First, we thank all the reviewers for their invaluable assessment of our paper in this challenging time. As they agree, the general idea of AdvFlows is sound and promising, and the paper is well-written and self-contained. In the following, we address some of the questions raised by the reviewers as much as time and space allows.

Overview The final goal of designing adversarial attacks is gaining a better insight into the pitfalls of DNNs, ultimately alleviating such threats. In this regard, designing attacks with a statistical flavor is extremely valuable as they: 1) provide a unifying framework of modeling DNNs' adversarial vulnerability, and more importantly, 2) help in establishing the required connection with mature fields like high-dimensional statistics to use their results in finding the ultimate solution to making DNNs more robust. Having these in mind, we have come up with AdvFlow that can be viewed as an important step in this direction.

Why NFs and not GANs? The ability of *Normalizing flows* (NF) for efficient inference and sampling, as well as their straightforward and stable training, made them an ideal candidate for our purpose of designing a black-box attack with a statistical perspective. Note that *generative adversarial networks* (GAN) have many disadvantages for use in the current framework: 1) It is known that GANs suffer from *mode collapse*, where they fail to represent different modes of data equally well. In contrast, flow-based models are trained to maximize the log-likelihood, and as such, they cover different modes of data better. 2) Finding the latent space representation of data in GANs requires solving a non-convex optimization problem by back-propagating through the model for every new attack. However, the proposed NF models are invertible by design, and to find the latent space representation of an image, one only needs to query the model. 3) More importantly, GANs neither represent an explicit distribution nor enable inference and density computation. The current design, however, enables further investigation of the attacker distribution properties in the future.

Attack strength The primary purpose of the current work is to convey the idea of blending statistical methods like normalizing flows and adversarial attacks so that we can better understand such threats. Thus, we aimed to compare with recent, but widely recognized black-box attacks for comparison. Nevertheless, by doing more rigorous hyper-parameter tuning or adding extra variables (like σ as correctly indicated by R2), the results can be improved further.[1]

Ablation study on adversarial example detectors To provide more reliable evidence that AdvFlow's distributional properties are fooling the adversarial example detectors, we perform the following ablation study. First, we use an untrained (denoted by un.) AdvFlow model that is initialized randomly. Then, we use the trained version (denoted by tr.) of the same architecture to perform black-box attacks. Using examples generated by these two models, we then train adversarial example detectors to spot the adversaries from clean images. In the paper, we used the Mahalanobis detector [26], a well-known SOTA adversarial example detector. For the sake of completeness, we also add LID [31] (the previous SOTA) and Res-Flow [58] (the recently introduced SOTA) alongside Mahalanobis detector. We compare our results with $\mathcal{N}$ATTACK, *which also approaches the black-box adversarial attack from a distributional perspective* for a fair comparison. The results are given in Table 1. As shown, only if we pre-train our method on clean data, we can fool the detectors. This is indicating that the attacker's distributional properties are fooling the detectors.

Performance comparison with SimBA [59] Note that SimBA [59] was not included in the original manuscript as it is designed for efficient ℓ_2 attacks. At the time of writing the paper, it was not clear how it can be generalized to ℓ_∞. Not until after the NeurIPS deadline did the authors include a generalized version for ℓ_∞, alongside the explanations.[2] We repeat the CIFAR-10 experiments of the paper using the recent version of SimBA-DCT, and report the results in Table 2. For a fair comparison, we compute the average and median of queries on examples where both methods have succeeded. As seen, we get similar results to Table 2 of the paper, outperforming SimBA in defended baselines.[3]

Table 1: Adv. example detection on CIFAR-10.

	AUROC(%)↑		
Detector	$\mathcal{N}$ATTACK	Ours (un.)	Ours (tr.)
LID [31]	78.69	84.39	**57.99**
Mah. [26]	97.95	99.50	**66.85**
Res. [58]	97.90	99.40	**67.03**

Table 2: Performance comparison with SimBA [59] on CIFAR-10.

	Success Rate(%)↑		Query Avg. ↓		Query Med. ↓	
Defense	SimBA	Ours	SimBA	Ours	SimBA	Ours
Vanilla	**99.98**	**99.42**	**238.08**	949.55	**126**	400
FreeAdv	35.52	**41.21**	497.97	**458.35**	256	**200**
FastAdv	35.07	**40.22**	**469.15**	**477.77**	245	**200**
RotNetAdv	35.63	**40.67**	499.75	**453.26**	267	**200**

[1] Note that some of the current SOTA results in black-box adversarial attacks come from the attacker's knowledge about the gradients of the target classifier using substitute models. However, once the target changes its training procedure (e.g., from vanilla to adversarial training), the performance of such methods drop significantly. In contrast, **our method is trained only on clean data and does not depend on any substitute network**. As such, it has a considerable advantage against these methods that are currently prevalent.

[2] See the official repo. of SimBA, where it clearly is indicated that the ℓ_∞ attack is added on 2020/06/22, after NeurIPS deadline.

[3] The results of Table 1 and 2 (as well as SVHN) will be added to the camera-ready version.

[58] Zisselman and Tamar. "Deep Residual Flow for Out-of-Distribution Detection." *CVPR*, 2020.
[59] Guo et al. "Simple Black-box Adversarial Attacks." *ICML*, 2019.