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 1 on 2024-7-22 (Invited Report)

Strengths

- The authors provide a tight asymptotic characterization of the learning of Random Features Models (RFM) on a random polynomial target function, in various data/width/dimension regimes.
- They outline and identify data (resp. width) limited regimes where the RFM reduces to a kernel (resp. polynomial regression) method, as well as a non-trivial width data regime, exhibiting in particular an interpolation peak phenomenon.
- The derivation relies on the replica method from statistical physics and several random matrix theory arguments and approximations. All steps are rather clearly justified, motivated, and discussed.
- The analytical findings are supported by convincing numerics.
- The consequences/takeaways of the analytical results are discussed, notably in terms of overfitting and expressive power.

We thank the referee for the careful reading of the manuscript and for the fair assessment of its strength points.

Weaknesses

I am listing a few minor points, typos or questions in the "changes requested" section. Here, I am listing some of my main questions and concerns.

- l. 154 (definition of the teacher). The authors consider a random teacher function, which doesn't allow to investigate simple and natural target functions such as $\|\mathbf{x}\|^2$, Hermite polynomials or spherical harmonics. This prevents in-depth connection and comparison to related results on kernel learning, e.g. [22]. How important is averaging over the teacher in the derivation? [57] (which I am aware is contemporaneous to the reviewed paper, and has a arXiv released after the first arXiv version of the present work) for example seem to be able to accomodate deterministic targets, at least for the learnable polynomial space.

  We agree that the choice of the teacher could be made more general. However, at least in the asymptotic regime we are considering, some of the examples the referee is suggesting are equivalent to our choice: a term proportional to $\|\mathbf{x}\|^2$ is converging to a constant for D large (this term is a constant for all D in the setting of [22], which is considering input data on the unit sphere, while in our case of i.i.d. standard normal vectors this is only true by law of large numbers). As for deterministic teachers: the average over the teacher is performed starting from equation (37) (Eq. (39) in the update draft), where the coefficients θ are still appearing explicitly. Without this average, the Gaussian integral over the weights $\mathbf{w}$ in this equation can still be performed, leaving a function of the student's features F, the teacher coefficients θ and the order parameters (and their conjugates). In this respect, the average of the teacher is used to simplify this expression, making only simple combinations of the random features appear through the matrices $C^{\odot\ell}$, that in the asymptotic limit we are considering depend only on the Stieltjes transformations (49) (Eq. (51) in the update draft). While not fundamental in principle, the use of a random teacher is thus crucial in writing in a relatively simple and amenable form the system of equations for the order parameters in our approach, which is then giving the explicit generalization curves reported in the paper. The mapping of our approach with [57] is left for future work.

- In the otherwise rather complete related works section, l. 122, maybe the works of [Zavatone-Veth and Pehlevan, 2024] and [Schroder et al, 2024] on deep structured RFMs, and that of [Defilippis, Loureiro, Misiakiewicz 2024] on dimension-free characterizations of RFMs could be included. I am aware some of these works are contemporaneous or appeared after the release of the first arXiv version of the present work, though prior to the present submission/version, and would leave the decision to the authors and the editor.

  We thank the referee for pointing out these relevant references. We added [Defilippis, Loureiro, Misiakiewicz 2024] in the related works section of the revised manuscript, and [Zavatone-Veth and Pehlevan, 2024], [Schroder et al, 2024] in conclusion, where we mention the possibility of extending our work to study structured data.

- p. 10: I understood the discussion, but it should ideally be clarified. In particular, could the equivalent model be written in terms of Hermite polynomials instead of Wick products, to connect with related equivalent maps e.g. [22]? Also, adding a short appendix explicitly showing how the features (26) admit population covariance (24), if this is the case, would be helpful.

  We improved the exposition on this part, writing the equivalent polynomial model also explicitly in terms of the Hermite basis (current Eq. (30)). In particular, Appendix C is answering the last point raised by the referee, and has been now properly referenced in the main text.

- Some technical approximations (l. 307, l. 243-246, further elaborated in the questions in "requested changes") are merely stated without sufficient discussion. I have not fully understood how these statements are supported, or if they are heuristic assumptions, and feel like further discussion is needed in these passages.

  We clarify point by point these statements below.

Report

I am overall in favor of acceptance. While I have not carefully gone through every reported technical steps, the overall derivation seems scientifically sound. The question explored is of interest, and my concerns are primarily on some aspects of the exposition of the results, although the overall quality of the writing is largely good and clear.

