Paper ID: 2097. Title: A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent.

We would like to thank the reviewers for their positive support and for their thorough and helpful remarks. The final submission of the paper will reflect their suggested revisions. The typographical errors will be fixed, and more details will be added to clarify the discussions and the proofs. Before addressing the reviewers' concerns individually, we wish to insist that, built upon previous efforts (e.g., of Mei and Montanari), one of the major objectives of this article is to extend the existing double descent analysis to a *more practical setting*. To this end, we proposed to analyze the popular random Fourier feature method, and to work with more generic data models. This allows us to conclude that double descent is *intrinsic* to random feature model and is *independent of the underlying data model*, as long as our mild technical assumptions are met. For the three most confident reviewers we focus on a few clarifying remarks. The fourth least confident reviewer had an anomalously-low score, and we focus on using his/her remarks to help clarify our main results for ML readers more generally.

Reviewer #1: We thank the reviewer for the positive support and constructive feedback.

Reviewer #2: Our result is a natural extension of the analysis of Mei and Montanari (1908.05355) and their results can be retrieved by taking data uniformly distributed on the unit sphere (which is a popular example of concentrated random vectors in our Assumption 2). By specifying the data and target model, Mei and Montanari reached more explicit results and established the double descent test curve. Our results hold for a much broader range of data models, and are thus of more practical interest. The proposed analysis, despite depending on the data kernel matrices, is still fully capable of characterizing the double descent phenomena and matches real-world experiments. More discussions will be made to better distinguish this work from previous efforts.

Reviewer #3: Assumption 2 does not impose any constraints on the number of training or test data in each class and is needed to bound the operator norms of matrices of the type $\mathbf{Q}\mathbf{\Sigma}_{\hat{\mathbf{X}}}^{\mathsf{T}}/\sqrt{n}$ and $\mathbf{Q}\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\mathbf{\Sigma}_{\hat{\mathbf{X}}}/n$. While we have the natural control $\|\mathbf{Q}\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}/\sqrt{n}\|^2 \leq \|\mathbf{Q}\mathbf{\Sigma}_{\mathbf{X}}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{X}}\mathbf{Q}/n\| \leq \lambda^{-1}$, it is in general not true under Assumption 1 if we replace $\mathbf{\Sigma}_{\mathbf{X}}$ by $\mathbf{\Sigma}_{\hat{\mathbf{X}}}$. This is needed in both $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in the proof of Theorem 3 in Appendix D, for instance in the first approximation of $\mathbf{Z}_1$ to bound the difference when we replace $\mathbf{I}_2 + \frac{1}{n}\mathbf{U}_i^{\mathsf{T}}\mathbf{Q}_{-i}\mathbf{U}_i$ by its expectation (with respect to $\mathbf{W}$). More details will be added to the discussions and proofs to clarify more explicitly when and how Assumption 2 is used.

With respect to the divergent behavior of the test error as $\lambda \to 0$, it is indeed due to the two-by-two matrix $\mathbf{\Omega}$ (instead of $\bar{\mathbf{Q}}$) in Theorem 3 that scales like λ^{-1} as $\lambda \to 0$ for $2N = n$. This is briefly discussed in Remark 3 and Section 3.3, with a proof in Lemma 5 of the appendix. We will state these results more explicitly in the final version of the submission.

Reviewer #4: We would like to clarify what appear to be several misunderstanding on the part of the review, and these constructive comments will be incorporated into the final version of this submission to help clarify the following issues.

Significance of this work: the theoretical analysis in the large n, p, N regime (with in particular the number of random features N not *much larger* than the sample size n) proposed in this work is, by itself, of considerable practical significance. While random feature techniques are proposed to alleviate the computational burden of large kernel matrices, and one thus expect to take $N < n$, our analysis shows in this practical $N \sim n \sim p$ regime that there is a significant mismatch between results obtained with the popular random Fourier feature and the "expected" Gaussian kernel. This is numerically supported by Figure 1 and theoretically explained by our Theorem 1-3. We thus argue that the simple substitution of the random Fourier Gram matrix by the Gaussian kernel matrix can be hazardous in most random feature-based methods, in the more practical $n \sim N$ regime.

Simpler characterization of double descent: as a consequence of the under- to over-parametrization phase transition behavior of the resolvent $\bar{\mathbf{Q}}$ discussed in Section 3.2, we observe in Remark 3 that the double descent test curve is a direct consequence of this phase transition and more precisely, of the singular behavior of the two-by-two matrix $\mathbf{\Omega}$ (that scales like λ^{-1} as $\lambda \to 0$ at $2N = n$, with a proof in Lemma 5 of the appendix) in the second term of $\bar{E}_{\text{test}}$ in Theorem 3. We will state these results more explicitly in the final version of the submission.

With respect to the empirical results in Section 3, artificial noise is only added to the training data in Figure 5 of Section 3.3. There, the objective of adding Gaussian noise is to study, in a qualitative manner, the impact of training-and-test data similarity on the double descent test curve. More discussions will be added to better clarify this point.

We agree with the reviewer that, by taking the training loss to be $\frac{1}{n}\|\mathbf{y}-\mathbf{\Sigma}^{\mathsf{T}}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|^2$ (instead of $\|\mathbf{y}-\mathbf{\Sigma}^{\mathsf{T}}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|^2$ proposed by the reviewer), with $\lambda > 0$ and $\mathbf{\Sigma} = \sigma(\mathbf{W}\mathbf{X})$ the feature of data $\mathbf{X}$, we implicitly choose the scaling of the regularization λ. This scaling is chosen here so that the feature Gram matrix and the regularization are set "on even ground", in the sense that with, say, linear activation $\sigma(t) = t$ we have $\boldsymbol{\beta} = \frac{1}{n}\mathbf{W}\mathbf{X}(\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{X} + \lambda\mathbf{I}_n)^{-1}\mathbf{y}$, so that $\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{X}$ and $\lambda\mathbf{I}_n$ are both of operator norm of order $O(1)$, with standard Gaussian $\mathbf{W}$ under Assumption 1.

We thank the reviewers again for their time and help in improving our contribution.