Response to feedback & suggestions

Kevin T. Grosvenor and Ro Jefferson

scipost-202110-00024

Here we include a detailed response to the questions, comments, and requested changes in the attachment to Report #2 by Dr. Harold Erbin; our response to general comments is provided with the resubmission. For clarity, we have included the original text of said attachment in blue, with our responses to each point in black. Note that, relative to the original Report, an effort has been made to update page/equation numbers where necessary to match the resubmission (version 2).

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.

Authors' response: We agree with the referee that we do indeed make many assumptions along the way. Our intention, as we have now made explicit in the beginning of section 2, was to introduce assumptions only as needed in order to keep the analysis as general as possible. However, the paper certainly benefits from having these collected in one place. Since this necessarily involves a great deal of notation and technical detail, we have relegated this to a new appendix (A.1), rather than obstructing the flow of the paper by expanding the already long introduction. We have pointed the reader to said appendix at the very beginning of section 2, where we have also elaborated on the general strategy in order to help the reader keep the big picture in mind. We thank the referee for these helpful suggestions, which we believe have improved the work.

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).

Authors' response: As the referee states, the continuum limit only makes sense for a large number of layers, i.e., large T. So, in fact, one should be *less*, not more, uncomfortable after taking the $T \to \infty$ limit. There is no difficulty in interpretation here: when one takes the continuum limit, one immediately and implicitly states that one is interested in phenomena that occur on the scale of a large number of layers. (Formally of course, this simply amounts to recovering the initial SDE from the Ito discretization, as we have noted in footnote 12, where we also commented explicitly about there being a perfectly sensible continuum limit in the layer direction, but not in the width direction (number of neurons)). We have however added additional clarifications about the large T (and large N) regime of applicability in the introduction, appendix A, and multiple footnotes.

(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 par. 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)?

Authors' response: We agree that time-translation symmetry is an important assumption, and have added a detailed discussion about this in the new appendix A.1 (see our response to point 1). We have also added more details to the explanation about this in the introduction.

(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, par. 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, par. 2, p. 43). As explained above, taking a non-compact continuous time makes the interpretation as a neural network difficult.

Authors' response: In fact, we stated this in the abstract as well as in the introduction when discussing the organization of the paper, and have now added a more detailed discussion earlier in the introduction, namely that the appropriate expansion parameter is T/N as in the earlier work [2]. This statement also appears again at the very start of section 4, where appropriate. This should suffice to introduce the expansion parameter to the reader in advance of sec. 4.3.2. More intuition for the tradeoff between depth and width was also given on page 50 as well as in the Discussion (page 60), and is elaborated on in more detail in the new appendix A.

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)?

Authors' response: In the context of the previous point, $\gamma T/N$ is the relevant scale in the problem. If the variations δN in N are small compared to N, then these variations will not matter. Technically however, we want $\delta N \ll \gamma T$ so that the effect of $\delta N/N$ can be neglected relative to $\gamma T/N$. This is automatically satisfied in the $N \to \infty$ limit, *unless* one also lets $\delta N \to \infty$. This is a logical possibility, but an extremely unnatural one which has not been entertained by any previous works in the context of the infinite-width limit.

In the particular example of autoencoders, one would have to check how well this approximation is met, namely, is the difference between the width of the wide and narrow layers small compared to the number of layers? If so, then our theory should apply reasonably well within the bulk of the network, including for autoencoders. If this is not the case – as would be generically true for autoencoders with a very narrow layer – then our theory would only apply to the bulk of the network away from both the boundaries and the narrow layer. In short, the narrow layer present in standard autoencoders is obviously outside the regime with which we are concerned, as detailed in several places as mentioned above.

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 $K_B\big(?g \sum_\alpha \tilde{z}^\alpha\big)$."

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?

Authors' response: Following the standard convention, g_t denotes the discrete case, while $g(t)$ denotes the continuous case. However, $g(t, x)$ is simply a typo and should be $g(t)$; we are grateful to the referee for catching this. More importantly, we are also very grateful to the referee for catching the issue with g seemingly being dependent on the state of the network, despite being the same between two copies. This was poor notation on our part, which we have now amended; as the referee intuits, g is treated as an external parameter which is independent of the current state of the system. See the new comment below (2.3) (we have also remarked on this in the new appendix A).

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 hence- forth consider the case with $\gamma > 0$."

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

Authors' response: This is a typo: MLPs are characterized by $\gamma = 1$, not $\gamma = 0$, as stated below (3.39). Again we thank the referee for catching this.

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.

Authors' response: There is no assumption here, we are merely being explicit about the order in the Taylor expansion to which we work. We have discussed this in more detail below (4.48) and in the Discussion. Nevertheless, we agree with the referee that the language below (4.48) can be confusing: *tanh* is a choice, but the order in the Taylor series is an approximation. We have amended the text accordingly, and also added further discussion about truncating the Taylor expansion in appendix A.1, see below.

(c) sec. 4.1, p. 37, below eq. (4.48), par. 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.

