We thank the reviewers for the detailed and insightful reviews. As noted by the reviewers, our work 1) develops generalization bounds for neural nets with improved depth dependency, and 2) "contributes novel tools and techniques for generalization theory." We answer most of the questions and will incorporate the feedbacks into the final version.

[R1,R4] Empirical Comparison of Bounds:

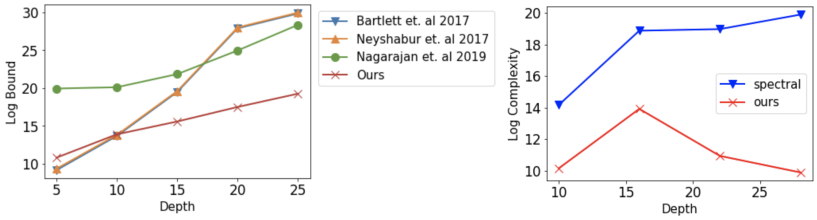

Figure 1: Left: Log generalization bounds (without $1/\sqrt{n}$) for fully connected networks trained on MNIST vs. depth. Right: Log leading terms for spectral vs. our bound on WideResNet trained on CIFAR10 using different depths.

In Figure 1, we address questions about empirical evaluation of our bounds. First, on the left we train a fully-connected neural net on MNIST and measure the our bound and bounds in previous papers.[1] The plot indicates that our bound is generally several magnitudes smaller than existing bounds based on product of matrix norms. Second, on the right we compare leading terms of our bound for WideResNet: $\frac{\sum_i \max_{x\in P_n} \|h_i(x)\|_2 \max_{x\in P_n} \|J_i(x)\|_{op}{}^2}{\gamma}$, where i ranges over the layers, with that of the bound of [Bartlett et al., 2017]: $\prod_i \|W^{(i)}\|_2/\gamma$. We note that our complexity measure scales better with depth for WideResNet than fully-connected nets, which indicates that architecture influences Jacobian and hidden layer norms.

[R1]: "It would be nice if the authors could discuss the challenges to extending it to ReLU networks."

• The primary challenge is that Theorem 5.1 requires the augmented indicators on the Jacobian norms to be themselves Lipschitz w.r.t. the hidden layers. However, the Jacobians of relu networks are piecewise constant and therefore not Lipschitz – thus, to control the change in the Jacobians, we must condition that the pre-activations are bounded away from 0, leading to the dependency on inverse pre-activations in Nagarajan and Kolter [2019] and our Theorem I.1.

[R2]: "It might be better to identify some scenarios in which we can prove that [Jacobians] are indeed small (polynomial in depth) ... the authors can include more intuitions to explain why the interlayer Jacobian should be small in practice."

• It seems to be a very challenging open question to rigorously prove that the Jacobian norms will be small (as this would require characterizing the solution obtained by gradient descent), but many empirical findings support that the interlayer Jacobian on training data should be small in practice compared to the product of spectral norms (e..g, [Arora et al., 2018, Figure 1], Nagarajan and Kolter [2019], Novak et al. [2018]). Intuitively speaking, for linearized deep nets, the Jacobian norm is of the form $\|W^{(k)} \cdots W^{(j)}\|_{\text{op}}$, which could be much smaller than $\|W^{(k)}\|_{\text{op}} \cdots \|W^{(j)}\|_{\text{op}}$ when there is cancellation within the weights.

[R4]: "Shouldn't the upper bound on ν in line 859 turn up somewhere else in the proof?"

• We implicitly used the following fact: suppose $\exists \delta > 0$, such that $\forall x, y$ with $\|x - y\| \le \delta$, $|\tilde{z}(x) - \tilde{z}(y)| \le \tau \|x - y\|$, then, $|\tilde{z}(x) - \tilde{z}(y)| \le \tau \|x - y\|$ is true for any x, y. This can be proven by dividing the line between x and y into segments of length δ and applying the given property on each segment. (We will clarify more in the next revision.)

[R4]: "precisely which part of the proof here makes a similar argument about the Jacobian norms on arbitrary unseen inputs?... I suspect this is implicitly taken care of by Lemma G.1"

• The intuition is correct – the same argument as Lemma G.1 shows that the product of indicators on Jacobian and hidden layer norms is globally Lipschitz, and therefore the the product of indicators generalizes to unseen inputs.

[R4]: "(a) have a theoretical discussion on the worst/best/expected dependence of the terms in this bound on width & depth and (b) empirical observations of the same."

• Our bounds have explicit polynomial dependence on depth, but no explicit dependence on width – they instead depend on the (2, 1) and (1, 1) weight matrix norms. Theoretically, this is comparable to previous bounds [Bartlett et al., 2017, Neyshabur et al., 2017], which also depend on such norms. Empirically, for deep networks it's unclear how these bounds depend on width in practice. (Neyshabur et al. [2018] study this but for shallow depth 2 networks.)

[R4]: "confused by ... training set P_n in Thm D.2. My understanding ... there's no notion of a training set as such..."

• Your understanding is exactly correct - σ, t are arbitrary parameters in Theorem D.2 and only set to their values computed on P_n when we apply the union bound of line 697/698. We will clarify this point in future revisions.

[1] Experimental details: We use BatchNorm layers (to allow training with larger depth) with a hidden layer width of 40 on 1000 randomly sampled images from MNIST. At test time, BatchNorm is an affine transformation,so we merge it into the adjacent linear layer and then compute the bound. We restrict the training set for faster computation of bounds and easier optimization.

[2] We compute leading terms only for faster computation. h_i denotes hidden layer i. J_i is the output Jacobian w.r.t. to layer i. Our bound as stated in the paper technically does not apply to ResNet because the skip connections complicate the Lipschitz augmentation step. This can be remedied with a slight modification to our augmentation step, which we omitted for simplicity.