We thank the three reviewers for their helpful feedback. We were happy to see that the reviewers were generally positive about the manuscript and its contributions: *"There are major algorithmic and empirical contributions in this paper."* (R1), *"seem to take reproducibility seriously"* (R2) and *"The paper is well written"* (R2).

Further improvement: comparison to Adam-PGD (R1): We now include Adam-PGD in our evaluations. It's better than PGD but still mostly outperformed by our attack (Figure 1, a).

Further improvement: extension to L_0 and L_1 (R1): We extended our approach to L_0 and L_1. Here, our attack shows even larger improvements over SOTA than on L_∞ and L_2, see Figure 1 (b, c). The comparison to EAD is still running but the preliminary results look similar. Full results will be included in the manuscript.

Comparison to ECOS and SCS (R1): Solvers like ECOS and SCS run into problems in our setting because they solve general cone problems. They usually compute a sparse QR decomposition of a constraint qualification matrix. For ImageNet, the matrix has a height and width of $224 \cdot 224 \cdot 3 + 1 \approx 150000$, for which even state-of-the-art commercial QR solvers struggle with numerical problems. By making use of the special structure of the problem we need to solve (only one equality constraint and simple box constraints), we can avoid factoring any matrices. For L_2 on ImageNet, this decreases runtime from 10-20s to 2-20ms per iteration, underlining the relevance of our algorithmic contributions.

Percentages, definition of query, runtime (R2): The percentages reported on pg. 4 are model accuracies under attacks bounded by the stated constraints (e.g., $L_\infty < 0.3$). A query includes one forward and backward pass of the attacked model, yielding model decision and gradient. The runtime difference between a standard gradient attack and our attack is $< 5\%$: the computational complexity of our attack is negligible compared to the model evaluation.

Clarity and Contributions (R3): R3 pointed out some clarity issues with regard to our distinction between black-box and white-box attacks. We believe that due to this issue we failed to convey the contributions of our work to R3. The original boundary attack is black-box as it only requires the final model decisions (classifications) to craft adversarials. Our contribution is to adopt the high-level idea of the boundary attack for a gradient-based white-box attack. Compared to the original boundary attack, which often needs 100000 queries to craft reasonably small adversarials (Brendel et al. 2018, Figs 6,7), our version usually requires 10 to 1000 queries until convergence (which is why we focus on the comparison with other white-box attacks). Note that our attack is not a simple adaption of the original boundary attack (which does not estimate the boundary but just makes random steps). We here formulated a completely new algorithm that is able to use the gradient information by solving a box-constraint trust-region problem. To solve this subproblem we had to develop highly specialized algorithms (see our discussion above). In addition, we developed attacks for L_0, L_1, L_2 and L_∞ while the original boundary attack works only for L_2. Our proposed attacks also drastically differ from all existing white-box attacks: while virtually all existing attacks start around the original image and follow the gradient towards the closest adversarial (the attacks differ mainly in the optimizer, loss function or clipping), we here start from a point far away from the original image and follow the boundary to minimize the adversarial distance.

Evaluation of attack budget (R3): Most papers evaluate attacks with a fixed budget, making it difficult to understand how well the attack works for smaller or larger budgets. To show the full picture, we plot the success for all budgets between 1 and 1000 queries (Fig. 3). For each budget (x-axis) we show model success when using the optimal hyperparameters for this budget. This leads to a fairer comparison since we can't cherry pick the attack budget.

Further improvement: Variability over samples (R3): Below, we show for the MNIST-Madry model in the L_2 case distortion curves with additional curves for individual samples (will be in the final paper for all models).

Robustness to gradient masking (R3): Gradient masking denotes the phenomenon where the decision landscape is very flat around evaluated samples. However, at some point the decision has to change and naturally, here the gradients will be even larger than without gradient masking. By walking along the decision boundary, we are exactly in this region of maximal gradients. Finding the decision boundary is also robust since we use a simple binary search.

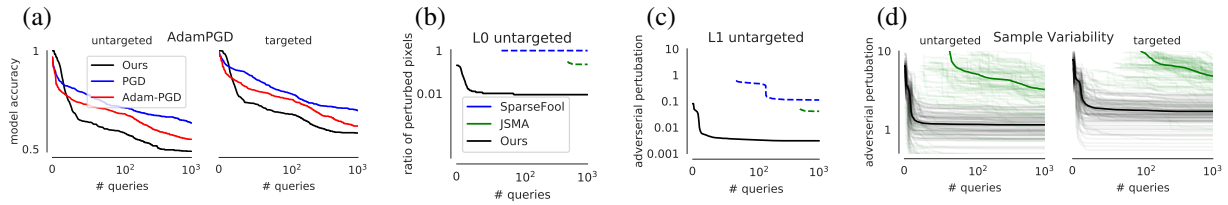

Figure 1: All query distortion/accuracy curves for Madry-MNIST: (a) L_∞ metric, comparing our proposed attack (black) with PGD (blue) and Adam-PGD (red). (b) L_0 metric, untargeted case. (c) L_1 metric, untargeted case. (d) L_2 metric, comparing our proposed attack (black) and C&W (green) with additional curves for individual target images.