Article Report: scipost-202110-00024 (v1)

Date:	12/12/2021
Title:	The edge of chaos: Quantum field theory and deep neural networks
Author(s):	Kevin T. Grosvenor, Ro Jefferson
Arxiv:	2109.13247

Summary This paper aims at deriving the quantum field theory associated with a statistical ensemble of recurrent or fully connected neural networks from first principles. The idea of the paper is to start with the description of the ensemble in terms of a stochastic differential equation to which one can associate a path integral, and reinterpret it as a field theory. Using ideas from statistical physics, it is then possible to compute different parameters such as the Lyapunov exponent and correlation length in terms of the network parameters (such as the standard deviations of the weights and biases) to characterize the properties of the network.

This is an important topic in its infancy, and this paper deserves careful attention. However, the paper in its current form is difficult to follow and the computations do not seem as general as suggested in the abstract and introduction. It is also difficult to see what concrete applications can be. As a consequence, before recommending it for publication, I would recommend a major revision of the presentation.

Strengths

1. Study from first-principle the NN-QFT correspondence, which is an important topic.
2. Computations are overall clear and detailed.

Limitations

1. The presentation is very unbalanced.
2. Many assumptions are introduced along the way without proper discussion.
3. There are no numerical tests to support the approach of the paper.
4. It is not clear how the results are connected to concrete applications.

Review I have some general remarks about the methodology of the paper:

1. The authors introduce many assumptions at various stages of the paper. While they are necessary to perform analytic computation, it seems that they restrict a lot the original claims (from the abstract and introduction) and make obscure the physical meaning of the computations. In particular, each assumption is introduced as a minor

technical assumption without taking into account the bigger perspective. While the authors come back on some assumptions in the conclusion, I feel that they should be discussed earlier, possibly all at the same place before starting the computations.

From a more general point of view, I think it is important for physicists applying physics methods to computer science to remember that, in the end, what matters is to make contact with the "real" world and not just build an abstract formalism. For example, a lot of work has been done at the end of the '80 to understand neural networks with statistical physics, but this has played no role in the recent resurgence of neural networks and the design of new architectures.

2. sec. 2 to 4: I would suggest clarifying the role of the time t, the different assumptions which are made, and what it means physically. While the assumptions are clearly needed to make progress, it is important to understand what it means for concrete neural networks and how much we can expect the current computations to be valid for numerical experiments.

   (a) sec. 2.1: The neural network is originally described by a set of hidden layers indexed by $t \in \{0, \dots, T\}$. To reach eq. (2.9), the continuous-time limit is taken. While it is a useful assumption and allows to reuse the formalism of path integral for stochastic process, this is not easy to interpret. I can see how it would make sense for a large set of layers, but I am still slightly uncomfortable (and even more after taking $T \to \infty$, see below).

   (b) sec. 2.3, p. 16, above eq. (2.41): The authors introduce the assumption that

   "(...) the system exhibits time translation symmetry (...)"

   While this is necessary for most of the computations of the paper, this is a very strong assumption and I would suggest spending more time discussing it and what this means (this is done in the conclusion, p. 59 §2 but this appears far too late given the importance of this condition). Boundaries are breaking time translation invariance, so it means that time becomes either periodic or non-compact (the second being chosen later). What does this mean in terms of neural networks (beyond what is said in the conclusion)?

   (c) sec. 4.3 and 4.4: this section assumes $T \to \infty$ such that T/N is fixed since the expansion is made in terms of the T/N. However, the fact that the expansion parameter is T/N is specified quite late, only in sec. 4.3.2, §1, p. 43. More problematic, the fact that $T \to \infty$ is specified only in footnote 37, p. 43 (though it is also said that it is kept finite to act as a regulator, sec. 4.3.2, §2, p. 43). As explained above, taking a non-compact continuous time makes the interpretation as a neural network difficult.

3. sec. 2.1, p. 6, below eq. (2.2): How restricting is the assumption of taking all layers to have the same width? Could we miss some effects for which relative changes in layer widths could be important (like autoencoders)?

4. sec. 2 and 3: There seems to be a problem with the definition of g. First, in sec. 2.1, g is a function of h and x, see eq. (2.3) or (2.13), and the shortcut $g_t := g(h_t, x_t)$ is introduced above (2.7). Later, it is written as $g(t)$, see (2.34): while it seems to be just a generalization of the previous shortcut, my interpretation of (2.34) is that it is not the case, since below the equation it is written:

   "(...) we retain the freedom to choose the diffusion coefficient $g(t, x)$.

   which is a different notation from which h has disappeared.

