We sincerely thank all reviewers for their meaningful and detailed comments that will help improve our work. In the following, we address the raised concerns one by one:

Reviewer 1. We thank Reviewer 1 for his/her constructive feedback. (i) Linear- and kernel models: We consider rank collapse to be a phenomenon that is specific to the compositional structure of deep networks and hence shallow models are naturally robust against it. (ii) Binary classification: we agree that mapping all data points to a line does not impede this per se but we would like to point out that the direction of the rank collapse after initialization is in fact random and *independent* from the data, which makes it very unlikely to coincide with the principal direction in the data. Particularly so in very high dimensional neural networks. Fig. 1 shows that rank collapse remains an issue when reducing CIFAR10 to a binary classifcation task (using class 0 and 1 only). (iii) Networks wider than the input: Good point, we will make sure to clarify this in the paper. If $N < c\sqrt{d}$ for some constant c, then the rank indeed remains at $\Omega(N)$.

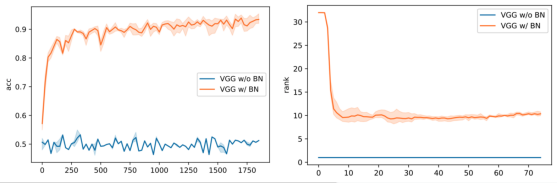

Figure 1: VGG19, two class CIFAR, SGD

Reviewer 2. We thank Reviewer 2 very much for sharing his/her thoughts on our work. (i) Building new DNNs: We would like to emphasize that the main goal of the work at hand was to deepen the understanding of batch normalization and its crucial interplay with random i.i.d. initialization. Furthermore, as we show in Section 4, rank collapse is not necessarily a problem of the architecture per se but it is closely linked to the way the networks are initialized. As can be seen in Figure 4, a sophisticated (rank preserving) initialization can even outperform BN in very deep networks. However, we do agree that developing new architectures with rank collapse in mind is indeed a very interesting direction to follow up on. As we elaborate next, we think our findings are already of interest to the community, as they offer a novel view on BN which we have not seen in the literature.

(ii) Broadness of impact: We believe that the work at hand makes a significant contribution to a better theoretical understanding of the delicate interplay between architecture, initialization and optimization algorithm. Such a theoretically sound perspective can serve as the basis for further practical advancements on all three of these fronts. One such advancement can be found in Section 4 which includes an accelerating pre-training step that is developed directly from our theoretical analysis. Furthermore, since batch normalization itself is one of the major architectural developments of the last decade with successful application in countless settings, we do believe that this result can be an important step towards a systematic improvement of optimization methods for neural networks based on theoretical intuitions.

Reviewer 3. We are very grateful for the comments made by Reviewer 3. (i) γ: Indeed, our analysis needs γ to be small (depending on the width), which comes from the fact that our proof technique relies on low-order Taylor approximations. However, we note that γ is, importantly, independent of L! Furthermore, as can be seen in Fig. 3 this seems to be an artefact of our proof technique, since the result holds empirically even for networks without skip connections ($\gamma = \infty$).

(ii) Overfitting: Well spotted! In our examples (Fig. 11) it seems to us that the overfitting tends to happen whenever a certain accuracy is achieved on the training set, regardless of the actual method that is used. Please see Fig. 2 to confirm this intuition: Here, one can see that when reaching $\sim 75\%$ training accuracy (epoch 100 for pre-training and 200 for pre-training) both methods yield a similar test loss (~ 1.75). This is not necessarily surprising as the networks considered are MLPs that might have a harder time at finding generalizable patterns in image data compared to e.g. convnets.

(iii) Extensions to CNNs: This is definitely a very interesting follow up directions. In fact we are currently looking into extending our theoretical result to convolutional neural nets. The notion of rank is a bit more subtle in convolutional layers as the hidden presentation are third order tensors but we still observe that (after unfolding the tensors) randomly initialized convnets suffer from rank collapse unless batch normalization layers are in place (please see Figure 1 above as well as Fig. 11 in the appendix of our paper).

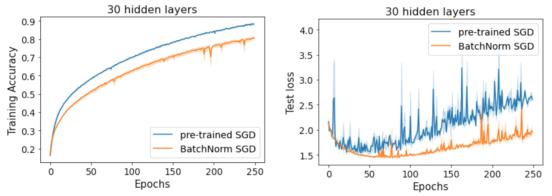

Figure 2: Fig 11 from paper for more epochs