introduce correlations between different space points to 'stitch together' the L 0-dimensional theories (associated to the SI fields) into a d-dimensional FT. This requires modifying the theory in some way so that there are non-trivial correlations between field values at different points.

One way to do so is to define new field variables $\Phi(x_I)$ as a function of the SI fields $\phi(x_I)$,

$$\Phi(x_I) = \Phi\left(\phi(x_1),\ldots,\phi(x_L)\right). \tag{4.44}$$

[10] We remind the reader that space derivatives are ill-defined, as $\phi(x)$ is discontinuous everywhere. However, derivatives with respect to $J(x_I)$ are still well defined.

As the value of Φ at site x_I in principle depends on the values of ϕ at all space points, $\Phi(x_I)$ and $\Phi(x_J)$ are correlated in general, even when $I \neq J$. The statistics of $\Phi(x_I)$ are then determined by the functional form of (4.44), as well as the statistics of $\phi(x_I)$. However, such a general formulation (4.44) is unwieldy, and we therefore simplify the picture.

We will describe a family of architectures where $\Phi(x_I)$ is constructed by a simpler ansätz, a smearing of $\phi(a)_{a\in\{x_1,\cdots,x_L\}}$ across all space points, and write down a necessary and sufficient condition to satisfy cluster decomposition. Consider the architecture,

$$\Phi(x_I) = \int_{-\infty}^{\infty} \mathrm{d}a f(x_I - a)\,\phi(a) \tag{4.45}$$

for some continuous and differentiable function $f(x_I - a)$. First, note that although a generic draw of $\phi(a)$ is discontinuous due to independence across different points in space, $\Phi(x_I)$ is rendered continuous by the smearing. Furthermore, if the function f is nonzero everywhere, $\Phi(x_I)$ will have correlations between all pairs of lattice sites.

We wish to check whether cluster decomposition is satisfied, and therefore need to compute correlation functions of $\Phi(x)$. The Φ-correlators are given by

$$\begin{aligned} G_{\Phi}^{(n)}(x_1,\ldots,x_n) &= \mathbb{E}_{\phi}\left[\Phi(x_1)\cdots\Phi(x_n)\right], \\ &= \mathbb{E}_{\phi}\left[\prod_{i=1}^{n}\int \mathrm{d}a_i\, f(x_i - a_i)\,\phi(a_i)\right]. \end{aligned} \tag{4.46}$$

As f does not depend on ϕ, we can carry out the expectation value over ϕ to obtain

$$G_{\Phi}^{(n)}(x_1,\ldots,x_n) = \int \prod_{i=1}^{n} \mathrm{d}a_i\, f(x_i - a_i)\, G_{\phi}^{(n)}(a_1,\ldots,a_n), \tag{4.47}$$

where $G_{\phi}^{(n)}(a_1,\ldots,a_n)$ is the n-point correlation function of ϕ. The only contribution to the connected correlator of $\Phi(x)$ comes from the connected piece of $G_{\phi}^{(n)}(a_1,\ldots,a_n)$ with n delta functions[11],

$$G_{c}^{(n)}(x_1,\ldots,x_n) = \kappa_n^{\phi} \int \mathrm{d}x \prod_{i=1}^{n} \mathrm{d}a_i\, f(x_i - a_i)\,\delta(x - a_i). \tag{4.48}$$

Evaluating the integral, we obtain

$$G_{c}^{(n)}(x_1,\ldots,x_n) = \kappa_n^{\phi} \int \mathrm{d}x \prod_{i}^{n} f(x_i - x). \tag{4.49}$$

Cluster decomposition is satisfied if and only if (4.49) asymptotes to zero in the limit where the separation between any two of the space points x_I, x_J is taken to ∞. Any smearing function $f(x)$ that decays faster than $1/x$ asymptotically satisfies this condition[12].

4.5.1. Example: Gaussian smearing

We now present an example with a particular choice of the smearing function f and show that the resulting theory satisfies cluster decomposition. Let

$$f(x) = e^{-\frac{x^2}{\beta}}, \tag{4.50}$$

$$\Phi(x) = \int \mathrm{d}a\, e^{-\frac{(x-a)^2}{\beta}}\,\phi(a), \tag{4.51}$$

for some $\beta > 0$. As before, we will consider a case where $\phi(x)$ at different space points are identically distributed, with cumulants κ_{ϕ}^{n}.[13] Following equation (4.49), the cumulants of $\Phi(x)$ are then given by

[11] The remaining terms factorize and do not contribute to the connected correlator.

[12] Note that the SI theory automatically satisfies cluster decomposition as the connected correlator, cf (4.40), vanishes unless all space points coincide.

[13] As $\phi(x)$ are identically distributed, the cumulants do not depend on the space coordinates x.

$$G_c^{(n)}(x_1,\dots,x_n) = \kappa_n^{\phi}\int \mathrm{d}x \prod_{i=1}^{n} e^{-\frac{(x_i-x)^2}{\beta}},$$
$$= \kappa_n^{\phi}\sqrt{\frac{\pi\beta}{n}}\exp\left[M_{ij}x_i x_j/\beta\right], \tag{4.52}$$

where

$$M_{ij} = \begin{cases} \frac{2}{n}-2, & \text{if } i=j \\ \frac{2}{n}, & \text{otherwise} \end{cases} \tag{4.53}$$

This matrix is negative semidefinite, with eigenvalues $\lambda_1 = \cdots = \lambda_{n-1} = -\beta/2$, $\lambda_n = 0$, and the eigenvector corresponding to λ_n is $(1,\cdots,1)$. Consequently, the cumulant $G_c^{(n)}(x_1,\dots,x_n)$ vanishes when any of the x_i are taken to be large, unless they coincide $x_1 = \cdots = x_n$. This theory thus satisfies cluster decomposition.

The dependence of the connected correlators (4.52) on the space coordinates x_i is completely determined by the choice of smearing function f, while their magnitudes depend both on f as well as the cumulants κ_ϕ^n. Although our main motivation here has been to engineer NN architectures that satisfy cluster decomposition, smearing layers offer great flexibility in manipulating the connected correlators and might be useful in designing NN with other desired properties.

5. Conclusions

In this paper we continued the development of NN FT (NN-FT), a new approach to FT in which a theory is specified by a NN architecture and a parameter density. This description enables a parameter space description of the statistics, yielding a different method for computing correlation functions. For a more detailed introduction to NN-FT, see the introduction and references therein.

We focused on three foundational aspects of NN-FT: non-Gaussianity, actions, and locality. Via the CLT, many architectures admit an $N\to\infty$ limit in which the associated NN-FT is Gaussian, i.e. a generalized free FT. In the ML literature, these are called NNGPs. In section 2 we demonstrated that interactions arise from parametrically violating assumptions of the CLT, yielding non-Gaussianities arising from $1/N$-corrections, as well as the breaking of statistical independence and the identicalness assumption. These interactions are apparent via parameter-space calculations of connected correlation functions, but manifest themselves as non-Gaussianities in the field density $P[\phi] = \exp(-S[\phi])$. In section 3 we developed a technique that allows for the action to be computed from the connected correlation functions, via connected Feynman diagrams. This is an inversion of the usual approach in FT: we compute coupling functions in terms of connected correlators, rather than the other way around. The technique was applied to NN-FT, including an analysis involving the parametric non-Gaussianities we studied. In section 4 we studied how to design architectures that realize a given action. We do so by deforming an NNGP by an operator insertion that, from a function-space perspective, corresponds to a deformation of the GP action. However, since we know the architecture we may also express the deformation in parameter space, in which case the non-Gaussianity associated to a given deformation of the action has a natural interpretation as a deformation of the NN parameter density. That is, the interactions arise from independence breaking. We apply this technique to induce local interactions, and derive an architecture that realizes ϕ^4 theory as an infinite NN-FT.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

We thank Sergei Gukov, Mathis Gerdes, Jessica Howard, Ro Jefferson, Gowri Kurup, Joydeep Naskar, Fabian Ruehle, Jiahua Tian, Jacob Zavatone-Veth, and Kevin Zhang for discussions. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). This work was performed in part at Aspen Center for Physics, which is supported by National Science Foundation Grant PHY-2210452. A M thanks ECT* and the ExtreMe Matter Institute EMMI at GSI, Darmstadt, for support in the framework of an ECT* Workshop during which part of this work has been completed. J H is supported by NSF CAREER Grant PHY-1848089.

