Reply to referees

We would like to thank both reviewers for their positive feedback, constructive criticism and suggestions. We collect in the present document the invited reports we received and our reply. We use two different colors to highlight the **referees**' words and **ours**.

Anonymous Report 2 on 2024-9-20 (Invited Report)

Summary of contributions

This paper focuses on the random features model (RFM), a single-hidden layer neural network where only the second layer weights are learned. Recent studies of this model are numerous, as it exhibits some learning phenomena observed in more complex neural networks, while being amenable to theoretical analysis. Denoting D the dimension of the data, N the size of the hidden layer, and P the number of data points, the paper aims at describing the phenomenology of learning in RFMs in a wide range of polynomial scaling for (D, N, P). Assuming Gaussian i.i.d. input data and a random polynomial teacher for the input-output data, and a RFM student, the authors derive, under a series of approximations, a set of equations for the order parameters of this problem. This allows to compute the generalization error achieved by the student, in a wide range of polynomial scaling regimes $N \sim D^L$, $P \sim D^K$. These results are strengthened by numerical simulations illustrating the predictive power of the theory.

On a technical level, the analysis relies on several approximation steps which can be justified in these scaling regimes. The authors first map the RFM to an equivalent noisy polynomial model with a degree varying as a function of the hidden layer size N (Section 4). This allows to then make use of the replica method of statistical physics to derive the aforementioned set of equations for the order parameters, while keeping track of all terms that might be relevant in the polynomial scaling regimes $N \sim D^L$, $P \sim D^K$ (Section 5). This latter derivation is however not straightforward, as the authors rely in particular on another important approximation regarding a class of large random matrices, for which they provide an analytical justification (Section 5 and 6).

Main comments

I read through the main text in some detail (following the computations and arguments, but not always checking every step), but did not check the content of the appendices in detail. Overall, I found the paper well-written and pleasant to read: the authors took good care of explaining well notations (cf. Table 1), and laying out the computations. Moreover, while I am not an expert on random feature models, the discussion of the existing literature is thorough, covering in particular important references beyond physics, in mathematics and theoretical computer science. It is well emphasized that the main contribution of the paper is to extend the framework of the replica method of statistical mechanics to study the generalization error in random feature models in the generic polynomial scaling regime $P \sim D^K$, $N \sim D^L$, beyond the previously-known proportional regime $P \sim N \sim D$. While the theory involves several approximations, the numerical experiments are convincing with regard to the predictive power of the equations derived by the authors, which seem to capture the behavior of the learning performance in a variety of different scaling regimes.

My global opinion of the paper is therefore positive: the authors manage to derive an effective approximate theory for learning in RFMs which yields good results in much more general regimes than previously known, and illustrate it with numerical experiments. I have no doubts that such conclusions will be of interest to the community. However, I also think that several points (especially technical but important assumptions and approximations used in the derivation) need to be better discussed, see below for some examples. For these reasons, I would tend to recommend the paper for acceptance in SciPost if the concerns I detail below are addressed in a revision.

We thank the referee for the time spent on our draft and for the positive feedback. We address their main concerns below.

1. I was slightly confused by the role of the readout activation ϕ. Indeed, the loss function $(y - \lambda)^2$ of eq. (10) fits the data y to the pre-activation λ, while the generalization error measures the error as

$[y-\phi(\lambda)]^2$, so the population loss (that the student aims at minimizing through its empirical version) is not related to the generalization error. In a classification task, is there a motivation for considering this loss?

Our main motivation for considering the purely quadratic loss (which has been used in the past to train classifiers, as, for example, implemented by the class RidgeClassifier in scikit-learn) is that for this choice the pattern contribution to the free energy S_P can be evaluated analytically at any β (see eq G.5, which is valid before the large β limit and for any loss; equation (54) (current draft) follows by evaluating the integral in λ by Gaussian integration). There is no fundamental reason to restrict the theory to this loss, at the cost of evaluating the integral in λ in Eq. G.5 via saddle point for large β. The λ at $\beta \to \infty$ will be then given by the extremum over λ of the function at the exponent of the integrand in eq. G.5 (this is essentially the proximal operator of the loss). We have explicitly stated this possibility in the revised manuscript.

2. Around line 200, it would be useful for the convenience of the reader to detail a bit more (possibly in appendix) the CLT-type arguments behind the Gaussian approximation for $p(\nu, \lambda)$, besides the references given in line 209.

   We stated more clearly that we used the Gaussianity of $p(\nu, \lambda)$ as an ansatz to be checked a posteriori. Given the sophistication of the mathematical literature on the subject, we feel that even a short review of it is beyond the scope of our paper, and we limit ourselves to cite the papers we are aware of. We added to this list current Ref. [36], which provides a sketch of a proof in Appendix A.2, where the moments of the variables λ are evaluated and the leading order diagrams identified as the Gaussian ones.

