We thank all reviewers for the constructive critique and would be happy to follow their suggestions, in particular, improve the presentation of our theoretical analysis.

Insights beyond mean field analysis. Our analysis provides full distribution information on the joint outputs. All other works study averages (or usually approx. thereof) and mostly focus on parts of the output distribution only (e.g., they either study single inputs or the correlation for two typical inputs). They never provide the full picture and cannot exclude other parameter choices to improve on the initialization with Gaussian weights. In addition, we explain why the squared signal norm is the relevant variable to study - because it is the only information that is transmitted from one layer to the other. While the He initialization preserves this quantity layer-wise *on average*, the distribution becomes more skewed for increasing depths and the center of max. probability decreases (see Fig. 1 (a)). Furthermore, the distribution of the cosine similarity explains why moderately deep and wide ReLU networks can be trained despite negative results by mean field (MF) analysis based on correlations. In addition, we hypothesize that reducing the effective number of nodes in a layer contributes to the success of DropOut and DropConnect. Next, we explain why an initialization with parameter sharing leads to improved signal propagation.

A detailed comment for Reviewer #3: Thm. 2 is not difficult to derive but certainly not standard in MF theory. There, the normal distribution originates from the MF limit. In contrast, here we understand that the output distribution is completely determined by the empirical covariance matrix of inputs. Higher order moments in the input distribution have no influence on the output distribution.

Relation between the theoretical analysis and our initialization proposal. Our theoretical analysis holds for general activation function ϕ. Our specialization to ReLUs identifies several problems that we first try to mitigate by different parameter choices $\sigma_{w,l}$, $\sigma_{b,l}$. This turns out to be impossible. Hence, we solve it with the GSM by effectively setting $\phi(x) = x$ at initialization for half of the nodes, while we disregard the other zero half. Note that the entries in W_0 are iid Gaussian and fulfill our assumptions. We could repeat our analysis for linear ϕ and show that, e.g., input correlations are preserved. This is rather obvious however. Instead, we refer to the rich literature on linear neural networks at initialization. Especially, [SX] suggests the choice of orthogonal W_0 for dynamical isometry. We therefore test this choice in addition.

Additional literature discussion. We agree that we have to extend our literature discussion. Closest to our work is [B] about gradient shattering. Yet, its main focus is on resNets and convNets. We compare with the few results on fully-connected feed forward layers here. They observe a phenomenon that relates to cosine similarity by an argument that links signal forward propagation with gradient descent. However, they study it in a setting where they cannot distinguish between the effect of vanishing gradients and increasing correlations (see Fig. 4 in [B] belonging to $\sigma^2 = 1/N$). As a result, they observe decreasing correlations, while we have a problem with increasing ones. Furthermore, they make a claim about the exponentially decaying covariance (Thm. 1 in [B], now with He $\sigma^2 = 2/N$) without regarding layer width and for typical inputs, while we consider finite layer width and all possible inputs. "Typical" is defined as required for their proving strategy and seems to be common in networks with batch normalization, but not in networks without, which we study. Yet, [B] provides a parameter sharing solution similar to ours, but for convolutional neural networks. It is based on an idea by [S] that was inspired by empirical observations of trained convolutional filters that learn linear networks. Thus, [B] and [S] provide two additional arguments for the proposed parameter sharing solution: batch normalization cannot avoid the problems related to shattered gradients and linear models are good starting points because some layers might need only little adjustment during training.

[HR] and [H] are rigorous studies of the average signal or gradient properties in finite width networks for *single* inputs to avoid vanishing/expoding gradients, while we study the whole joint output distribution, thus also with respect to different inputs. In consequence, they do not encounter the problems associated with the cosine similarity. Indeed, Cor. 1 on p. 8 of [HR] is similar to our Eq. (6). Yet, [HR] assumes the same σ_w and σ_b on all layers and therefore does not ask for alternative parameter choices as we do. [CS] studies with Eq. (1) a similar integral as required for our Thm. 5, but considers the MF limit (in the paragraph after Eq. (8)). [Y] discusses an alternative initialization by shifting the ReLU with a non-zero bias b_i in a MF context with batch normalization. Initially, we thought about following a similar approach, but the main problem is that the bias must depend on the input batch and is thus similar to batch normalization and computationally more intensive.

[SX] Saxe et al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. ICLR'14.

[B] Balduzzi et al. The Shattered Gradients Problem: If resnets are the answer, then what is the question. ICML'17.

[S] Shang et al. Understanding and Improving Convolutional Neural Networks via Concatenated ReLUs. ICML'16.

[Y] Yang et al. Mean field theory of batch normalization. ICLR'19.

[HR] Hanin and Rolnick. How to Start Training: The Effect of Initialization and Architecture. NeurIPS'18.

[H] Hanin. Which Neural Net Architectures Give Rise to Exploding and Vanishing Gradients? NeurIPS'18.

[CS] Cho and Saul. Kernel methods for deep learning. NIPS'09.