We thank the reviewers for their comments and remarks. We are also grateful for the errata, missed references, as well as the questions posed by the reviewers which help us explain our results with more clarity.

Relation to ML-VAMP: A few reviewers wished that the paper could discuss the significance of the ML-Mat-VAMP method over ML-VAMP from [11, 28, 29]. The ML-VAMP algorithm considers only vector-valued quantities in each layer, while the ML-Mat-VAMP considers matrix-valued unknowns. We show ML-Mat-VAMP can analyze a far broader set of applications including Multi-output GLMs, multi-task learning, not possible with ML-VAMP. Extending the proofs to the matrix case is non-trivial as it requires understanding the interaction between columns in each layer.

Also, in addition to analyzing inference problems, we show that the ML-Mat-VAMP can enable studying learning and generalization error of 2-layer NNs. Remarkably, our analysis can provide exact predictions in this case. Previous AMP methods such as ML-VAMP could only study the generalization error of single-output GLMs [ESAP$^+$20]. The application of ML-Mat-VAMP and generalization error in 2-layer NNs is also non-trivial as it requires an interesting recasting of the learning problem into an inference problem.

Response to Reviewers # 1 and # 4. The reviewers raise excellent questions requesting the contrast between this work and [1]. The problem considered in our Section 5 is the same committee machine model from [1], with an important difference that our State Evolution (SE) analysis holds for a far broader class of data matrices – ones which are Rotationally Invariant (not just uncorrelated Gaussian features) – which include correlated non-Gaussian feature vectors leading to poorly conditioned data matrices. Due to the lack of space in the rebuttal document, we are unable to demonstrate this via plots. We shall include these experiments on learning committee machines under correlated non-Gaussian features into the main paper (This case is not explained by the model in [1]). The code for generating the figures and a Python implementation for the ML-Mat-VAMP will be made available on a public github repository. Moreover, our results also holds for the several other multi-layer models detailed in Section 2.

Response to Reviewer # 2. Thank you for your remarks. The supplementary material has the full details regarding the proof. If accepted, we are allowed to add 1 extra page in the main paper. Per your suggestions, we would add more details about the assumptions and definitions of asymptotic weak limits discussed in Section 4 as well as a summary of the proof. We would also simplify SE for some models of interest, e.g. those mentioned in Section 2. Regarding the assumptions, the asymptotic results hold in the case where the number of rows $\{N_\ell\}_{\ell=1}^L \to \infty$ such that $\lim_{N_0 \to \infty} \frac{N_\ell}{N_0} = \beta_\ell = \mathcal{O}(1)$, but the number of columns satisfies $d = \mathcal{O}(1)$. When applying this model to analyze learning in 2-layer NNs, this is equivalent to the case with input features $p \to \infty$, number of samples $N \to \infty$ such that $\lim_{N,p \to \infty} \frac{p}{N} = \beta = \mathcal{O}(1)$, and the number of hidden units $d = \mathcal{O}(1)$. To our knowledge, this regime of 2-layer NNs has not been analyzed in the recent papers on *double descent* in wide networks [LXS$^+$19].

Response to Reviewer # 3. We thank the reviewer for their comments. It is true that a large body of general purpose solvers are available for the inference tasks considered here (e.g. gradient descent methods for MAP inference and MCMC methods for MMSE inference). However, these methods are notoriously difficult to analyze exactly due to the non-convex nature of the problem and dependencies on various factors such as the step size and initialization. The main benefit of ML-Mat-VAMP is *not* that it out-performs these methods. Instead, the main value is that ML-Mat-VAMP offers rigorous and exact predictions on performance in certain high-dimensional regimes. In addition, we show empirically the fixed points of ML-Mat-VAMP agree with standard methods (e.g. Adam). Hence, the paper provides a tool for predicting the performance of commonly-used methods as well.

Regarding Broader Impact: It was our understanding that this section was meant for papers introducing implementations of empirical models trained on large public datasets such as GPT-2,3. However on being pointed out by the reviewers, we realize that our work also serves a broader purpose of bringing interpretability to Neural Network based models which significantly impacts the NeurIPS as well as the broader scientific community. We shall address our thoughts in this regard if our manuscript is accepted.

References

[ESAP$^+$20] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson K Fletcher. Generalization error of generalized linear models in high dimensions. In *ICML*, 2020.

[LXS$^+$19] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8572–8583, 2019.