3. In Section 4, I was not able to directly understand why the effective truncation of eq. (24)-(25) suggests that one can represent the RFM as an effective noisy polynomial student. Is this because if one were to consider such a polynomial student, the kernel of eq. (20) would then have the form given in (24)-(25)? As this is at the heart of the results of this section (and of the paper, since this is directly used in the replica calculation later), I believe the authors should add clarifications to this argument.

   Yes, as mentioned already by the referee, the reason to consider a polynomial student of that form is precisely to match the statistics of the truncated kernel. We clarified this point in the text.

4. The conclusion of Section 4 is an extension of the Gaussian equivalence principle of [GLR + 22] to the case $N \sim D^L$ for some $L > 0$. However the authors also mention that they take in this section the limit $P \to \infty$ "for the purpose of arguing" and later on use its results for more general values of P (see eq. (34), where it is used in the replica computation). Could the authors discuss more why this principle remains valid in the replica computation even if the limit $P \to \infty$ is not taken before $N, D \to \infty$?

   Section 4 is written in terms of the solutions of the optimization problem (9) *at a given instance of the training set*, and so we use the large-P limit in order to introduce the true kernel $\mathcal{K}$ instead of the empirical one $\bar{\mathcal{K}}$ via the law of large numbers (which is indeed not justified as $P \sim N$). In the replica calculation, however, the expectation over the data distribution is taken explicitly (see (32), current draft), so that the true kernel appears naturally once Gaussianity of $(\lambda^a)_a$ at given $\mathbf{w}$, F, θ is assumed, independently on the values of N, P, that only need to be large in order to evaluate the order parameters at the saddle point.

5. In (43) the authors use the replica symmetric ansatz. Is this justified here simply by the convexity of the problem?

   Yes, the RS ansatz is justified as long as the optimization problem we are considering (Eq. 9) is convex. As the parameters $\mathbf{w}$ enter linearly in the function λ, the only requirement is the convexity of the loss function $\mathcal{L}$ with respect to λ.

6. The assumption that the row spaces of $C^{\odot \ell}$ and $C^{\odot k}$ (for $\ell \neq k$) are almost orthogonal is crucially used in the derivation, however I don't believe it is justified in detail in the text. Is it related to the later approximation of $C^{\odot \ell}$ by removing terms with equal indices (eq. (47)), and then taking them to be Wishart matrices?

The two assumptions (orthogonal row spaces for $k \neq \ell$ and Marchenko-Pastur distributed spectra) are distinct but follow the same argument. By removing terms with equal indices and taking the spectrum of matrices $C^{\odot \ell}$ distributed with a Marchenko-Pastur law, we are saying that these matrices can be written as sums of outer products of almost orthogonal vectors (the vectorized tensors $\mathbf{F}_{\boldsymbol{\alpha}}^{\otimes \ell}$). In the same way, by taking inner products of tensors with different degrees, we obtain suppressed contributions. We updated the manuscript to clarify these points, following also suggestions from Referee #1 (see updated Sec. 6)

7. In the saddle point equations of Section 5.2, the authors keep track of many quantities that depend on (D, N, P) without taking them to their asymptotic limit (said asymptotic limits being studied in Appendix H). This allows to tackle different scaling regimes with a single set of equations, and to obtain a much better agreement with experiments than what is given by the asymptotic limits. Could this surprisingly good agreement be analytically justified by analyzing the magnitude of the finite-size corrections to eq. (50)?

   The agreement is indeed surprising, considering that we kept all the leading order terms in N, P and D^ℓ for all ℓs, discarding corrections that can be of the same order of some of these terms. Formally, the parameters that have to be large in order to justify the evaluation of the order parameters at the saddle point are N and P, entering the variational free energy. The scaling of D is used to identify where to truncate the kernel. Currently we do not know how to justify the agreement in this setting.

8. Section 6 seems a bit repetitive with respect to the discussion around eq. (47). Since the authors mainly focus in Section 6 on the Wishart approximation for $C^{\odot \ell}$, it might be clearer to re-organize this discussion around where this approximation is used.

   We feel that the few lines around eq (47) ((49) current draft) require more discussion, as also apparent from the round of reviews. At the same time, we prefer not to make heavier the already technical derivation in section 5.1, postponing the discussion to Section 6. We rephrased some sentences to clarify our aim and the scope of Section 6, without major re-organizations.

9. In section 6, the authors mention two "cornerstones" of their analysis. However, it seems to me that other important assumptions are used in the derivation, such as the column space orthogonality assumption, or the extension of the Gaussian equivalence principle. Is there a reason why the authors chose to focus on these two assumptions here?