Appendix A. Continuum Hermite polynomials

Let us first recall the definition of continuum Hermite polynomials for convenience,

$$H(\phi, x_1, \cdots, x_n) = (-1)^n e^{S_G} \frac{\delta}{\delta\phi(x_1)} \cdots \frac{\delta}{\delta\phi(x_n)} e^{-S_G}. \tag{A.1}$$

Defining

$$S_i = \frac{\delta S_G}{\delta\phi(x_i)}, \qquad S_{i,j} = \frac{\delta^2 S_G}{\delta\phi(x_i)\delta\phi(x_j)}, \tag{A.2}$$

the first six Hermite polynomials are,

$$\begin{aligned}
&H_1(\phi, x_1) = S_1, \\
&H_2(\phi, x_1, x_2) = S_1 S_2 - S_{1,2}, \\
&H_3(\phi, x_1, x_2, x_3) = S_1 S_2 S_3 - S_{1,2} S_3[3], \\
&H_4(\phi, x_1, x_2, x_3, x_4) = S_1 S_2 S_3 S_4 - S_{1,2} S_3 S_4[6] + S_{1,2} S_{3,4}[3], \\
&H_5(\phi, x_1, x_2, x_3, x_4, x_5) = S_1 S_2 S_3 S_4 S_5 - S_{1,5} S_2 S_3 S_4[10] + S_{1,2} S_{3,4} S_5[15], \\
&H_6(\phi, x_1, x_2, x_3, x_4, x_5, x_6) = S_1 S_2 S_3 S_4 S_5 S_6 - S_{1,6} S_2 S_3 S_4 S_5[15] \\
&\qquad\qquad + S_{1,2} S_{3,4} S_5 S_6[45] - S_{1,2} S_{3,4} S_{5,6}[15],
\end{aligned} \tag{A.3}$$

where the square brackets denote sums over all terms with a given index structure, for example $S_{1,2}S_3[3] = S_{1,2}S_3 + S_{1,3}S_2 + S_{2,3}S_1$.

Appendix B. Details of examples

B.1. ReLU-net cumulants at finite N, i.i.d. parameters

Let us study the output distribution of a single hidden layer network at width N, ReLU activation function, $d = d_{\text{out}} = 1$, given by

$$\phi(x) = W_i^1 R(W_{ij}^0 x_j) \;\; \text{where } R(z) = \begin{cases} z, \text{ for } z \geqslant 0 \\ 0, \text{ otherwise} \end{cases}. \tag{B.1}$$

The parameters are sampled i.i.d. $W^0 \sim \mathcal{N}(0, \frac{\sigma_{W_0}^2}{d})$, $W^1 \sim \mathcal{N}(0, \frac{\sigma_{W_1}^2}{N})$, and bias $= 0$. The 2-pt function is $G^{(2)}_{c,\text{ReLU}}(x,y) = \sigma_{W_0}^2 \sigma_{W_1}^2 \big(R(x)R(y) + R(-x)R(-y)\big)/2$, and higher order cumulants are

$$\begin{aligned}
G^{(4)}_{c,\text{ReLU}}(x_1, \cdots, x_4) = \frac{1}{N} \Bigg(&\frac{15\sigma_{W_0}^4 \sigma_{W_1}^4}{4d^2} \left(\sum_{j=\pm 1} R(jx_1)R(jx_2)R(jx_3)R(jx_4) \right) \\
&- \frac{\sigma_{W_0}^4 \sigma_{W_1}^4}{4d^2} \left(\sum_{\mathcal{P}(abcd)} \sum_{j=\pm 1} R(jx_a)R(jx_b)R(-jx_c)R(-jx_d) \right) \Bigg),
\end{aligned} \tag{B.2}$$

$$\begin{aligned}
G^{(6)}_{c,\text{ReLU}}(x_1, \cdots, x_6) = \frac{1}{N^2} \Bigg[&\frac{225\sigma_{W_0}^6 \sigma_{W_1}^6}{2d^3} \left(\sum_{j=\pm 1} R(jx_1)R(jx_2)R(jx_3)R(jx_4)R(jx_5)R(jx_6) \right) \\
&- \sum_{\mathcal{P}(abcdef)} \left(\frac{9\sigma_{W_0}^6 \sigma_{W_1}^6}{4d^3} \left(\sum_{j_1=\pm 1} R(j_1x_a)R(j_1x_b)R(j_1x_c)R(j_1x_d) \right) \left(\sum_{j_2=\pm 1} R(j_2x_e)R(j_2x_f) \right) \right. \\
&\left. - \frac{\sigma_{W_0}^6 \sigma_{W_1}^6}{4d^3} \left(\sum_{j_1=\pm 1} R(j_1x_a)R(j_1x_b) \right) \left(\sum_{j_2=\pm 1} R(j_2x_c)R(j_2x_d) \right) \left(\sum_{j_3=\pm 1} R(j_3x_e)R(j_3x_f) \right) \right) \Bigg],
\end{aligned} \tag{B.3}$$

where $\mathcal{P}(abcd)$ denotes all combinations of non-identical a, b, c, d drawn from $\{1, 2, 3, 4\}$, and similarly for $\mathcal{P}(abcdef)$.

B.2. Cos-net cumulants at finite N, non-i.i.d. parameters

The output of a single hidden layer, finite N, fully connected feedforward network with cosine activation function is given by,

$$\phi(x) = W_i^1 \cos(W_{ij}^0 x_j + b_i^0). \tag{B.4}$$

For i.i.d. parameters, e.g. $W^1 \sim \mathcal{N}(0, \sigma_{W_1}^2/N)$, $W^0 \sim \mathcal{N}(0, \sigma_{W_0}^2/d)$, and $b^0 \sim \text{Unif}[-\pi, \pi]$, the 2-pt function is given by $G_{c,\text{Cos}}^{(2)}(x,y) = \frac{\sigma_{W_1}^2}{2} e^{-\frac{1}{2d}\sigma_{W_0}^2 (x-y)^2}$. For simplicity, we focus on the $d=1$ case; the statistical independence of first linear layer weights can be broken by a hyperparameter $\alpha_{\text{IB}} \ll 1$, then the correlated weight distribution is

$$\mathcal{P}(W^0) = c \exp\left[-\sum_i \left(\frac{(W_i^0)^2}{2\sigma_{W_0}^2} + \frac{\alpha_{\text{IB}}}{N^2} \sum_{i_1, i_2} (W_{i_1}^0)^2 (W_{i_2}^0)^2 \right) \right], \tag{B.5}$$

where c is a normalization constant. The cumulative non-Gaussian effects due to finite width and non-i.i.d. parameters alter all correlation functions, including the 2-pt function at finite width. Using perturbation theory at leading order in α_{IB}, the 2nd and 4th cumulants are evaluated as the following,

