General response:

We wish to thank the reviewers for their valuable feedback. We are pleased that they generally appreciated the novelty of the analysis presented in this paper, which sheds light on many recent observations in the double-descent literature. We also appreciate extensive and thoughtful comments regarding clarity and pointers to relevant literature that was missing. We will implement all of them which we think will greatly improve the quality of the text, we also include more specific discussions on the main points below (**R1** **R2** **R3** **R4**).

We would like to start by emphasizing that the purpose of our paper is to analyze and distinguish two types of overfitting which are both attracting significant interest and are often conflated with one another. From this point of view, our paper is less about the existence of multiple peaks but more about (1) differentiating the sources and properties of the two types of peaks and (2) investigating how they interact with the nonlinearity of the activation function (which implicitly regularizes the linear peak: as suggested by **R2** we will emphasize and add more details on this).

We believe these contributions will help understand more complex cases, especially in light of recent works which also study the presence of multiple peaks. On the one hand, data distributions whose covariances have block structure can give rise to multiple *linear peaks*[1]. On the other hand, random feature regression with the NTK features of a two-layer network displays two *nonlinear* peaks due to the block structure of the covariance of the NTK features[2]. The work by Adlam and Pennington, which appeared after submission time and was mentioned by **R1** , provides an answer to **R2** 's question about the possible existence of several nonlinear peaks.

R2 and **R3** raise valid concerns about how our picture generalizes to more complex data distributions and learning algorithms. Despite its simplicity, the setup we consider is rich enough to capture the two kinds of overfitting as desired. As far as the data distribution is concerned, we considered the unstructured iid case in the main text because it is the historical model to describe the two kinds of overfitting. Appendix C briefly discusses the case of MNIST, but the phenomenology of the linear peak becomes significantly more complex in structured datasets as illustrated by the work of Chen et. al. As for the impact of the learning algorithm, an unclear sentence pointed out by **R4** might be underselling how realistic our teacher-student framework is: by '2-layer networks', we mean 3-layer networks with 2 *hidden* layers. Furthermore, we have verified that adding extra layers has little impact on the results, a point which we will further stress.

Specific responses:

@**R1** **Parameter-wise vs. sample-wise**: We are not sure to understand the first weakness stated by the reviewer. Double descent was initially investigated parameter-wise but several works have subsequently studied it sample-wise. We certainly agree with the reviewer that changing the width and the dataset size is not symmetric. We vividly illustrate this fact by presenting a full visualization of the 2D phase-space, which highlights the role played by the input dimension. **'Aren't both peaks due to small eigenvalues in $Z^T Z$?'**: The linear peak is indeed related to small eigenvalues (orange curve of fig.5), but not due to strictly vanishing eigenvalues like the nonlinear peak (purple curve of fig.5). In fact, the smallest eigenvalue decreases monotonously sample-wise (for $N \leq P$) reaching zero in correspondence of the non-linear peak. An explanation for this counter-intuitive behavior is provided in appendix B. We will clarify this point.

@**R2** **Bias-variance decomposition**: We understand the concerns of the reviewer about the relevance of the section presenting a BV decomposition. The aim of this section was to highlight the fact that the linear peak is only caused by noise variance, which is why it vanishes in absence of noise in contrast to the nonlinear peak. However we acknowledge that the density of the figures may drown the message conveyed, and will consider simplifying them or moving some material to the appendix.

@**R3** **The case of MNIST**: Our contribution indeed leverages pre-existing theory, but we believe it brings along new insights and will do our best to better highlight the key takeaways. Regarding the MNIST experiment, as we will explain better, the absence of the linear peak is due to the very small intrinsic dimension of such a simple and structured dataset. A more thorough investigation of the impact of the structure of the dataset is left for future work.

@**R4** **Role of the nonlinearity**: We agree that the wording 'purely nonlinear' ($r = 0$) should be clarified. In the high-dimensional regime we focus on, the linear part of an activation function is measured by $\zeta = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[z\sigma(z)\right]$, which vanishes for even functions such as $\sigma(x) = |x|$ but also $\sigma(x) = x^2$. The non-smoothness of generalization behavior around the point of pure linearity ($r = 1$) is also an interesting question which we will expand in SM. As for the divergence of the bias in Mei and Montanari's work[3], it is due to the fact that the bias still contains the diverging initialization variance. The latter was disentangled to yield a well-behaved bias[4].

[1] L. Chen, Y. Min, M. Belkin, and A. Karbasi. Multiple descent: Design your own generalization curve.

[2] B. Adlam, J. Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization.

[3] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve.

[4] S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime.