Thank you all for helpful reviewing. We clarify our viewpoint with respect to some of your questions.

Is the assumption about data spectrum unrealistic? The reviewer 1 pointed out that our analysis is based on "an unrealistic spectrum where the eigenvalues are highly degenerative". However, we think there may be some misunderstanding. Our formulation is not limited to some specific eigenvalue distributions. The macroscopic equations (3) and (5) we derived have no limitation about the number of distinct eigenvalues. Besides, just because there are one or two distinct eigenvalues (as considered in section 4), does not mean that the data distributes within ≤ 2 dimensional subspace. For instance, if there are two distinct eigenvalues, $\lambda_1(>0)$ of multiplicity N_1 and $\lambda_2(>0)$ of multiplicity N_2, the data span N_1+N_2 dimensional subspace. Note that N_1+N_2 can be 2, the case with very degenerated input distribution, and it can be N, the case with non-degenerated input distribution.

Though it is unusual that N_i eigenvalues strictly coincide in real dataset, we think that it is meaningful to simplify the eigenvalue distribution to a simple one that has controllable first and second moments, for the main purpose of understanding the relationship between macroscopic data statistics and learning dynamics in an interpretable way.

Difference from recent works on learning dynamics. As the reviewer 1 pointed out, some recent works including Saxe et al. [2019] analyze the relationship between learning dynamics and data structure (spectrum) with linear networks. However, linear networks do not exhibit the plateau phenomenon which is our main interest, since there is no need to occur specialization of weights in linear nets (see Aubin et al. [2018], for example). To our best knowledge, it is the first time for the plateau phenomenon to be discussed in relation to the statistical nature of the data. We are willing to add a mention about this in our camera-ready manuscript if this work is accepted.

On initial conditions of weights. The importance of initial conditions on successful learning has been shown in several works, including Schoenholz et al. [2016]; they suggests that if the initial condition is close to the edge of chaos, the depth scale of the network will become larger and the network will becomes more trainable even if it consists of dozens of layers. However, we think that this effect is limited in our situation, because what we focus on is shallow networks which consist of only two layers.

Meaning of order parameter Ω. We introduced $\Omega^{(e)}$ as grouping the e-th order parameters of the first layers as $\Omega^{(e)} := (Q^{(e)}, R^{(e)}, T^{(e)})$, rather than as a new order parameter. Likewise, we defined χ in order to group together the order parameters of the second layers. The equation shown above p.6 containing $f^{(e)}$, g and h is for emphasizing the dependency between order parameters (the specific form of $f^{(e)}$, g and h is written in the equations (3) and (5)); the important thing is that the 'speed' of e-th order parameters of first layers depends *only* on 1-st and $(e+1)$-th order parameters of first layers, $\Omega^{(1)}$ and $\Omega^{(e+1)}$ (and order parameters of second layers χ). Together with the equation $\Omega^{(d)} = \cdots$ in p.6, which indicates that d-th order parameters can be represented by $(<d)$-th order parameters, we can obtain the closed system of differential equations of order parameters which consist of $\Omega^{(0)}$, $\Omega^{(1)}$, $\ldots$, $\Omega^{(d-1)}$ and χ.

Contribution and takeaway of our works. Plateau phenomenon has been researched mainly with regards to the structure of neural networks, and the relationship with the data has been overlooked (Saad and Solla [1995], Amari et al. [2018]). However, our work focuses on the statistics of the data learned and its impact on the learning dynamics, including plateau phenomenon. The statistical mechanical method we used is a powerful tool for analyzing learning dynamics of nonlinear neural networks. Our main contribution is to extend this method to generalized cases where the input data has arbitrary covariance and to suggest the framework for understanding the relationship between data statistics and plateau phenomenon. The takeaway is the following: the plateau phenomenon specific to learning dynamics of nonlinear neural network is heavily dependent to the data learned. By deriving the learning dynamics with generalized data statistics, we can develop the theory of plateau phenomenon in more realistic settings.

References

S. Amari, T. Ozeki, R. Karakida, Y. Yoshida, and M. Okada. *Neural computation*, 30(1):1–33, 2018.

B. Aubin, A. Maillard, F. Krzakala, N. Macris, L. Zdeborová, et al. In *NeurIPS*, pages 3223–3234, 2018.

D. Saad and S. A. Solla. *Phys. Rev. E*, 52(4):4225, 1995.

A. M. Saxe, J. L. McClelland, and S. Ganguli. *PNAS*, 116(23):11537–11546, 2019.

S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. *arXiv preprint arXiv:1611.01232*, 2016.