$$G_{c,\text{Cos}}^{(2)}(x_1, x_2) = \frac{\alpha_{\text{IB}} \sigma_{W_0}^4 \sigma_{W_1}^2 e^{-\frac{\sigma_{W_0}^2 (\Delta x_{12})^2}{2}}}{2N} \left[\left(1 + \sigma_{W_0}^2 (\Delta x_{12})^2\right) - \frac{\left(1 - 5\sigma_{W_0}^2 (\Delta x_{12})^2 + \sigma_{W_0}^4 (\Delta x_{12})^4\right)}{N} \right], \tag{B.6}$$

$$\begin{aligned} G_{c,\text{Cos}}^{(4)}(x_1, x_2, x_3, x_4) = \frac{\sigma_{W_1}^4}{8N} \sum_{\mathcal{P}(abcd)} \Bigg[& \left(-2e^{-\frac{\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2\right)}{2}} + 3e^{-\frac{\sigma_{W_0}^2 (\Delta x_{ab} + \Delta x_{cd})^2}{2}} \right) \\ & + \frac{\alpha_{\text{IB}} \sigma_{W_0}^4}{N} \Bigg(-6e^{-\frac{\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2\right)}{2}} + 3e^{-\frac{\sigma_{W_0}^2 (\Delta x_{ab} + \Delta x_{cd})^2}{2}} + 3\sigma_{W_0}^2 (\Delta x_{ab} + \Delta x_{cd})^2 \\ & \times e^{-\frac{\sigma_{W_0}^2 (\Delta x_{ab} + \Delta x_{cd})^2}{2}} - 2\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2\right) e^{-\frac{\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2\right)}{2}} \\ & - 2\sigma_{W_0}^4 (\Delta x_{ab})^2 (\Delta x_{cd})^2 e^{-\frac{\sigma_{W_0}^2 \left((\Delta x_{ab})^2 + (\Delta x_{cd})^2\right)}{2}} \Bigg) \Bigg], \end{aligned} \tag{B.7}$$

where $\Delta x_{ij} := x_i - x_j$. The Fourier transformation of this cumulant at $\alpha_{\text{IB}} = 0$ is

$$\begin{aligned} \tilde{G}_{c,\text{Cos}}^{(4)} = \frac{3\pi^{3/2} \sigma_{W_1}^4 \sqrt{d}}{2\sqrt{2} N \sigma_{W_0}} \Bigg[& e^{-\frac{p_1^2 d}{2\sigma_{W_0}^2}} \left(\delta^d(p_1 + p_2) \delta^d(p_1 + p_3) \delta^d(p_4 - p_1) + \delta^d(p_2 - p_1) \delta^d(p_1 + p_3) \right. \\ & \left. \times \delta^d(p_1 + p_4) + \delta^d(p_1 + p_2) \delta^d(p_3 - p_1) \delta^d(p_1 + p_4) \right) \Bigg] - \frac{\pi \sigma_{W_1}^4 d}{2N\sigma_{W_0}^2} \Bigg[e^{-\frac{(p_1^2 + p_2^2) d}{2\sigma_{W_0}^2}} \delta^d(p_1 + p_4) \delta^d(p_2 + p_3) \\ & \times e^{-\frac{(p_1^2 + p_2^2) d}{2\sigma_{W_0}^2}} \delta^d(p_1 + p_3) \delta^d(p_2 + p_4) + e^{-\frac{(p_1^2 + p_3^2) d}{2\sigma_{W_0}^2}} \delta^d(p_1 + p_2) \delta^d(p_3 + p_4) \Bigg] + p_1 \leftrightarrow p_2, p_3, p_4, \end{aligned} \tag{B.8}$$

where use the convention $e^{i(p_1 x_1 + p_2 x_2 + p_3 x_3 + p_4 x_4)}$.

Next, we present another example where non-Gaussianities arise due to both finite width and non-i.i.d. parameters.

B.3. Gauss-net at finite N, non-i.i.d. parameters

We define the Gauss-net architecture as a single hidden layer, width N, feedforward network with exponential activation function, and an overall normalizing factor, such that the output is

$$\phi(x) = \frac{W_i^1 \exp(W_{ij}^0 x_j + b_i^0)}{\sqrt{\exp[2(\sigma_{b_0}^2 + \frac{\sigma_{W_0}^2}{d} x^2)]}}. \tag{B.9}$$

For i.i.d. parameter distributions, $W^0 \sim \mathcal{N}(0, \frac{\sigma_{W_0}^2}{d})$, $W^1 \sim \mathcal{N}(0, \frac{\sigma_{W_1}^2}{N})$, and $b^0 \sim \mathcal{N}(0, \sigma_{b_0}^2)$, the 2-pt function is $G_{c,\text{Gauss}}^{(2)}(x,y) = \frac{\sigma_{W_1}^2}{2} e^{-\frac{1}{2d}\sigma_{W_0}^2 (x-y)^2}$, identical as Cos-net. We break the statistical independence of the first

linear layer weights similar to the previous example, at $d=1$. Then, the 2nd and 4th order cumulants at leading order in $\alpha_{\rm IB}$ are,

$$G^{(2)}_{c,\rm Gauss}(x_1,x_2) = -\frac{\alpha_{\rm IB}\sigma^4_{W_0}\sigma^2_{W_1}}{2N}e^{-\frac{\sigma^2_{W_0}(\Delta x_{12})^2}{2}}\left[\left(\sigma^2_{W_0}X^2_{12}-1\right)-\frac{\left(1+5\sigma^2_{W_0}X^2_{12}+\sigma^4_{W_0}X^4_{12}\right)}{N}\right], \tag{B.10}$$

and,

$$\begin{aligned}G^{(4)}_{c,\rm Gauss}(x_1,x_2,x_3,x_4) = &\frac{3\sigma^4_{W_1}\exp\left(-\frac{\sigma^2_{W_0}\left(x_1^2-2x_1(x_2+X_{34})+x_2^2-2x_2X_{34}+(\Delta x_{34})^2\right)}{2}\right)}{4N}\\ &+\frac{\alpha_{\rm IB}\sigma^4_{W_0}\sigma^4_{W_1}}{4N^2}\left(3\exp\left(-\frac{\sigma^2_{W_0}\left(x_1^2-2x_1(x_2+X_{34})+x_2^2-2x_2X_{34}+(\Delta x_{34})^2\right)}{2}\right)\right.\\ &-3\sigma^2_{W_0}(X_{12}+X_{34})^2\exp\left(-\frac{\sigma^2_{W_0}\left(x_1^2-2x_1(x_2+X_{34})+x_2^2-2x_2X_{34}+(\Delta x_{34})^2\right)}{2}\right)\\ &-\sum_{\mathcal{P}(abcd)}\left(3e^{-\frac{\sigma^2_{W_0}\left((\Delta x_{ab})^2+(\Delta x_{cd})^2\right)}{2}}-\sigma^2_{W_0}\left(X^2_{ab}+X^2_{cd}\right)e^{-\frac{\sigma^2_{W_0}\left((\Delta x_{ab})^2+(\Delta x_{cd})^2\right)}{2}}\right.\\ &\left.\left.+\sigma^4_{W_0}X^2_{ab}X^2_{cd}e^{-\frac{\sigma^2_{W_0}\left((\Delta x_{ab})^2+(\Delta x_{cd})^2\right)}{2}}\right)\right)-\sum_{\mathcal{P}(abcd)}\frac{\sigma^4_{W_1}}{4N}e^{-\frac{\sigma^2_{W_0}\left((\Delta x_{ab})^2+(\Delta x_{cd})^2\right)}{2}},\end{aligned} \tag{B.11}$$

where $X_{ij}:=x_i+x_j$, and $\Delta x_{ij}:=x_i-x_j$.