   Second, below (3.3) it is written

   "(...) the term on the last line is the contribution from the common stochasticity $K_B(-g \sum_\alpha \tilde{z}^\alpha)$."

Since g depends on x and h (in principle) and both are double-copied (in eq. (3.3), both variables have an index α), then why is g common to both copies?

5. sec. 3.1, p. 26, below (3.33): The authors claim that their computations apply to fully connected networks (MLP). However, this seems to be a borderline case:

   > *"(...) in order to study networks at the edge of stability, we will henceforth consider the case with $\gamma > 0$."*

   whereas MLP are characterized by $\gamma = 0$, see below (2.16) p. 10.

6. Similarly, the authors write that they work for general activation functions. However, they make several hypotheses that restrict a lot the domain of application:

   (a) sec. 4, p. 31, eq. (4.13): The authors fix $\phi(h) = \varphi(h) = \tanh h$ and perform a Taylor expansion, keeping the first two non-trivial terms. At this stage, this is fine because it looks like the methods would generalize to other functions by just replacing the Taylor expansion.

   (b) sec. 4.1, p. 35, below eq. (4.36):

   > *"This would lead (...) in the Taylor expansion."*

   This paragraph is confusing because it looks like a choice, where it is really an assumption. As stated, more comments appear elsewhere (below (4.48) maybe?), but the tone is confusing.

   (c) sec. 4.1, p. 37, below eq. (4.48), §1:

   > *"Since $\phi(h) \in [-1, 1]$, we expect the next order term to be less important (...)"*

   which is a crucial assumption because it is not true for most activation functions. In particular, there is no small parameter in the argument of $\phi(h)$ so it is not consistent to truncate its expansion to a finite order (see for example the kind of computations with an exponential interaction in [1, p. 11]), except in the very specific case above where the value of the function is bounded.

   (d) sec. 4.4: How legal is it to omit higher-order corrections from the Taylor expansions? The argument from the previous point is approximately acceptable, but this looks more dubious for general activation functions. I understand that it may be very difficult to include other interactions, but it is important to discuss what is expected: would it be possible to classify the graphs according to the order of interactions they contain, or will all graphs be mixed (I expect the second case since there is no expansion parameter)?

   I would suggest to either work with more general activation functions by using a general Taylor expansion instead of (4.13):

   $$\phi(h) = \phi_0 + h\,\phi'(0) + h^2\,\phi''(0) + \cdots \tag{1}$$

   and to explain clearly the impact of the truncation in sec. 4.4, either to state in the abstract/introduction that they work only with $\phi(h) = \tanh h$ (or even more precisely with a cubic odd polynomial).

7. sec. 5, p. 58, §3:

   > *"This leads to the question whether such a theory is renormalizable: we have shown that the infinite series of corrections to the two-point function converge at weak coupling in T/N (...)"*

   This seems to be contradicted by the term in T^2/N found by the authors (which diverges as $T, N \to \infty$, with T/N fixed), see (4.76).

Clarity Overall, the text is quite unbalanced: it looks like the authors focus a lot on algebraic manipulations at the expense of conceptual explanations. Here is a list of suggestions for improving the presentation:

8. It seems that a lot of content in sections 2 and 3 has been reproduced from [45] and it is not always easy to understand what are the new contributions from the authors. Hence, I would suggest stating clearly what are the new results and formula of this paper. Moreover, if section 2 is strongly inspired from [45], it could be made a bit shorter (though it is useful to have it self-contained), see the comments below.

9. While I appreciate papers where computations are spelled out in a clear way and where the reader can easily follow each step, I found that it was slightly too explicit in the current paper. In particular, given the absence of figures and motivations for some conceptual aspects, this makes the paper looks quite unbalanced. Also since the paper is very long and seems to take materials from other sources like [45], I would suggest reducing the length of some computations. Here are a few examples:

   (a) sec. 2.2, p. 11 eq. (2.23): the second and third lines are just completing the square, this is so trivial that it can be omitted.

   (b) sec. 4, p. 33, e. (4.28): it would be much simpler to just take the Fourier transform of (4.24).

   (c) sec. 4, p. 34, e. (4.30): second equality is not necessary (the only change compared to the next line is $\mathrm{i} = -1/\mathrm{i}$).