   Indeed, we updated the beginning of section 6 stating all the assumptions taken so far, and clarifying that the scope of the section is to justify the ones regarding the ensemble of random matrices. The others have been considered more carefully throughout the text, following the referees' suggestions.

10. In the conclusion, the authors mention considering trained neural networks as an open direction. Perhaps some other open directions are more easily reachable: do the authors believe that these methods could be extended to more correlated models of data, or non-convex losses for instance?

    We appreciate the suggestion. We think indeed that our derivation can be extended to consider structured data (Gaussian mixtures, hidden manifold models, object manifolds, etc). In these cases, it can be relevant to check how the intrinsic dimension D_0 of the data plays with the other dimensions entering the problem, that is D, N and P.

    As for non-convex losses: it is true that the RS ansatz can be justified a priori only as long as the loss is strictly convex, but (i) it provides an approximation even for RSB problems and (ii) the teacher is planting in the training set a low-energy configuration that in many other known cases is enough to effectively convexify the problem (see for example the case of the linear classifier with hinge loss, which generically presents a RSB phase diagram in the case of random labels [Franz, Sclocchi, Urbani] but is in practice RS when the labels are assigned by a teacher [Loffredo, Pastore, Cocco, Monasson]). For these reasons, we believe that our analysis can be relevant even for non-convex losses.

    We added these ideas to the discussion section, together with some references already exploring this direction.

Minor comments and questions

I list here some minor comments and questions.

1. The ELU activation is mentioned in Figure 1 before being defined.

   We reported the definition in Fig. 1 to improve readability.

2. To clarify the setting, it might help to mention in the introduction (where the main conclusions are mentioned) that this work considers a student which learns by minimizing a square loss, which differentiates this work e.g. from Bayesian students, or other choices of loss functions.

   As mentioned before, there is no fundamental reason to restrict our derivation to the square loss, if not to make formulas more explicit. With other choices of losses, the large-β limit can still be obtained from a saddle point integral. We updated the manuscript to clarify that we are considering a classical empirical risk minimization problem, as opposed to the Bayesian setting.

3. In Section 5, the authors compute the asymptotics of the log-partition function as a way to obtain the asymptotic values of the different order parameters defined in eq. (14), from where they finally get the generalization error, since they approximate $p(\nu, \lambda)$ by a Gaussian whose moments are given by said order parameters. I was thus a bit confused by the presentation of eqs. (30) and (31) which makes it seem that the authors directly compute $p(\nu, \lambda)$ (even introducing its replicated version), while I don't believe this is used anywhere after that.

   The presentation of (30) and (31) (current (32), (33)) is written to connect with the derivation of the explicit formulas of the generalization error (15) and (16), coming from (13). We intended this as a motivation for replicas, as some readers might not find obvious why the only thing needed is the log-partition function.

4. In line 295, the authors say that the Fourier conjugate of t a goes to 0 in the large-N limit, giving reference to Gardner's seminal work [Gar88]. It would be useful to add a short paragraph (possibly in appendix) to explain why this is the case.

   The reason is that the terms depending on $\hat{m}^{(0)}$ in the variational free energy can be written as $\sqrt{N}\hat{m}^{(0)}m^{(0)} + N(\hat{m}^{(0)})^2 K$, for some $O(1)$ function K not depending on $\hat{m}^{(0)}$, so that the saddle point in $\hat{m}^{(0)}$ gives $\hat{m}^{(0)} \sim 1/\sqrt{N}$, which is small in the asymptotic limit we are considering. We added this argument in a footnote.

Some typos –

1. Line 41: "a the lazy-training"

2. Lines 172-173 do not read well.

3. $\mathbf{w}^\star_{\mathcal{T}}$ instead of $\mathbf{w}^\star$ in eq. (9), to match future notations (here and in several places).

   As $\mathbf{w}^\star$ is an implicit function on $\mathcal{T}$, θ, and F, we omit altogether the label $\mathcal{T}$ and stated this fact at the first occurrence, after Eq. 9

4. Line 181: the sentence starts right after the equation?

5. Line 187: "the computation partition function"

6. Line 205: "this quantities"

7. Line 211: "the Gardner's"

8. Line 226: "th"

9. Line 293 and 326: "indexes"

10. Line 326: "make make"

11. Line 349: A sentence is not finished.

    We thank the referee for pointing out these typos.

0.1 References

[Gar88] Elizabeth Gardner. The space of interactions in neural network models. Journal of physics A: Mathematical and general, 21(1):257, 1988.

[GLR + 22] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In Mathematical and Scientific Machine Learning, pages 426–471. PMLR, 2022.