At $\alpha_{\rm IB}=0$, the Fourier transformation of this cumulant becomes the following

$$\begin{aligned}\tilde{G}^{(4)}_{c,\rm Gauss} = &\frac{\pi^{3/2}\sigma^4_{W_1}}{2\sqrt{2}N^2d^{3/2}\sigma_{W_0}}\left[e^{-\frac{p_1^2d}{2\sigma^2_{W_0}}}\left(d^2N-dp_1^2\sigma^2_{W_0}+2\sigma^4_{W_0}\right)\left(\delta^d(p_1+p_2)\delta^d(p_1+p_3)\right.\right.\\ &\left.\left.\times\,\delta^d(p_4-p_1)+\delta^d(p_2-p_1)\delta^d(p_1+p_3)\delta^d(p_1+p_4)+\delta^d(p_1+p_2)\delta^d(p_3-p_1)\delta^d(p_1+p_4)\right)\right]\\ &-\frac{\pi\sigma^4_{W_1}}{2N^2\sigma^2_{W_0}d}\left[\left(d^2(N+p_1^2p_2^2)-2d(p_1^2+p_2^2)\sigma^2_{W_0}+6\sigma^4_{W_0}\right)\left(e^{-\frac{(p_1^2+p_2^2)d}{2\sigma^2_{W_0}}}\delta^d(p_1+p_4)\delta^d(p_2+p_3)\right.\right.\\ &\left.\times\,e^{-\frac{(p_1^2+p_2^2)d}{2\sigma^2_{W_0}}}\delta^d(p_1+p_3)\delta^d(p_2+p_4)\right)+\left(d^2(N+p_1^2p_3^2)-2d(p_1^2+p_3^2)\sigma^2_{W_0}+6\sigma^4_{W_0}\right)e^{-\frac{(p_1^2+p_3^2)d}{2\sigma^2_{W_0}}}\\ &\left.\times\,\delta^d(p_1+p_2)\delta^d(p_3+p_4)\right]+p_1\leftrightarrow p_2,p_3,p_4,\end{aligned} \tag{B.12}$$

using the same convention as Cos-net.

B.4. Non-Gaussianity from non-identical parameter distributions

We discussed examples of NN architectures where non-Gaussianities arise at various widths, from the choice of identical but correlated parameter distributions. In addition to this, it is possible to violate CLT through independently drawn dissimilar NN parameter distributions; this too induces non-Gaussianities in NN output distributions. Let us present an architecture where non-Gaussianities at infinite width limit arise due to dissimilar *and* independent parameter distributions. Consider the NN architecture with output

$$\phi(x_k)=\sum_{j=-N}^{N}e^{-\frac{j^2}{\sigma^2}}W^L_jh^{L-1}_j(x_k)+b^L, \tag{B.13}$$

with parameters drawn from $W^1\sim\mathcal{N}(0,\sigma^2_{W_L})$, $b^L\sim\mathcal{N}(0,\sigma^2_{b_L})$, and $h^{L-1}_j(x_k)$ denotes the output of jth neuron in $(L-1)$th hidden layer, from input x_k. The presence of the prefactor $e^{-\frac{j^2}{\sigma^2}}$ at the jth node of final linear layer leads to dissimilarities in the final layer parameter distributions. Let us study the first three leading order cumulants at $\lim N\to\infty$,

$$G^{(2)}_c(x_1,x_2)=\lim_{N\to\infty}\sum_{j=-N}^{N}e^{-\frac{2j^2}{\sigma^2}}\sigma^2_{W_L}\mathbb{E}[h^{L-1}_j(x_1)h^{L-1}_j(x_2)]=\sqrt{\frac{\pi}{2}}\sigma\,\sigma^2_{W_L}\mathbb{E}[h(x_1)h(x_2)], \tag{B.14}$$

$$G_c^{(4)}(x_1,\cdots,x_4)=\sqrt{\frac{\pi}{4}}\sigma\,\sigma_{W_L}^4\Big[3\,\mathbb{E}[h(x_1)\cdots h(x_4)]-\sum_{\mathcal{P}(abcd)}\mathbb{E}[h(x_a)h(x_b)]\mathbb{E}[h(x_c)h(x_d)]\Big], \tag{B.15}$$

and

$$G_c^{(6)}(x_1,x_2,x_3,x_4,x_5,x_6)=\sqrt{\frac{\pi}{6}}\sigma\,\sigma_{W_L}^6\Big[15\,\mathbb{E}[h(x_1)h(x_2)h(x_3)h(x_4)h(x_5)h(x_6)]-3\sum_{\mathcal{P}(abcdef)}$$
$$\big(\mathbb{E}[h(x_a)h(x_b)h(x_c)h(x_d)]\mathbb{E}[h(x_e)h(x_f)]-2\,\mathbb{E}[h(x_a)h(x_b)]\mathbb{E}[h(x_c)h(x_d)]\mathbb{E}[h(x_e)h(x_f)]\big)\Big]. \tag{B.16}$$

We used $h(x):=h^{L-1}(x)$, and identities $\mathbb{E}[(W_j^L)^6]=15\,\sigma_{W_1}^6$, $\mathbb{E}[(W_j^L)^4]=3\,\sigma_{W_1}^4$. All these cumulants are nonvanishing at $\lim N\to\infty$; similarly, one can show that other higher order cumulants are non-vanishing too, adding non-Gaussianities to the output distribution.

Appendix C. CGF and Edgeworth expansion for NNFT

We express the output of a single hidden layer width N NN as a sum over N continuous variables

$$\phi(x)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}h_i(x), \tag{C.1}$$

where $h_i(x)$ are the outputs of each neuron before they get summed up into the final output.

C.1. Finite N and I.I.D. parameters

The CGF for i.i.d. parameters $P(h;\vec{\alpha}=\vec{0})=\prod_{i=1}^{N}P_i(h_i)$ become the following

$$\begin{aligned}W_{\phi(x)}[J]=&\log\mathbb{E}\left[e^{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int dxJ(x)h_i(x)}\right]\\=&\log\left[\prod_{i=1}^{N}\int Dh_i\,P_i(h_i)\,\exp\left(\frac{1}{\sqrt{N}}\int dxJ(x)h_i(x)\right)\right]\\=&N\log\left[\sum_{r=0}^{\infty}\prod_{i=1}^{r}\int dx_i\,\frac{G_{h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{r!N^{r/2}}\right]\\=&\sum_{r=0}^{\infty}\prod_{i=1}^{r}\int dx_i\,\frac{G_{c,h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{r!N^{r/2-1}}\end{aligned} \tag{C.2}$$

where $J(x)$ and $h_i(x)$ are the source current and output of ith neuron, respectively. In the second last step, we have used the following relation,

$$\sum_{r=0}^{\infty}\prod_{i=1}^{r}\int dx_i\,\frac{G_{h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{r!N^{r/2}}=e^{\sum_{r=0}^{\infty}\frac{1}{r!N^{r/2}}\int\left(\prod_{i=1}^{r}dx_i\right)G_{c,h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}. \tag{C.3}$$

Lastly, we use $W[J]=\sum_{r=0}^{\infty}\left(\prod_{i=1}^{r}\int dx_i\right)\frac{G_c^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{r!}$ to obtain

$$G_c^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)=\frac{G_{c,h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{N^{r/2-1}}, \tag{C.4}$$

with a N-scaling of cumulants, as expected.