10. sec. 1: The authors indicate that this paper is part of the NN-QFT correspondence, it is hard to how it is related to earlier papers such as [7-9] which stated the correspondence clearly.

11. sec. 1, p. 3, §2: From the opening sentence

    > *"In this work, we explicitly construct (. . .)"*

    it looks like the authors are building this field theory for the first time. However, it appears that it was done before in [44-45], which could be cited there.

12. sec. 2: I think it would be useful to summarize in words the method followed in sec. 2 to build the field theory (introducing auxiliary fields, etc.) at the beginning of the section, such that the reader has an idea of where the paper is going. Currently, it looks a lot like a series of formal manipulations and it is hard to get an intuition of why we do this and where we want to stop.

13. sec. 2.1, p. 6: The starting point of the whole paper is equations (2.1) and (2.2) which are stated without any motivation or intuition. Given how central it is to the paper and how other points are over-detailed (like completing a square), it would be very useful to spend some time introducing the model. In particular, how it is related to recurrent or fully connected networks? What is the interpretation of the different parameters A, B, etc.? At the top of p. 7, there is a brief note on MLP, however, this is not sufficient to completely characterize MLP; in fact more information is given below (2.16) after modifying (2.2) to (2.16): hence, I think it would be very useful to show how the MLP emerges as a concrete example. Figures could also help.

14. sec. 2.1, p. 6: It is not clear if the last hidden state h_T corresponds to the output layer or not. The output layer is not a hidden layer, so it seems that h_T should not be the last layer, but then how do we read the output values?

15. sec. 2.1, p. 6, eq. (2.2): Why introduce this functional form for f instead of (2.16) which is used in most of the paper? The form (2.2) for f does not seem important for sec. 2.1 and using a unique form would simplify the discussion and reduce information

overload. Moreover, the comment below (2.16) seems to indicate that the latter is more common in the literature.

16. sec. 2.1, p. 6: It is not clear if the last hidden state h_T corresponds to the output layer or not. The output layer is not a hidden layer, so it seems that h_T should not be the last layer, but then how do we read the output values?

17. sec. 2.2, p. 10, eq. (2.16): What is the intuition for / interpretation of the new parameter γ? Is it a fixed number (real, positive?), a statistical variable, a matrix? The paragraph below (2.23) seems to indicate that it is an arbitrarily fixed real "constant", but this should be explained earlier.

18. sec. 2.2, p. 10, eq. (2.18): This equation would be better below (2.15).

19. sec. 2.2, p. 12, eq. (2.25): I am quite confused by the phrase:

    > *"introducing the N^2-local field variables $\mathfrak{A}$, etc."*

    $$\mathfrak{A}(t_1, t_2) := \frac{\sigma_A^2}{N} \sum_j h_j(t_1) h_j(t_2) \tag{2}$$

    Indeed, it looks like $\mathfrak{A}$, etc., are each a single bi-local field: they depend on two times $t_1, t_2 \in [0, T]$, and the sum over $j = 1, \dots, N$ means that there is a single component. Given the sentence, I would have expected to see (still) bi-local matrix fields $\mathfrak{A}_{ij}(t_1, t_2) = h_i(t_1) h_j(t_2)$, as one sees for the general use of the Hubbard-Stratonovich transformation [2, sec. 21.6] (though given the structure of the Lagrangian it makes sense to introduce a single bi-local field).

20. sec. 2.2, p. 12, footnote 15: I am confused by the statement of the footnote: it says that

    > *"(...) N should be sufficiently large for the Gaussian distributions to be valid (...)"*

    but below (2.29) it is written that:

    > *"(...) up to this point, the result (2.28) is exact, subject to working with the ensemble average $\langle Z \rangle_{X,b}$."*

    The first sentence and the last part of the second seem to indicate some approximation, which seems to be in tension with saying that it is "exact". Maybe the interplay between the two sentences could be clarified.

21. sec. 2.2, p. 12, eq. (2.32): The input $x_i(t)$ vectors seem to be fixed and independent of h_i and $\tilde{z}_i$ and are not integrated over in the path integral (see previous equations), so what is the meaning of computing the expectation values? Moreover, the authors could note that the auxiliary fields $\mathfrak{B}$ and $\mathfrak{U}$ have a slightly different role compared to $\mathfrak{A}$ and $\mathfrak{W}$ since they are built out of non-dynamical variables but still introduced as new fields in the path integral.

22. sec. 2.3, p. 17-18, eq. (2.45) to (2.47): The paragraph above (2.45) is not very clear. Moreover, it would be helpful to explain in more detail how one arrives at the Gaussian measure (2.47) starting from the non-Gaussian measure (2.35).

