

1 Before we address the comments raised by each reviewer in turn, we would like to clarify two key points:

2 **Why normalized ResNets are trainable:** We do not argue that BN can train deep ResNets because it stabilizes the
3 hidden activations on the forward pass. In fact we show in table 1 that stable forward signal propagation is not sufficient
4 (see “divide by $\sqrt{2}$ ” experiment). Our argument is that, when BN is used, the scale of hidden activations at initialization
5 on the skip path grow proportional to the square root of the depth. The growth of the hidden activations on the skip path
6 is beneficial, because the outputs of normalized residual branches have scale $O(1)$, and therefore the residual blocks in
7 deep BN-ResNets are dominated by the skip path. *Residual blocks dominated by the skip path on the forward pass*
8 *will also preserve signal propagation on the backward pass.* This is because the backward propagated signal through
9 the ℓ -th residual branch will be downweighted by $O(\sqrt{\ell})$ relative to the skip path, and therefore the backward signal
10 through the deep normalized residual block is also dominated by the skip path. To provide further evidence for our
11 argument, we have verified empirically that none of the following schemes are able to train a 1000-2 Wide-ResNet:

- 12 1. Placing a single BN layer before the softmax (without including BN layers on residual branches).
- 13 2. Including BN layers on the residual branch but multiplying the residual block by a factor $\alpha \leq \sqrt{1/2}$.
- 14 3. Including BN layers on the residual branch and adding a BN layer after the skip and residual branches merge.

15 In all of these experiments, the residual branch contributes equally to the output of the residual block at initialization (or
16 it dominates if $\alpha \ll \sqrt{1/2}$). The network thus becomes harder to train as depth increases, as predicted by our analysis.

17 **Fixup initialization:** The two main contributions of our paper are to explain why deep BN-ResNets are trainable, and
18 to provide a detailed empirical study of the benefits of BN in popular architectures. This empirical study clarifies that
19 large learning rates are not the primary benefit of BN (as claimed by previous papers). The Fixup paper does not study
20 either of these topics, and we therefore believe our paper has a distinct and significant contribution alongside Fixup.

21 The success of SkipInit also demonstrates that one of the key claims made in the Fixup analysis is false. The authors
22 argue that, in order to train ResNets without BN, one must downscale the weights on the residual branch (even if the
23 hidden activations are already suppressed). Our experiments with SkipInit demonstrate that it is sufficient to downscale
24 the hidden activations. Although SkipInit and Fixup achieve similar performance, SkipInit is simpler to implement.

25 **Reviewer 1:** Our argument is *not* solely based on the forward pass (see above). Placing a BN layer before the softmax
26 achieves 94.1% test accuracy for the 100-2 Wide-ResNet on CIFAR-10, but it cannot train the 1000-2 Wide-ResNet.

27 On line 512 of appendix C, we explain that the approximation $\mathcal{B}(x^0)$ is Gaussian is tight when the batch size $B \gg 1$.
28 Appendix C also assumes large width, and we cite previous work where we build on their results. We do not assume
29 large width in section 2.1. We discuss the work of Balduzzi et al. and Zhang et al. in detail in section 6. See above for
30 further discussion of Zhang et al. (Fixup). We would like to clarify that Balduzzi et al. do not propose that downscaling
31 the residual branch at initialization is sufficient to remove BN. We will expand on these points in the updated text.

32 **Reviewer 2:** We agree that comparing raw learning rates could be misleading, since one can rescale the learning rate by
33 changing the model parameterization. However this does not affect our argument. Our experiments show that, for small
34 batch sizes, the optimal learning rate with or w/out BN (or with SkipInit) is significantly smaller than the largest stable
35 learning rate. In all cases, the optimal learning rate is only close to the largest stable learning rate when the batch size is
36 large. After changing the model parameterization, the optimal learning rate would still be smaller than the largest stable
37 learning rate for small batch sizes. We therefore conclude that increasing the largest stable learning rate (or improving
38 the model conditioning) is not the main benefit of BN for small batches. We will clarify this point in the updated text.

39 We will also include a discussion of the suggested references in the updated text, and we will restructure section 2.1 to
40 improve the clarity of presentation. Introducing biases is beneficial because they ensure the expressivity of the model
41 does not fall when BN is removed. Please see the discussion above which clarifies how our work differs from Fixup.

42 **Reviewer 3:** Please see the summary of our analysis of deep ResNets above. Intuitively, ResNets will preserve signal
43 propagation on both the forward and the backward pass if the residual blocks are dominated by the skip connection.
44 Therefore deep normalized ResNets are trainable since BN downscales the activations on the residual branch. However
45 if we multiply the outputs of residual blocks by $\sqrt{1/2}$, then the residual branch and the skip path will both contribute to
46 the signal on the backward pass. Deep networks of this form will not be trainable with Gaussian weights.

47 We agree that in principle there could be other benefits of BN which also enable it to train deep networks. However
48 we have tried several variants of BN-ResNets (see discussion above), and every variant we have tried which does not
49 downscale the residual branch at initialization has not been able to train when the depth is large. For example, if we
50 multiply the output of normalized residual blocks by $\alpha = \sqrt{1/2}$, our 100-2 BN-ResNet achieves 94.5% test accuracy
51 on CIFAR-10 but the 1000-2 BN-ResNet does not train. The performance degrades further for coefficients $\alpha < \sqrt{1/2}$.

52 **Reviewer 4:** Thank you for the positive feedback on our work. We will clarify that each α is a learnable parameter.