C.2. Correlated parameters at finite N

Let $\vec{\alpha}=\{\alpha_1,\cdots,\alpha_q\}$ be parameters breaking statistical independence between neurons. For small $\vec{\alpha}$, one can write

$$P(h;\vec{\alpha})=P(h;\vec{\alpha}=\vec{0})+\sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q}\frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}\partial_{\alpha_{s_1}}\cdots\partial_{\alpha_{s_r}}P(h;\vec{\alpha})\Big|_{\vec{\alpha}=0}. \tag{C.5}$$

One can define the rth derivative as $\partial_{\alpha_{s_1}}\cdots\partial_{\alpha_{s_r}}P(h;\vec{\alpha})=P(h;\vec{\alpha})\mathcal{P}_{r,\{s_1,\cdots,s_r\}}$; the recursive relation satisfied by $\mathcal{P}_{r,\{s_1,\cdots,s_r\}}$ is

$$\mathcal{P}_{r+1,\{s_1,\cdots,s_{r+1}\}}=\frac{1}{r+1}\sum_{\gamma=1}^{r+1}(\mathcal{P}_{1,s_\gamma}+\partial_{\alpha_{s_\gamma}})\mathcal{P}_{r,\{s_1,\cdots,s_{r+1}\}\setminus s_\gamma}. \tag{C.6}$$

With this, the NN parameter distribution can be expressed as

$$P(h;\vec{\alpha})=P(h;\vec{\alpha}=\vec{0})+\sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q}\frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}P(h;\vec{\alpha})\mathcal{P}_{r,\{s_1,\cdots,s_r\}}\big|_{\vec{\alpha}=0} \tag{C.7}$$

Next, let us derive the CGF for the NN functional distribution,

$$\begin{aligned}W_{\vec{\alpha}}[J]&=\log\left[\int Dh\,P(h;\vec{\alpha})\,e^{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\int dx\,h_i(x)J(x)}\right]\\&=\log\left[\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\left(1+\sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q}\frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}\mathcal{P}_{r,\{s_1,\cdots,s_r\}}\big|_{\vec{\alpha}=0}\right)e^{\frac{1}{\sqrt{N}}\int dx\,h_i(x)J(x)}\right]\right]\\&=\log\left[e^{W_{\text{free}}[J]}+\sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q}\frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int dx\,h_i(x)J(x)}\cdot\mathcal{P}_{r,\{s_1,\cdots,s_r\}}\big|_{\vec{\alpha}=0}\right]\right].\end{aligned} \tag{C.8}$$

The last line is obtained using

$$\begin{aligned}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int dx J(x)h_i(x)}\right]&=\exp\left(N\sum_{r=0}^{\infty}\int\prod_{i=1}^{r}dx_i\,\frac{G_{c,h_i}^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)}{r!N^{r/2-1}}\right)\\&=e^{W_{\text{free}}[J]}.\end{aligned} \tag{C.9}$$

At $\lim N\to\infty$, we obtain $W_{\text{free}}[J]=\int dx_1dx_2\frac{J(x_1)\,G_{c,h_i}^{(2)}(x_1,x_2)\,J(x_2)}{2}$.

The partition function of a FT is related to its CGF as

$$Z[J(x)]=\mathbb{E}[e^{i\int J(x)\phi(x)}]=\prod_{i=1}^{N}\int DhP_i(h_i)\,e^{\frac{i}{\sqrt{N}}\int dxJ(x)\,h_i(x)}. \tag{C.10}$$

Under the transformation $J\to iJ$, the CGF becomes,

$$W[J]=\sum_{r=1}^{\infty}\int\prod_{j=1}^{r}dx_j\,\frac{i^r}{r!}G_c^{(r)}(x_1,\cdots,x_r)J(x_1)\cdots J(x_r)=:\sum_{r=1}^{\infty}\int\prod_{j=1}^{r}dx_j\,\frac{i^r}{r!}G_c^{(r)}J_{\underline{r}}, \tag{C.11}$$

the inverse Fourier transform of which, up to renormalization, is

$$\begin{aligned}P[\phi]&\propto\int DJ\,e^{W[J]-i\int dxJ(x)\phi(x)}\\&=\int DJ\,e^{\sum_{r=1}^{\infty}\int dx_1\cdots dx_r\frac{i^r}{r!}G_c^{(r)}(x_1,\cdots,x_r)J_{\underline{r}}-i\int dxJ(x)\phi(x)}\\&=\int DJ\,e^{\sum_{r=3}^{\infty}\int dx_1\cdots dx_r\frac{i^r}{r!}G_c^{(r)}(x_1,\cdots,x_r)J_{\underline{r}}}e^{-i\int dxJ(x)\phi(x)}\;e^{\frac{i\int dx_1G_c^{(1)}(x_1)J_{\underline{1}}}{1!}-\frac{\int dx_1dx_2G_c^{(2)}(x_1,x_2)J_{\underline{2}}}{2!}}\\&=\int DJ\,e^{\sum_{r=3}^{\infty}\int dx_1\cdots dx_r\frac{(-1)^r}{r!}G_c^{(r)}(x_1,\cdots,x_r)\,\partial_{\underline{r}}}e^{-i\int dxJ(x)\phi(x)}\\&\quad\times e^{i\int dx_1J(x_1)G_c^{(1)}(x_1)-\frac{1}{2}\int dx_1dx_2J(x_1)G_c^{(2)}(x_1,x_2)J(x_2)},\end{aligned} \tag{C.12}$$

where $\partial_{\underline{r}}=\frac{\delta}{\delta\phi(x_1)}\cdots\frac{\delta}{\delta\phi(x_r)}$. Next, we evaluate the integral associated with the Gaussian process,

$$\int DJe^{-i\int dx_1J(x_1)\phi(x_1)+i\int dx_1J(x_1)G_c^{(1)}(x_1)-\frac{1}{2}\int dx_1dx_2J(x_1)G_c^{(2)}(x_1,x_2)J(x_2)}, \tag{C.13}$$

using a change of variables $J'(x) \to J(x) + i\int dx'\, G_c^{(2)}(x,x')^{-1}[\phi(x') - G_c^{(1)}(x')]$ that keeps the measure of the source $DJ \to DJ'$ invariant. We obtain

$$\begin{aligned}
-S_G &= -i\int dx J(x)[\phi(x) - G_c^{(1)}(x)] - \frac{1}{2}\int dx_1 dx_2\, J(x_1)\, G_c^{(2)}(x_1,x_2)\, J(x_2) \\
&= -\frac{1}{2}\int dx_1\, dx_2 \left[J(x_1) + i\int dx'\, G_c^{(2)}(x_1,x')^{-1}[\phi(x') - G_c^{(1)}(x')]\right] G_c^{(2)}(x_1,x_2)\left[J(x_2)\right. \\
&\quad \left. + i\int dx''\, G_c^{(2)}(x_2,x'')^{-1}[\phi(x'') - G_c^{(1)}(x'')]\right] - \frac{1}{2}\int dx''\, dx'\, dx_1\, dx_2\, [\phi(x'') - G_c^{(1)}(x'')] \\
&\quad \times G_c^{(2)}(x'',x_2)^{-1}\, G_c^{(2)}(x_2,x_1) G_c^{(2)}(x_1,x')^{-1}\, [\phi(x') - G_c^{(1)}(x')] \\
&= -\frac{1}{2}\int dx_1\, dx_2\, J'(x_1)\, G_c^{(2)}(x_1,x_2)\, J'(x_2) - \frac{1}{2}\int dx dx'\, [\phi(x) - G_c^{(1)}(x)]\, G_c^{(2)}(x,x')^{-1} \\
&\quad \times [\phi(x') - G_c^{(1)}(x')]
\end{aligned} \tag{C.14}$$

An integration over J' results in the distribution

$$e^{-\frac{1}{2}\int dx dx'\,[\phi(x) - G_c^{(1)}(x)]\, G_c^{(2)}(x,x')^{-1}\,[\phi(x') - G_c^{(1)}(x')]},$$

such that

$$P[\phi] = e^{\sum_{r=3}^{\infty}\int dx_1\cdots dx_r\, \frac{(-1)^r}{r!} G_c^{(r)}(x_1,\cdots,x_r)\,\partial_L}\, e^{-\frac{1}{2}\int dx dx'\,[\phi(x) - G_c^{(1)}(x)]\, G_c^{(2)}(x,x')^{-1}\,[\phi(x') - G_c^{(1)}(x')]}. \tag{C.15}$$

We obtain perturbative corrections around the Gaussian field density by expanding the first exponential term in (C.15) as a series; contributions from higher order cumulants become increasingly less significant in most cases.