23. sec. 3: Can you provide more intuition for using a double copy? I am quite confused by how to interpret it, especially in the context of neural networks.

    For example, I don't understand eq. (3.2): the parenthesis says

    > *"(...) between two identically-prepared copies of the system (...)"*

but one still considers different trajectories. So by "identically prepared", do you mean "same parameters for the path integral" but then we consider different trajectories (= solutions) finishing at different times?

24. sec. 3.1, p. 20, footnote 20: Writing $[h^(1), h^2(s)]$ is superfluous: it is obvious that $h^\alpha(t)$ is a real function and not an operator (which have not been used at all in this paper) and having a commutator here is more confusing than enlightening.

25. sec. 3.2, p. 25, eq. (3.23): The symbol $d(t)$ is ambiguous: the previous definition of d in had $d(t_1, t_2) = d(\tau)$, and below in (3.24) we have $\tau = t - s$.

26. sec. 3.2, p. 25, eq. (3.24): The symbol T is already used to denote the upper limit of t, so it would be clearer to introduce another symbol.

27. sec. 3.2, p. 26, above eq. (3.32): This is confusing because it sounds that all solutions to the time-independent Schrödinger equation are bound states in the present case, however, the sentence below (3.38) seems to indicate that scattering states are also possible. It would help to clarify this point and maybe discuss the properties of V'' (which allows it to have bound states) below (3.32).

28. sec. 3.2, p. 27, below eq. (3.38): I am confused by this paragraph and (3.38): the first sentence that the solution (3.38) is not the ground state, but the paragraph concludes by saying that it characterizes the edge of stability which in turn is a condition on the ground state, see below (3.33).

29. sec. 4, p. 28, §1:

    > *"(...) deviations from Gaussianity require the addition of corrections terms, corresponding to the fact that higher cumulants no longer vanish. In the language of QFT, these correspond to loop corrections to the leading order or tree-level result above."*

    This sentence is strange: for a free (Gaussian) QFT, cumulants (connected) Green functions vanish since the Green functions are given purely by disconnected two-point functions following Wick theorem. Thus, interactions give non-vanishing contributions to the cumulants. I am sure that the authors are aware of these facts, but this is not how the paragraph reads.

30. sec. 4.1, p. 37, below eq. (4.48), §2:

    > *"(...) we shall see that the weak coupling condition (...)"*

    Just to be sure, the "weak coupling" means T/N?

31. sec. 4.1, p. 37, below eq. (4.48), §3:

    > *"While this can be done analytically, the resulting expression is exceedingly lengthy and not particularly enlightening, so we refrain from including it here."*

    In view of all the details given in the paper, why omit here a formula?

32. sec. 4.2, p. 40:

    > *"(...) the double-line notation here closely resembles that introduced by 't Hooft (...)"*

    This is not a resemblance, this is exactly the same idea (a matrix field has two indices so a ribbon propagator; an n-tensor field has n strands).

33. sec. 4.4.2, p. 50, below eq. (4.76):

> *"Therefore, we will obtain a factor of $\delta(\omega = 0)$ when inverse Fourier transforming the product $c(\omega)^2$."*

The factor $\delta(0)$ appears already before the Fourier transform since $c(\omega) \sim \delta(\omega)$, so $c(\omega)^2 \sim \delta(0)\delta(\omega)$.

There are a few minor typos:

1. sec. 2.1, p. 7, eq. (2.5): Missing index on h and x in the arguments of f and g.
2. sec. 4, p. 30, eq. (4.10): $\langle \tilde{z}_i t)\rangle \rightarrow \langle \tilde{z}_i\,(\,t)\rangle$
3. sec. 4.1, p. 33, between eq. (4.26) and (4.28): there are various typos, most instances of x should be replaced by t except $\hat{f}(x) \rightarrow \hat{f}(\omega)$.
4. sec. 4.3.1, p. 42, eq. (4.53): there is an extra comma after $\mathrm{d}\tau''$

References

[1] T. Bautista and A. Dabholkar. "Quantum Cosmology Near Two Dimensions." In: *Physical Review D* 94.4 (2016), p. 044017. DOI: 10.1103/PhysRevD.94.044017. arXiv: 1511.07450.

[2] S. Weinberg. *The Quantum Theory of Fields, Volume 2: Modern Applications.* Cambridge University Press, May 2005.