Two reviewers had high scores. We will not focus on their minor comments here but will incorporate their suggestions. One reviewer was undecided and had an intermediate score due to detailed questions we will address. The lowest score was given by a reviewer who had rather general concerns. These are all important for follow-up work, as we describe.

R1: Is computing boundary thickness (BT) straightforward? A: Estimating BT (Eq.(1)) on the adversarial direction does not need the exact projection required for margin (Eq.(3)), so it is straightforward, compared to measuring the exact margin. One can obtain the Std Dev of BT estimate using repeated runs. E.g., Fig 4 uses 10 runs and each uses 320 samples. We tried increasing from 320 samples to 1280, which reduces the Std Dev further from 0.037 to 0.020.

R2: Fig 5b. A: The y-axis is the generalization using non-robust features only, as done by [33] (details in Suppl Sec H).

R3: Clean accuracy (acc). A: Clean acc is critical, and we will report it in the main text. However, the robustness of noisy mixup does not arise purely due to increased regularization. To demonstrate, we vary the weight decay in mixup to obtain different levels of regularization and report clean versus robust acc in Figure-R 1 (1). Noisy mixup outperforms mixup with varying regularization levels. Table 1 clean acc: (CIFAR10) mixup/noisy mixup is 96.0/94.4, (CIFAR100) mixup/noisy mixup is 78.3/72.2. The drop of clean acc is expected for robust models [Dan, Wei, Ravikumar, 2020]. E.g., we tried the same network (ResNet18) with adversarial training and achieved 57.6 clean acc on CIFAR100, which is about 20% drop. Another reason for the clean acc drop is the network size, because when we change ResNet18 to ResNet50, the clean acc drop reduces. Specifically, for mixup/noisy mixup using ResNet50, the clean acc is 79.3/75.5, the OOD acc is 53.4/55.5, the robust acc under PGD-20 is (1.0, 1.6, 3.1)/(4.3, 6.4, 10.3) with (8, 6, 4)-unit attack.

R3: Margin takes the minimum and boundary thickness (BT) takes average. A: The classical margin does take minimum, but average margin has also been considered. e.g., see Eq.(9) in [13]. When using their "sum aggregation", [13] averages over both samples and classes. Prop 2.2 stresses that it is the "min version" of BT that reduces to classical margin. We empirically compare BT to both average and minimum margin, see Fig 3c.

R3: Average margin measurement in Fig3c, and large change in BT from $\beta = 0.75$ to $\beta = 1$. A: The average margin in Fig 3c is indeed measured as the expectation. Figure-R 1 (2-6) shows the transition from $\beta = 0.9$ to 1.0. A huge and sudden phase-transition happens as β gets 1.0. This is why BT as we define performs better than margin.

R3: Geometric interpretation of BT. Steep boundary. A: Geometrically, BT refers to weighted average distance between pairs $(x_r, x_s) \sim p$, with weights defined by how much mass of posterior probability difference is captured between level sets $[\alpha, \beta]$. This implies the "slope" of decision function is somewhat related to BT, as pointed by R3. It is hard to solely attribute this to the slope because high dimensional optimization landscape on non-separable data can be tricky. Furthermore, the slope at the decision boundary is a local property, and the BT is more of a global metric that we agree is at least somewhat related to the former. In the toy case when the boundary is steep and classes are well-separated, margin (i.e. BT with $\alpha = 0, \beta = 1$) captures robustness well while $\alpha = 0, \beta = 0.75$ leads to thinner BT as the reviewer points out. But this motivation for margin is too simplistic and is not representative of optimization landscapes in practical problems. E.g., we observe that NNs with a steep boundary overfit and have low robustness, even in adversarial training (see the left-most points in Fig 3a that have small BT and low robustness). Our contribution is to capture robustness for practical NNs, and we provide two knobs (α and β) to generalize the simplistic concept of margin in a principled way. We observe through exhaustive empirical evidence and ablation studies that setting $\alpha = 0, \beta = 0.75$ is a much better indicator for robustness than margin (when $\alpha = 0, \beta = 1$).

R3: Isolating effect of introducing the noise class. A: To separate the pure effect of the additional class in noisy mixup, we follow R3's advice and measure ordinary mixup on CIFAR10 with the 11th class but only mix sample pairs within the first ten classes or within the noise class. The clean acc of this scheme is 95.3%, the OOD acc is 82.1%, the black-box robust acc is 55.5%, and the robust acc under PGD-20 attack with (8, 6, 4) units is (4.4, 6.5, 11.3). Comparing with Table 1, we see that noisy mixup is more robust in all of the aspects above.

R4: Boundary thickness (BT) definition is complicated; choices of hyper-parameters. Generalization theory. A: There is subtlety to our definition, but we have tried to explain the intuition and relationship of BT with other concepts, while being mathematically precise. See discussions with other reviewers and their reviews on clarity of the paper. Regarding choosing hyperparameters, we have provided extensive ablation study on choosing different hyper-parameters, including α, β (Section E.3.2 and Section F.3), and p (Section E.2). In all experiments, a thick boundary improves robustness. Generalization theory is important follow-up, but it is outside the scope of this paper. We hope this reviewer can conclude that our (empirical and theoretical) findings are worthy of being shared with the ML community.

R4: Other algorithms to explicitly increase BT. Noisy mixup is weakly connected. A: We have justified theoretically for using mixup to increase BT during training (Prop. A.1). Noisy mixup indeed explicitly increases BT because it increases BT both amongst training samples (like mixup) as well as between training samples and noise samples. We agree that there may be other better ways to increase BT for OOD errors than just using the mixup on noise class. However, we show that this simple scheme works reasonably well, and also show the corresponding increase in BT in Fig 4. We hope that by introducing BT to the community, our work motivates further work in this direction.

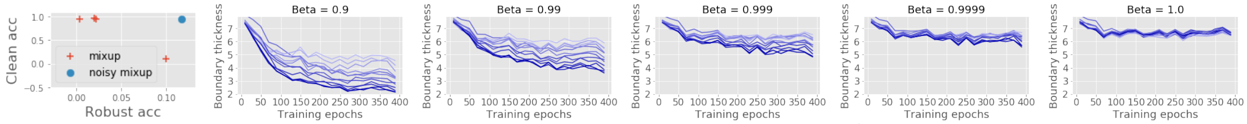

Figure-R 1: For R3: (1) Comparing noisy mixup and mixup with different weight decay. Varying the weight decay (ℓ_2 regularization) only cannot match the performance of noisy mixup. (2-6) Reimplementing Fig 3c with different β when $\alpha = 0$. There is a large range of β's in which boundary thickness can measure robustness.