C.3. 4-pt function at finite N, non-i.i.d. parameters

Next, we evaluate the 4-pt function of this NNFT with the following CGF

$$W_\phi[J] = \log\left[e^{W_{\phi,\vec{\alpha}=0}[J]} + \sum_{r=1}^{\infty}\sum_{s_1,\cdots,s_r=1}^{q} \frac{\alpha_{s_1}\cdots\alpha_{s_r}}{r!}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\cdot \mathcal{P}_{r,\{s_1,\cdots,s_r\}}\big|_{\vec{\alpha}=0}\right]\right]. \tag{C.16}$$

For appropriately small $\vec{\alpha}$, the ratio of the second term in the logarithm to the first is small, and one can Taylor expand $\log(1+x) \approx x$ to obtain,

$$W_\phi[J] = W_{\phi,\vec{\alpha}=0}[J] + \sum_{s=1}^{q}\frac{\alpha_s}{e^{W_{\phi,\vec{\alpha}=0}[J]}}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\cdot \mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]. \tag{C.17}$$

The 4-pt function is obtained as $G_c^{(4)}(x_1,\cdots,x_4) = \frac{\partial^4 W_\phi[J]}{\partial J(x_1)\cdots\partial J(x_4)}\big|_{J=0}$. We abbreviate

$$M = \sum_{s=1}^{q}\frac{\alpha_s}{e^{W_{\phi,\vec{\alpha}=0}[J]}}\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\cdot \mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right], \tag{C.18}$$

then, $G_c^{(4)}(x_1,\cdots,x_4) = \frac{\partial^4 W_{\phi,\vec{\alpha}=0}[J]}{\partial J(x_1)\cdots\partial J(x_4)}\Big|_{J=0} + \frac{\partial^4 M}{\partial J(x_1)\cdots\partial J(x_4)}\Big|_{J=0}$.

Next, we evaluate the fourth J-derivative of M and turn the source J off,

$$\begin{aligned}
\frac{\partial^4 M}{\partial J_1\cdots\partial J_4}\Big|_{J=0} &= \sum_{s=1}^{q}\frac{\alpha_s}{e^{W_{\phi,\vec{\alpha}=0}[J]}}\left(\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^d x_1\cdots d^d x_4\,\frac{h_i(x_1)\cdots h_i(x_4)}{N^2}\, e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\right. \\
&\quad + \sum_{\mathcal{P}(abce)}\left[\left(\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_b} - \frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a \partial J_b}\right)\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^d x_c d^d x_e \frac{h_i(x_c)h_i(x_e)}{N}\right.\right. \\
&\quad \left.\cdot e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\cdot \mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right] - \left(\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_b}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_c} - \frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a \partial J_b}\right. \\
&\quad \left.\left.\left.\cdot\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_c} + \frac{\partial^3 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a \partial J_b \partial J_c}\right)\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^d x_e \frac{h_i(x_e)}{\sqrt{N}} e^{\frac{1}{\sqrt{N}}\int d^d x h_i(x) J(x)}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\right]
\end{aligned}$$

$$+\Bigg(\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_1}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_2}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_3}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_4}+\frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a\partial J_b}\frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_c\partial J_e}$$
$$-\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_b}\frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_c\partial J_e}+\frac{\partial^3 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a\partial J_b\partial J_c}\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_e}-\frac{\partial^4 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_1\partial J_2\partial J_3\partial J_4}\Bigg)$$
$$\times\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[e^{\frac{1}{\sqrt{N}}\int d^dx h_i(x)J(x)}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]-\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^dx_b d^dx_c d^dx_e\frac{h_i(x_b)h_i(x_c)h_i(x_e)}{N^{3/2}}\right.$$
$$\left.\times e^{\frac{1}{\sqrt{N}}\int d^dx h_i(x)J(x)}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a}\Bigg]\Bigg)\Bigg|_{J=0}:=\vec{\alpha}\cdot\Delta G_c^{(4)}(x_1,\cdots,x_4),$$

where we use the abbreviation $J(x_i):=J_i$. In the mean-free case,

$$\frac{\partial W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a}\Big|_{J=0}=\frac{\partial^3 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a\partial J_b\partial J_c}\Big|_{J=0}=0, \tag{C.19}$$

$$\frac{\partial^4 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_1\partial J_2\partial J_3\partial J_4}\Big|_{J=0}=G_c^{(4),\text{i.i.d.}}(x_1,\cdots,x_4),\quad\frac{\partial^2 W_{\phi,\vec{\alpha}=0}[J]}{\partial J_a\partial J_b}\Big|_{J=0}=G_c^{(2),\text{i.i.d.}}(x_a,x_b). \tag{C.20}$$

Thus, the 4-pt function is

$$\begin{aligned}G_c^{(4)}(x_1,\cdots,x_4)&=G_c^{(4),\text{i.i.d.}}(x_1,\cdots,x_4)\\&+\sum_{s=1}^{q}\frac{\alpha_s}{e^{W_{\phi,\vec{\alpha}=0}[J=0]}}\Bigg(\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^dx_1\cdots d^dx_4\frac{h_i(x_1)\cdots h_i(x_4)}{N^2}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\\&+\sum_{\mathcal{P}(abce)}\Bigg[-G_c^{(2),\text{i.i.d.}}(x_a,x_b)\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\int d^dx_c d^dx_e\frac{h_i(x_c)h_i(x_e)}{N}\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\\&+\Big(G_c^{(2),\text{i.i.d.}}(x_a,x_b)G_c^{(2),\text{i.i.d.}}(x_c,x_e)-G_c^{(4),\text{i.i.d.}}(x_1,\cdots,x_4)\Big)\prod_{i=1}^{N}\mathbb{E}_{P_i(h_i)}\left[\mathcal{P}_{1,s}\big|_{\vec{\alpha}=0}\right]\Bigg]\Bigg),\\&=G_c^{(4),\text{i.i.d.}}(x_1,\cdots,x_4)+\vec{\alpha}\cdot\Delta G_c^{(4)}(x_1,\cdots,x_4)+O(\vec{\alpha}^2).\end{aligned} \tag{C.21}$$

at leading order.