Requested changes

I am listing below a number of typos, comments, and minor questions.

- l. 115 "as long as with finite dimensional outputs"
- l. 187 missing "of"
- l. 201 Slightly awkward phrasing, maybe "since x is a test pt, and is thus uncorrelated with ...." would be simpler and clearer.
- l. 226 missing "e"

  We corrected the above typos. We thank again the referee for his time in reading carefully the manuscript.

- l.243-246 Is there any (even non-rigorous) reason to expect the rank to be given by this minimum? Isn't it in full generality just an upper bound? More discussion would be helpful.

  We corrected a typo in the manuscript on line 244 (current l. 264), as the rank of the matrix $C^{\odot \ell}$ is generically equal to $\min\{D^\ell/\ell!, N\}$, as reported elsewhere in the text. The reason to expect this is that, neglecting possible outliers, these matrices can be built by summing the outer product of $D^\ell/\ell!$ almost orthogonal vectors, that is $\{\mathbf{F}_{\boldsymbol{\alpha}}^{\otimes \ell}\}_{\boldsymbol{\alpha}}$ (see current Eq. (60)). By "almost orthogonal" we mean that $\mathbf{F}_{\boldsymbol{\alpha}}^{\otimes \ell} \cdot \mathbf{F}_{\boldsymbol{\beta}}^{\otimes \ell} = o(N)$ as $N \to \infty$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$. Indeed, at least 1 index α must be different from 1 index β in order for the 2 multi-indices to be different, and their inner product is suppressed in this

asymptotic limit by CLT (the variables $F_{i\alpha}$ have zero mean). By "generically", we mean for typical instances of these matrices. We added this discussion to Section 6.

- Similarly, the statement that off-diagonal elements don't affect eigenvalues/vectors is a bit too fast, and further discussion would be helpful to support this.

  The statement can be made more clear comparing with the Wishart ensemble: a Wishart matrix has N elements on the diagonal that are O(1) and $\sim N^2$ elements outside the diagonal that are $O(1/\sqrt{D})$. These elements cannot be neglected in computing the spectrum as long as $N \sim D$, as apparent from the Marchenko-Pastur law. However, when $N/D \to 0$, the Marchenko-Pastur law concentrates around 1, as the elements outside the diagonal are too small (compared to their number) to impact the distribution of eigenvalues given by the diagonal alone. In other words, when D is scaling faster to infinity than N, by law of large numbers a Wishart matrix concentrates around the identity matrix. We added this argument to the draft.

- l. 343 incomplete sentence.

- l. 417 Instead of "overfitting the effective noise", isn't the model rather using the effective noise to overfit the teacher? The two phenomena are different.

  We think this is a matter of wording. Indeed, this is what we meant. We updated the manuscript following the suggestion of the referee.

- (55) Though I understand the approximation of neglecting diagonal terms, I am not sure to understand why the ℓ-th Hadamard power of C can be thought of as Wishart? In particular, it seems the corresponding matrix involved in the product doesn't have Gaussian entries, and the entries are also not mutually independent? Perhaps more discussion would help.

  In few points of the manuscript we were using the term "Wishart" improperly. The correct statement is that the ℓ-th Hadamard power of C (once the diagonal terms are neglected) has a spectrum asymptotically distributed according to the Marchenko-Pastur law. This property is true for more general matrices than the ones in the Wishart ensemble. In particular, in our case the lower moments of the matrices in the product (60) (of the current draft) do match the ones of a Gaussian independent ensemble (see also the discussion above on the orthogonality of the vectors $\mathbf{F}_{\boldsymbol{\alpha}}^{\otimes \ell}$). For formal results on this ensemble, see current Ref. [71]. We corrected the improper statements and reported this discussion in the manuscript.

- l. 307 Why are the row spaces assumed orthogonal? Again, more discussion would prove useful.

  A basis of the row spaces of these matrices is approximately given by the vectors $\mathbf{F}_{\boldsymbol{\alpha}}^{\otimes \ell}$ themselves. Inner products among 2 vectors of this kind with different ℓs are suppressed for large N because at least 1 index α will remain unpaired, as we mentioned above. Orthogonality of row spaces is also a reasonable assumption based on what we observed in the numerics, Appendix D.2. It is a useful assumption, as it allows to factorize the contributions of the matrices $C^{\odot \ell}$ giving a clearly interpretable theory as a result. We have modified the text to explicit this more clearly.