Authors' response: We have replaced $\phi(h) \in [-1, 1]$ with $\tanh(h) \in [-1, 1]$ to make this more specific. However, the boundedness of the activation function is not the point: rather, as we have now elaborated in the text, the relevant feature is that the expectation value of h is 0 and its two-point function is small and decays exponentially over time. Therefore it is only necessary to keep a few terms in the Taylor expansion around $h = 0$. The only crucial assumption is that $\phi(0) = 0$, which we have also elaborated on in appendix A, with references to a more detailed discussion in [2].

(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).

Authors' response: Again, one expects that higher-order terms in the Taylor expansion of $\phi(h)$ do not matter for any well-behaved function $\phi(h)$, i.e., a function that does not vary an enormous amount on the scale of the square-root of the two-point function of h; see previous point. However the suggestion of the referee that would could consider a general Taylor expansion is interesting, and we have included this in the discussion in appendix A. We have also elaborated on the choice of activation function in the main text, specifically near the beginning of section 2 where it is first introduced, cf. the new (2.2) and subsequent discussion.

7. sec. 5, p. 58, par. 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).

Authors' response: We have addressed this in detail in the several paragraphs following (4.76), and have also added more discussion in the new appendix A. Let us summarize the discussion here: the diagram that naïvely appears to scale as T^2/N in fact scales as $(\frac{\gamma T}{N})(\frac{\sigma_{b,\text{eff}}}{\gamma})^4(\gamma T)$, cf. eq. (4.79). As we then explain, the weak coupling regime is governed *both* by σ_w^2/γ^2 *and* by σ_b^2/γ^2. Accordingly, $\sigma_{b,\text{eff}}/\gamma$ must be small enough so as to compensate for the extra factor of γT for the perturbative expansion to be well-behaved, and thus this diagram effectively scales as T/N, not T^2/N. In practice, σ_b^2/γ^2 is usually set to the order of 10^{-2}, and therefore, $\sigma_b^4/\gamma^4 \sim 10^{-4}$, with network depths on the order of at least 10^2. So this assumption is satisfied with a couple orders of magnitude to spare, and even deeper networks can be accommodated by decreasing σ_b^2.

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.

Authors' response: We have substantially expanded the text around "we shall draw heavily..." to indicate the main additions that we make relative to the formalism in [45,46], and had already stated twice in the introduction that the core of the formalism is not new and that we build on these previous works. We believe that these statements should collectively be sufficient to acknowledge previous work and to clarify our contributions. As the referee points out, it is useful to have everything self-contained, and we have presented a coherent and complete exposition for both pedagogical clarity as well as to establish our notation, and for this reason we prefer not to shorten the exposition; see 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. 13 eq. (2.22): the second and third lines are just completing the square, this is so trivial that it can be omitted.

(b) sec. 4, p. 33, eq. (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 $i = -1/i$).

Authors' response: We thank the referee for these suggestions to streamline some of the algebraic manipulations, and have implemented these suggestions.

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.

Authors' response: We have discussed this in considerable detail in the introduction, where we have already devoted significant space to the relation to previous work. Additionally however, following the excellent suggestion of referee 1, we have added a new appendix A in which the elements of the correspondence (i.e., the precise mappings) are clearly enumerated, which we believe has substantially improved the presentation.

11. sec. 1, p. 3, par. 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.

Authors' response: We have modified this sentence to more clearly acknowledge these important works.

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.

Authors' response: We thank the referee for this excellent suggestion, and have added a roadmap at the beginning of section 2 to make it easier for the reader to follow and to give an idea of where the lengthy derivations are going.

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.

Authors' response: This is also a good point: we have added a statement emphasizing that (2.1) is formally the most general SDE one can write down in this context. We have also replaced (2.2) with what was previously (2.16), as indeed the latter is the more standard convention in the literature with which we work. Additionally, we have substantially expanded upon the subsequent explanation for the various component appearing in this expression, including the reduction to an MLP.

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?

Authors' response: As we have stated previously (both here and in the paper), we are studying phenomena in the bulk of the network away from the boundaries (since, as the referee has already pointed out, the boundaries break time-translation invariance). We have elaborated on footnote 8 to make this more clear (i.e., the real input layer may be thought of as h_{-1}, while the real output layer may be thought of as h_{T+1}; both are immaterial to the analysis).

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.

Authors response: See point 13 above.

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?

Authors' response: This is a duplicate of point 14.

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.

Authors' response: We have elaborated on the roles and interpretations of the various parameters when introducing the update function (2.2), including γ, just below it; see point 13. Some additional intuition is given in the new appendix A.

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

Authors' response: We agree and have implemented this suggestion.

19. sec. 2.2, p. 12, eq. (2.24): I am quite confused by the phrase:
"introducing the N^2-local field variables 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 $j = 1, \ldots, N$ means that there is a single component. Given the sentence, I would have expected to see (still) bi-local matrix fields $A_{ij}(t_1, t_2) = h_i(t_1) h_j(t_2)$, 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).

Authors' response: We apologize for the poor phrasing in this case. We have replaced N^2-local with bi-local.

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.

Authors' response: Here, we are using the customary language in the field of large-N QFT, where once the limit of large N has been explicitly stated, it is no longer repeatedly referred to as an approximation and is taken for granted. However we have removed the "subject to working with the ensemble average" in the event that this may have caused confusion.