Appendix D. Fourier transformation trick for $G_c^{(2)}(x,y)^{-1}$

Let us evaluate the expression

$$\int dy_1\cdots dy_n\,G_c^{(n)}(y_1,\cdots,y_n)\,G_c^{(2)}(y_1,x_1)^{-1}\cdots G_c^{(2)}(y_n,x_n)^{-1}, \tag{D.1}$$

when $G_c^{(2)}(y_i,x_i)^{-1}$ involves differential operators. The integrals over y_i cannot be directly evaluated as the eigenvalues of each $G_c^{(2)}(y_i,x_i)^{-1}$ are unknown. To avoid this problem, we substitute the operators and cumulant with their Fourier transformations,

$$\begin{aligned}&\int d^dy_1\cdots d^dy_n\,d^dp_1\cdots d^dp_n\,d^dq_1\cdots d^dq_n\,d^dr_1\cdots d^dr_n\,\tilde{G}_c^{(n)}(p_1,\cdots,p_n)\,\tilde{G}_c^{(2)}(q_1,r_1)^{-1}\\&\quad\cdots\tilde{G}_c^{(2)}(q_n,r_n)^{-1}\,e^{iy_1(p_1+q_1)+ir_1x_1\cdots+iy_1(p_n+q_n)+ir_nx_n}\\&=\int d^dp_1\cdots d^dp_n\,d^dr_1\cdots d^dr_n\,\tilde{G}_c^{(n)}(p_1,\cdots,p_n)\,\tilde{G}_c^{(2)}(-p_1,r_1)^{-1}\cdots\tilde{G}_c^{(2)}(-p_n,r_n)^{-1}e^{i\sum_{j=1}^{n}r_jx_j}.\end{aligned} \tag{D.2}$$

Here $\tilde{f}$ is the Fourier transformation of f, and we obtained the second line by evaluating y_i integrals to get $\delta^d(p_i+q_i)$, then integrating q_i variables.

When $G_c^{(2)}$ is translation invariant, we have $G_c^{(2)}(y_i,x_i)^{-1}\propto\delta^d(y_i-x_i)$, leading to further simplification of the above expression as,

$$\int d^dp_1\cdots d^dp_n\,\tilde{G}_c^{(n)}(p_1,\cdots,p_n)\,\tilde{G}_c^{(2)}(-p_1)^{-1}\cdots\tilde{G}_c^{(2)}(-p_n)^{-1}e^{-ip_1x_1\cdots-ip_nx_n}. \tag{D.3}$$

We exemplify this expression for Cos-net and Gauss-net architectures.

Appendix E. Gaussian processes: locality and translation invariance

Any Gaussian process (GP) can be described as a function space distribution given by action S,

$$S = \int dx\, dy\, f(x)\, G_c^{(2)}(x,y)^{-1} f(y), \tag{E.1}$$

where $G_c^{(2)}(x,y)^{-1}$ is the *precision function*, related to the GP kernel by the inversion formula

$$\int dy\, G_c^{(2)}(x,y)^{-1} K(y,z) = \delta(x-z). \tag{E.2}$$

A *local* GP can be defined as a family of functions with a completely diagonalizable precision operator, resulting in the action

$$S = \int dx f(x)\, G_c^{(2)}(x)^{-1} f(x), \tag{E.3}$$

with the inversion relation simplified into

$$G_c^{(2)}(x)^{-1} K(x,z) = \delta(x-z). \tag{E.4}$$

This can be seen by considering $G_c^{(2)}(x,y)^{-1} = \delta(x-y)\Sigma(x)$ and performing the integral over y in equation (E.2). A Gaussian process can always be written in a local basis, as we will show below.

E.1. Gaussian process action in the local basis

Any Gaussian Process $f(x)$, when evaluated at a discrete set of inputs $\{x_i\}_i$, forms a multivariate Gaussian distribution. The covariance matrix of a multivariate Gaussian is a real symmetric matrix, and thus can be diagonalized. We can use this diagonalization procedure on the Gaussian process distribution itself, thereby rewriting it with a kernel proportional to a Dirac delta function,

$$\begin{aligned} S &= -\frac{1}{2}\int d^d x_i\, d^d x_l f(x_i)\, G_c^{(2)}(x_i,x_l)^{-1} f(x_l), \\ &= -\frac{1}{2}\int d^d x_i\, d^d x_j d^d x_k d^d x_l f(x_i) V(x_i,x_j) D(x_j,x_k) V^{-1}(x_k,x_l) f(x_l), \\ &= -\frac{1}{2}\int d^d x_k \left[\int d^d x_i\, V(x_i,x_k) f(x_k)\right] \Sigma(x_k) \left[\int d^d x_l\, V^{-1}(x_k,x_l) f(x_l)\right], \\ &= -\frac{1}{2}\int d^d x\, \phi^T(x)\Sigma(x)\phi(x), \end{aligned} \tag{E.5}$$

where $\phi(x) := \int d^d y f(y) V(y,x)$ and last step of (E.5) is obtained by $x_k \to x$. $D(x,y)$ is defined as $D(x_i,x_l) = \delta(x_i - x_l)\Sigma(x_i) = \int d^d x_j d^d x_k\, V^{-1}(x_i,x_j) G_c^{(2)}(x_j,x_k)^{-1} V(x_k,x_l)$.

ORCID iDs

James Halverson https://orcid.org/0000-0003-0535-2622
Anindita Maiti https://orcid.org/0000-0002-4712-6626

References

[1] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436
Goodfellow I J, Bengio Y and Courville A 2016 *Deep Learning* (MIT Press)
[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L U and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Curran Associates, Inc.)
Silver D *et al* 2017 Mastering the game of Go without human knowledge *Nature* **550** 354
[3] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 Machine learning and the physical sciences *Rev. Mod. Phys.* **91** 045002
[4] Carifio J, Halverson J, Krioukov D and Nelson B 2017 Machine learning in the string landscape *J. High Energy Phys.* JHEP09(2017)157
Gukov S, Halverson J, Ruehle F and Sułkowski P 2021 Learning to unknot *Mach. Learn.: Sci. Technol.* **2** 025035
Davies A *et al* 2021 Advancing mathematics by guiding human intuition with AI *Nature* **600** 70
Gukov S, Halverson J, Manolescu C and Ruehle F 2023 Searching for ribbons with machine learning (arXiv:2304.09304)
[5] Neal R M 1995 Bayesian learning for neural networks *PhD Thesis* University of Toronto

