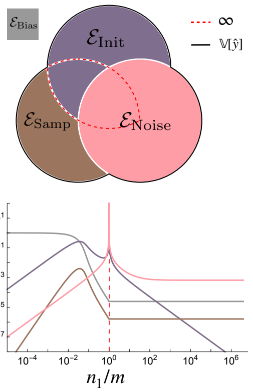
We thank the reviewers for their careful reviews during these trying times. We also appreciate their additional efforts considering this rebuttal, which we believe addresses all of their major concerns.

Prior work. The reviewers highlighted a few relevant related works that we were unaware of at the time of submission. We have added the below figure and a discussion of these papers to an updated version of the paper.



d'Ascoli et al. (2020). It was surprising to see a paper with such a similar problem setup and motivation at ICML (unfortunately, we missed the arXiv version). Our efforts were in fact completely independent and, as R4 points out, use entirely different techniques. These calculations are exceedingly technical — taking well over a year to complete — so please bear this in mind when considering if the two works are contemporaneous. We emphasize that we significantly advance the understanding put forward in d'Ascoli in at least four important ways. 1) We give a unique, symmetric decomposition of the variance that applies to *any* model with multiple sources of randomness. The growing zoo of published decompositions for RF regression can all be understood as special cases of our decomposition (including d'Ascoli; see figure to the right, which renders their decomposition in the setup of Fig. 1 of our paper). 2) Our decomposition resolves implicit ambiguities in the d'Ascoli approach — if they had investigated bagging, they would have found that it actually removes the divergence, which would seem to contradict their main conclusion that it is caused by label noise and parameter initialization. 3) Their analysis was restricted to unstructured RF models, whereas ours extends to the NTK (see Sec. S2). 4) Our techniques from RMT are significantly more generalizable; we believe the replica method cannot be simply extended to handle non-Gaussian or correlated matrices, whereas our free probability approach opens the door for such investigations in future work.

Yang et al. (2020). This paper defines the total bias and variance similarly to Neal *et al.* (2018), but they do not decompose the variance. Several of their observations, *e.g.* that label noise increases the total variance and can lead to double descent, are made precise through the decomposition in our paper.

Jacot et al. (2020). This paper studies the relationship between Gaussian RF models and Kernel Ridge Regression. Unlike our analysis, their bias-variance decomposition is conditional on the dataset X. The variance decomposition itself utilizes the law of total variance, and so can again be viewed as a special case of our fine-grained decomposition.

Simpler models. We believe the simplest tractable model that captures all the phenomena of double descent is the RF model we study here — for linear regression, the number of parameters cannot be varied without simultaneously changing the training data. Our results do apply to the RF model with $\sigma(x) = x$, but it is rank constrained by $\min\{n_0, n_1\}$ and does not properly model the over-parameterized regime. We have added a discussion of these points to the text.

R1. We chose the data distribution to be defined by a linear function for three reasons, which we have now clarified in the text: 1) it is the setup analyzed in the prior work that we compare to; 2) it reproduces the rich phenomena of double descent; and 3) the more general case of a nonlinear NN teacher can be reduced to this case, see Sec. S2.

The missing citation to Jacot *et al.* (2018) was an oversight and has been added to an updated version.

R2. We have not found any straightforward intuition for all of the variance terms in our decomposition. However, we feel that Example 1 is useful. Ultimately, if you ask how to break up a variance into disjoint components that can be summed together in a meaningful way, this is in some sense the "correct" way to do so. We hope our example of the variance reducing effects of ensemble and bagging methods demonstrate this. Finally, we have found that considering the variance of multivariate polynomials in iid Gaussian random variables can be illuminating.

We have expanded our discussion of Neal *et al.* to properly credit its many contributions to double descent and the related body of work. To be clear, that paper is an important contribution and was a strong motivation for our study.

R3. The many random variables involved make the notation of the variance terms tricky. The decomposition can be applied to classical models: D would be defined as in our paper as the training data and P would be any random parameters on which the algorithm depends. For linear regression, for example, P would be the initial weights if it is fit with GD, or $P = \emptyset$ if one instead utilized the closed-form expression for the predictor. Similarly, for decision trees, the definition of P depends on how the tree is constructed and whether any randomization is used.

We do not claim that previous works have any mathematical inaccuracies, merely that the interpretation of their results can lead to an ambiguous or incomplete picture. Our goal is to unify previous decompositions, which is especially important given the growing number of different approaches, which we discuss in the Prior work section.