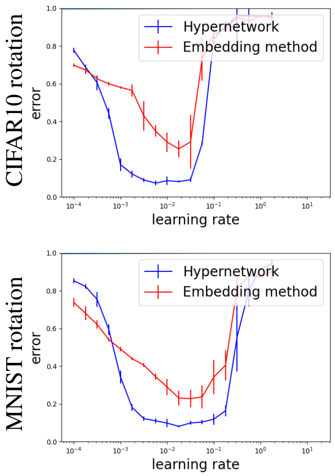

Figure I: Varying the lr.

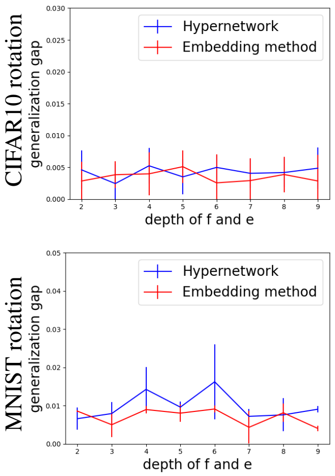

Figure II: Generalization gap.

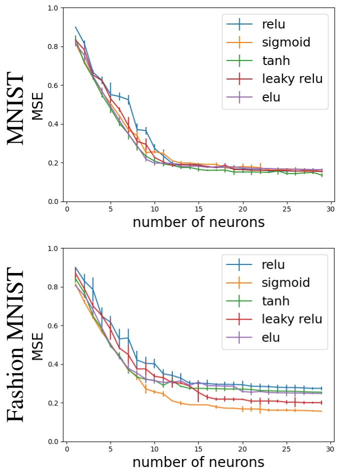

Figure III: Validating assmp 2.

To R1: **Popularizing the approach** We will add a summary of the results in more accessible language, focusing on the more surprising aspects of hypernetworks and how our framework sheds light on these. **A few undefined symbols** We apologize for that. This will be fixed in the next version of the paper. **Abstract** Thanks for the suggestion. In the next version of the paper, we will broaden the context of the abstract similar to the conclusions section.

To R2: **Clarity** We will make an effort to make the reading pleasant and would like to respectfully point out that the other three reviewers believe that the paper is well-written and easy to read. **Prior work** We tried to refer [5] in detail throughout the paper. For instance, in the related work section we explain the tools developed in [5] and how we extend them in our paper (Thm. 1). Additionally, we devoted most of Sec. 3 (L 162-182) to discuss the concepts and results of [5]. In [5] they prove the lower bound in Thm. 1 when assuming the existence of a continuous selector (see L 5-9, L 84-87). In our Thm. 1, we prove the lower bound for the case of neural networks, while proving the existence of a continuous selector. Since the existence of a continuous selector is also a cornerstone in the proofs of Thms. 2-5, the analysis in [5] is insufficient to prove these results. We will further discuss these issues in the next version of the paper. **Unsupported statements in the introduction** The word “suggests” is indeed misleading since the statement on minimal complexity is our definition of modularity, see also L 228-237. We will provide references to support the bottleneck statement, e.g., [21]. **The presentation of [Ha et al. 2016]** was meant to introduce the term “primary network”, while mentioning the usage for sequences. We apologize for being inaccurate, we will fix it in the next version of the paper. The definitions we use are given explicitly in L 111-116. **Clarifying defs and symbols** In the next version of the paper, we will clarify the definitions of accuracy and N-width. Unfortunately, we omitted the definition of BV, which is given in L 100-103 in the supplementary. $C^1(\mathbb{R})$ stands for the set of continuously differentiable functions over $\mathbb{R}$. **Hyperparams** We did not apply any sort of regularization/normalization (including dropout) on the two models to minimize the number of hyperparameters and since the baseline method is considered more stable than the hypernetwork, which implies that it may require less regularization. Following the review, we conducted a hyperparameter sensitivity test for the learning rate. We compared the two models in the configuration of Fig. 3(a-b) when fixing the depths of f and e to be 4. The results, shown in Fig. I clearly show that the hypernet outperforms the baseline for every learning rate in which the networks provide non-trivial error rates. **Missing Fig. 4** See comments to R4.

To R3: **Multidim case** The theory indeed scales to the multi-dimensional case. In this case, if the output dimension is independent of ϵ, we get the exact same bounds. Our colorization experiment in the supplementary corresponds to a multidimensional target function (with three outputs). **The generalization gaps** for the two models are similar, when varying the depth of f and e, same setting as Fig. 3(a-b), see Fig. II. We would explore this further. **Assumption 2** To empirically justify this assumption we trained shallow neural networks on MNIST and Fashion MNIST classification with a varying number of hidden neurons, using the MSE loss and one-hot encoding of the labels. As can be seen in Fig. III, the MSE loss strictly decreases when increasing the number of hidden neurons. This is true for a variety of activation functions. To theoretically justify this assumption, we will prove the following lemma:

Lemma 1 *Assumption 2 holds for $\mathbb{Y} = C([0,1])$ (i.e., the set of continuous functions $y : [0,1] \to \mathbb{R}$) and 2-layered ReLU networks approximators.*

To R4: **Fig. 4** is in the supplementary; the paper’s reference to it should have been to Fig. 1. We apologize for this.