[6] 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)
Novak R, Xiao L, Lee J, Bahri Y, Abolafia D A, Pennington J and Sohl-Dickstein J 2018 Bayesian convolutional neural networks with many channels are Gaussian processes (arXiv:1810.05148)
Garriga-Alonso A, Aitchison L and Rasmussen C E 2019 Deep convolutional networks as shallow Gaussian processes (arXiv:1808.05587)
[7] Yang G 2019 Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation (arXiv:1902.04760)
Yang G 2019 Tensor programs I: wide feedforward or recurrent neural networks of any architecture are Gaussian processes (arXiv:1910.12478 [cs.NE])
Yang G 2020 Tensor programs II: neural tangent kernel for any architecture (arXiv:2006.14548)
[8] Maiti A, Stoner K and Halverson J 2021 Symmetry-via-duality: invariant neural network densities from parameter-space correlators (arXiv:2106.00694 [cs.LG])
[9] Williams C K 1997 Computing with infinite networks *Advances in Neural Information Processing Systems* pp 295–301
[10] Halverson J 2021 Building quantum field theories out of neurons (arXiv:2112.04527 [hep-th])
[11] Naveh G, David O B, Sompolinsky H and Ringel Z 2021 Predicting the outputs of finite deep neural networks trained with noisy gradients *Phys. Rev.* E **104** 064301
[12] Halverson J, Maiti A and Stoner K 2021 Neural networks and quantum field theory *Mach. Learn.: Sci. Technol.* **2** 035002
[13] Fukushima K 1975 Cognitron: a self-organizing multilayered neural network *Biol. Cybern.* **20** 121
Fukushima K 1980 Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position *Biol. Cybern.* **36** 193
Rumelhart D E, Hinton G E and Williams R J 1985 Learning internal representations by error propagation *Technical Report* (California University San Diego La Jolla Institute for Cognitive Science)
[14] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278
LeCun Y, Haffner P, Bottou L and Bengio Y 1999 Object recognition with gradient-based learning *Shape, Contour and Grouping in Computer Vision* (Springer) pp 319–45
[15] Bruna J, Zaremba W, Szlam A and LeCun Y 2013 Spectral networks and locally connected networks on graphs (arXiv:1312.6203)
Henaff M, Bruna J and LeCun Y 2015 Deep convolutional networks on graph-structured data (arXiv:1506.05163)
Duvenaud D K, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A and Adams R P 2015 Convolutional networks on graphs for learning molecular fingerprints *Advances in Neural Information Processing Systems vol 28* ed C Cortes, N D Lawrence, D D Lee, M Sugiyama and R Garnett (Curran Associates, Inc.) pp 2224–32
Li Y, Tarlow D, Brockschmidt M and Zemel R 2015 Gated graph sequence neural networks (arXiv:1511.05493 [cs.LG])
Defferrard M, Bresson X and Vandergheynst P 2016 Convolutional neural networks on graphs with fast localized spectral filtering *Advances in Neural Information Processing Systems* pp 3844–52
Kipf T N and Welling M 2016 Semi-supervised classification with graph convolutional networks (arXiv:1609.02907)
[16] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*
Huang G, Liu Z, van der Maaten L and Weinberger K Q 2016 Densely connected convolutional networks (arXiv:1608.06993 [cs.CV])
[17] Bahdanau D, Cho K and Bengio Y 2014 Neural machine translation by jointly learning to align and translate (arXiv:1409.0473)
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5998–6008
[18] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift (arXiv:1502.03167)
Ba J L, Kiros J R and Hinton G E 2016 Layer normalization (arXiv:1607.06450 [stat.ML])
[19] Hron J, Bahri Y, Sohl-Dickstein J and Novak R 2020 Infinite attention: NNGP and NTK for deep attention networks *Int. Conf. on Machine Learning* pp 4376–86
[20] Dinan E, Yaida S and Zhang S 2023 Effective theory of transformers at initialization (arXiv:2304.02034 [cs.LG])
[21] Yaida S 2019 Non-Gaussian processes and neural networks at finite widths (arXiv:1910.00019)
[22] Antognini J M 2019 Finite size corrections for neural network Gaussian processes (arXiv:1908.10030 [cs.LG])
[23] Roberts D A, Yaida S and Hanin B 2022 *The Principles of Deep Learning Theory* (Cambridge University Press)
[24] Dyer E and Gur-Ari G 2020 Asymptotics of wide networks from Feynman diagrams (arXiv:1909.11304)
[25] Erdmenger J, Grosvenor K T and Jefferson R 2021 Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group (arXiv:2107.06898 [hep-th])
Grosvenor K T and Jefferson R 2022 The edge of chaos: quantum field theory and deep neural networks (arXiv:2109.13247 [hep-th])
[26] Erbin H, Lahoche V and Samary D O 2022 Non-perturbative renormalization for the neural network-QFT correspondence *Mach. Learn.: Sci. Technol.* **3** 015027
Erbin H, Lahoche V and Samary D O 2022 Renormalization in the neural network-quantum field theory correspondence (arXiv:2212.11811 [hep-th])
[27] Banta I, Cai T, Craig N and Zhang Z 2023 Structures of neural network effective theories (arXiv:2305.02334 [hep-th])
[28] Jacot A, Gabriel F and Hongler C 2018 Neural tangent kernel: convergence and generalization in neural networks *NeurIPS*
[29] Arora S, Du S S, Hu W, Li Z, Salakhutdinov R and Wang R 2019 On exact computation with an infinitely wide neural net (arXiv:1904.11955 [cs.LG])
[30] Du S S, Hou K, Póczos B, Salakhutdinov R, Wang R and Xu K 2019 Graph neural tangent kernel: fusing graph neural networks with graph kernels (arXiv:1905.13192 [cs.LG])
[31] Alemohammad S, Wang Z, Balestriero R and Baraniuk R 2021 The recurrent neural tangent kernel (arXiv:2006.10246 [cs.LG])
Alemohammad S, Balestriero R, Wang Z and Baraniuk R 2021 Enhanced recurrent neural tangent kernels for non-time-series data (arXiv:2012.04859 [cs.LG])
[32] Lee J, Xiao L, Schoenholz S S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2019 Wide neural networks of any depth evolve as linear models under gradient descent (arXiv:1902.06720)

[33] Huang J and Yau H-T 2020 Dynamics of deep neural networks and neural tangent hierarchy *Proc. 37th Int. Conf. on Machine Learning* ed H III Daumé and A Singh (PMLR) pp 4542–51
Aitken K and Gur-Ari G 2020 On the asymptotics of wide networks with polynomial activations (arXiv:2006.06687 [cs.LG])
[34] Bordelon B and Pehlevan C 2023 Dynamics of finite width kernel and prediction fluctuations in mean field neural networks (arXiv:2304.03408 [stat.ML])
[35] Zavatone-Veth J, Canatar A, Ruben B and Pehlevan C 2021 Asymptotics of representation learning in finite Bayesian neural networks *Advances in Neural Information Processing Systems* ed M Ranzato, A Beygelzimer, Y Dauphin, P Liang and J W Vaughan (Curran Associates, Inc.) pp 24765–77
[36] Seroussi I, Naveh G and Ringel Z 2022 Separation of scales and a thermodynamic description of feature learning in some CNNs (arXiv:2112.15383 [stat.ML])
[37] Krippendorf S and Spannowsky M 2022 A duality connecting neural network and cosmological dynamics (arXiv:2202.11104 [gr-qc])
[38] Fischer K, René A, Keup C, Layer M, Dahmen D and Helias M 2022 Decomposing neural networks as mappings of correlation functions *Phys. Rev. Res.* **4** 043143
Dick M, van Meegen A and Helias M 2023 Linking network and neuron-level correlations by renormalized field theory (arXiv:2309.14973 [cond-mat.dis-nn])
Huang H 2018 Mechanisms of dimensionality reduction and decorrelation in deep neural networks *Phys. Rev.* E **98** 062313
[39] Albergo M, Kanwar G and Shanahan P 2019 Flow-based generative models for Markov chain Monte Carlo in lattice field theory *Phys. Rev.* D **100** 034515
Abbott R *et al* 2022 Gauge-equivariant flow models for sampling in lattice field theories with pseudofermions *Phys. Rev.* D **106** 074506
Gerdes M, de Haan P, Rainone C, Bondesan R and Cheng M C N 2022 Learning lattice quantum field theories with equivariant continuous flows (arXiv:2207.00283 [hep-lat])
[40] Osterwalder K and Schrader R 1973 Axioms for Euclidean Green's functions *Commun. Math. Phys.* **31** 83
[41] Dedecker J, Doukhan P, Lang G, José Rafael L R, Louhichi S and Prieur C 2007 Central Limit theorem *Weak Dependence: With Examples and Applications* (Springer) pp 153–97
[42] Hanin B 2023 Random fully connected neural networks as perturbatively solvable hierarchies (arXiv:2204.01058 [math.PR])
[43] McCullagh P 1987 *Tensor Methods in Statistics* (Chapman and Hall)
[44] Weinberg S 1995 *The Quantum Theory of Fields* (Cambridge University Press)