21. sec. 2.2, p. 12, eq. (2.32): The input xi(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 B and U have a slightly different role compared to A and W since they are built out of non-dynamical variables but still introduced as new fields in the path integral.

Authors' response: Indeed the referee is correct that $x(t)$ is an external variable, as stated in footnote 10. This implies that the product in question simply moves outside the expectation value; we have added a new footnote 18 to avoid any confusion on this point.

Regarding the second comment, while there is a sense in which B and U are slightly different from A and W insofar as the former deal with external data, we do not believe this distinction is particularly meaningful as far as our analysis is concerned; for example, W and U play formally the same role in (4.3).

22. sec. 2.3, p. 19-10, eq. (2.44) to (2.46): The paragraph above (2.44) is not very clear. Moreover, it would be helpful to explain in more detail how one arrives at the Gaussian measure (2.46) starting from the non-Gaussian measure (2.35).

Authors' response: We have removed the potentially confusing remark "pursuant to our self-averaging assumption" to make clear that (2.44) is just the statement that the values are drawn from a bivariate Gaussian distribution (we assume that by (2.35), the referee means (2.44)). In any case, (2.46) is merely a rewriting of (2.44) in the standard form for a bivariate normal distribution. We have elaborated on the line between (2.45) and (2.46) to make this clear.

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?

Authors' response: We have added text at the beginning of section 3 to provide more intuition for the double-copy method, namely, we fix $h(0)$ to be the same for both copies, which have initially identical weights and biases as the latter have not yet been integrated out. We have also referred the reader to previous work by the Google Brain team and others, where a similar idea is used.

24. sec. 3.1, p. 22, footnote 22: 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.

Authors' response: We have implemented this suggestion.

25. sec. 3.2, p. 27, 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 $\tau = t?s$.

Authors' response: We thank the referee for catching this confusing typo. $d(\tau)$ should be $d(t_1, d_2)$, and we have changed $d(t)$ (previously used as a shorthand for the case $t_1 = t_2$) to $d(t, t)$ for clarity.

26. sec. 3.2, p. 27, 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.

Authors' response: Indeed, we did not mean to overload notation here; we have changed this lightcone coordinate to u.

27. sec. 3.2, p. 28, 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).

Authors' response: We are of course not saying that all solutions to the time-independent Schrödinger equation are bound states, but the solution that we are looking for is; we have added text below (3.31) to avoid this confusion. We have also rewritten the text between (3.37) and (3.39) to make the logic more clear. Additionally, while we have streamlined our presentation to focus only on the necessary elements of the argument, we have now referred the reader to the quite detailed discussion in [36] on this point should they desire more information about the potential.

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).

Authors' response: See previous point. Precisely at the edge of stability, when (3.38) is saturated, $y(\tau)$ is the ground state with $E = 0$. We apologize for the previously confusing wording.

29. sec. 4, p. 30, par. 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.

Authors' response: This is indeed what the statement says. We have replaced the word "correction" with "interaction", in the hopes that this is more clear.

30. sec. 4.1, p. 40, below eq. (4.48), par. 2:

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

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

Authors' response: No: at the beginning of that same paragraph, we explicitly say "weak 't Hooft coupling $\sigma^2_{w,\text{eff}} < \gamma^2$". We have consistently referred to σ^2_w (or, equivalently, the effective variance, cf. (4.48)) as the ('t Hooft) coupling, and to T/N as the expansion parameter. This has been previously stated in the abstract, introduction, and the very beginning of section 4. Nonetheless, it has also been further clarified in the new appendix A.

31. sec. 4.1, p. 37, below eq. (4.48), par. 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?

Authors' comments: As we have said, the expression is both very lengthy and not at all enlightening. If the expression were reasonably short and simple, we might indeed have included it even though it is certainly not a central result, but this is not the case. In contrast, the reason we have included the similarly lengthy expressions on pages 72-73 is because these are actually part of the main result. We do not believe that writing-out these expressions would improve the paper, which after all is already 8 pages longer than the (already long) original version.

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).

Authors' response: It is true that the strands in 't Hooft's notation stand for indices, and hence an n-tensor field has n strands. In our case however, the ribbons are there because the W and $\tilde{W}$ fields are bi-local (see point 19 above). We do not have matrix fields; the double-lines denote different times, not different indices. Furthermore, we do not have the same relationship between the genus of the surface on which a ribbon diagram can be embedded and the order in perturbation theory exhibited in 't Hooft's notation. We regard the close resemblance as interesting, but it would be misleading to identify them.

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 $c(\omega)^2 \sim \delta(0)\delta(\omega)$.

Authors' response: Agreed, this was poor phrasing; we have corrected the wording, and thank the referee for catching this.

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): $\langle \tilde{z}_i t) \rangle \qquad \langle \tilde{z}_i(t) \rangle$.
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 $f(x) \to f(\omega)$.
4. sec. 4.3.1, p. 42, eq. (4.53): there is an extra comma after $d\tau''$

Authors' response: We thank the referee for spotting these typos, and have corrected them along with a few additional typos documented in the list of changes.