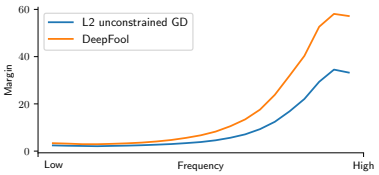
We would like to thank the reviewers for their valuable feedback, which we will duly consider and integrate in our revised manuscript. In this paper, we demonstrate that "the decision boundaries of a DNN can only exist as long as the classifier is trained with some features that hold them together", i.e., DNNs have an inductive bias towards invariance. Through "a multitude of insightful experiments", "with good attention to detail", we delve into this property which sheds light on open problems like "catastrophic forgetting" and "adv. robustness". The structure of our paper (R3) is designed to (i) "rigorously confirm" the existence of this inductive bias (Sec. 3), and (ii) further investigate its consequences on the sensitivity and dynamics of adv. training (Sec. 4). Throughout the paper, we back up all our claims, first, using controlled synthetic experiments, and then, "rigorously" verifying our hypotheses on real datasets with abundant empirical evidence. We clarify the main points raised by the reviewers here below.

Margin and features (R4) The main claim of our paper is that DNNs *only* create decision boundaries in regions where they identify discr. features in the training data. We further shed more light on the relationship between adv. examples and features studied in [3,4]. We show that there is a big relative difference in the large margin along the invariant dirs. and the smaller margin in the discr. dirs. Nevertheless, we never claim that, within the discr. dirs, margin is at all proportional to "discriminativeness". In fact, we agree that the margin associated to different discr. features can greatly vary (Fig. 4). Overall, however, we firmly believe that the invariant dirs. will always have the largest margin.

Causation (R4) The main difficulty for establishing causation in our paper is the fact that the discr. features of real datasets are not known. Hence, determining their role on the geometry of a trained DNN can only be done by artificially manipulating the data. We strongly believe that the experiments in Sec. 3.2 are enough to rule out the other two main factors that might explain our results: the network and the algorithm. Specifically, in the flipping experiments, flipping the data – *ceteris paribus* – also flips the margin distribution, thus demonstrating that the margins are necessarily caused by the information present in the data. The other interventions we do on the data (e.g., low-pass experiments) confirm that in the absence of information in a certain dir. the network becomes invariant along this dir. Therefore, guided by the principles of the scientific method, and supported by strong evidence, we believe that grounds for causation are properly established.

Choice of DCT (R1,R3) The DCT has a long application tradition in image processing due to its good approximation of the decorrelating transform (KLT). Furthermore, in previous studies on the robustness of deep networks to different freqs., the DCT was also the basis of choice [7] because it avoids dealing with complex subspaces. A more aligned basis with respect to the discr. features would probably show a sharper transition between low and high margins. However, finding such network-agnostic bases is a challenging task without knowing the features *a priori*. The DCT is not perfectly feature-aligned, but it seems to be a good choice for comparing different architectures, especially if we compare its results to those obtained using a random orthonormal basis where differences in margin cannot be identified (c.f. Sec. N in Supp. material). We will include this explanation in the revised version of our manuscript.

DeepFool (R3,R4) In the adv. robustness literature DeepFool is generally regarded as one of the most efficient methods to identify minimal adv. perts. Because we measure margin, norm-constrained attacks like PGD are not suitable for our study, and more complex attacks like C&W, or using unconstrained GD in the input space, are computationally much more demanding and harder to tune while finding very similar adv. perts. to DeepFool. Hence, DeepFool is more adequate for our work. We will include this explanation in the revised version of our manuscript. For completeness, we show a comparison of the median margin obtained on MNIST (LeNet) using DeepFool and a subspace-constrained GD attack: Even if the margins are slightly smaller for the *stronger* attack, the relative differences between regions (our quantity of interest) are the same for both attacks.



CIFAR10 margin (R2) Without knowing the mechanisms used by the network to select the discr. features of a dataset it is hard to give a full explanation of the low margins in the high frequencies of CIFAR10. Nevertheless, a possible reason for the different behaviour on CIFAR10 might be its low resolution. In fact, it is a common mistake that during the downsampling process no antialising filter is applied to the images before resizing, and hence some low-freq. information leaks to the high freqs. of the low-resolution images. This might explain why the network can identify some discr. features in the high-freq. spectrum of CIFAR10 (downsampling technique not specified in technical report [25]).

Filtering other datasets (R2) Indeed, the conclusion of the low-pass CIFAR10 experiment is generally applicable to other datasets, e.g., a LeNet trained on LP-MNIST (bandwidth of 14) has 0% drop in accuracy when tested on MNIST data. We will include this result in the appendix alongside Sec. C and Sec. H (results on HP-MNIST)

Low-pass adv. training (R2) As seen in Fig. 7, during adv. training, the energy of the perturbations has a very small high-freq. component, and it is predominantly concentrated in the low freqs. However, it seems that the small, but non-zero components in the high freqs. of the perturbations are necessary to improve robustness, as training only using low-freq. perturbations do not yield satisfactory robustness results.