Appendix

In this section we first show adversarial examples obtained by different ℓ_p attacks on MNIST and CIFAR10 data for visual comparison. These examples highlight the different behavior exhibited by each attack. We then report the query-distortion curves for all datasets, models and attacks used in this paper, showing that our attack outperforms current attacks on the ℓ_1 norm and rivals their performance on other norms, while typically converging with much fewer queries.

A1. Adversarial Examples

In Figs. 3-4, we report adversarial examples generated by all attacks against model M2 and C2, respectively, on MNIST and CIFAR10 datasets, in the untargeted scenario.

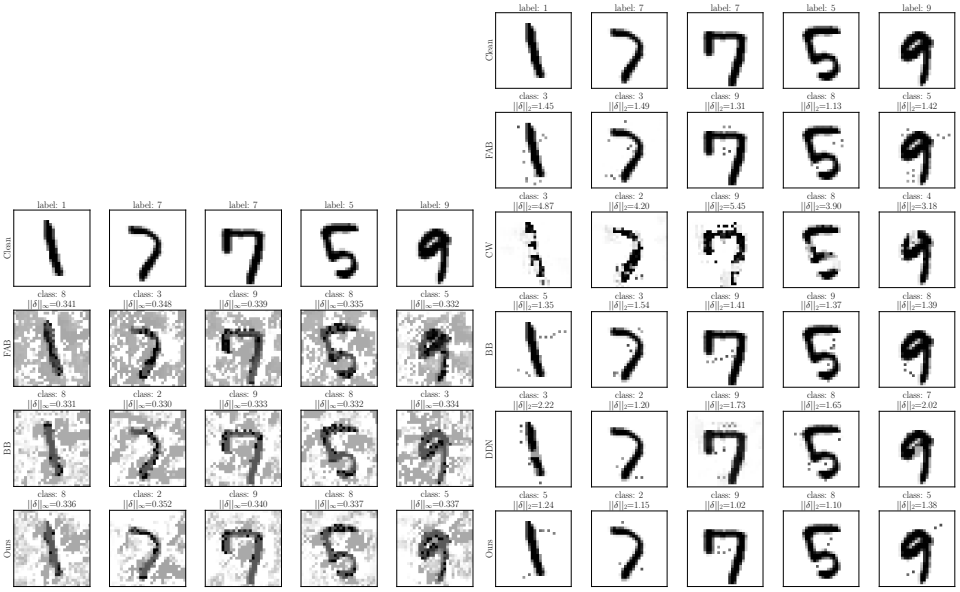
The clean samples and the original label are displayed in the first row of each figure. In the remaining rows we show the perturbed sample along with the predicted class and the corresponding norm of perturbation $\|\boldsymbol{\delta}^\star\|_p$. It is worth noting that the output class for different untargeted attacks is not always the same, which might sometimes explain differences in the perturbation sizes. An example is given in Fig. 4b, where the sample in the fourth column, labeled as "ship", is perturbed by most of the attacks towards the class "airplane", while in our case it outputs the class "dog" with a much smaller distance.

A2. Query-distortion Curves

In Sect. 3.2 we introduced the query-distortion curves as an efficiency evaluation metric for the attacks. We report here the complete results for all models, in targeted and untargeted scenarios.

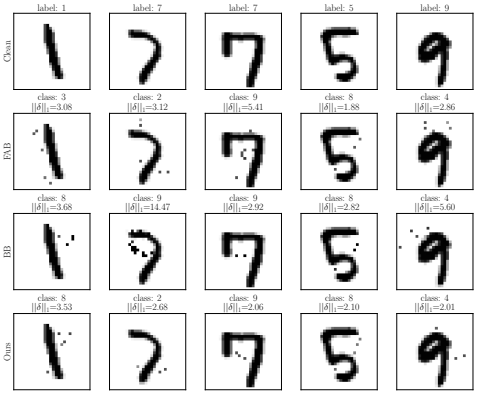
On the MNIST dataset, our attacks generally reach smaller norms with fewer queries, with the exception of M2 (Figs. 5-6), where it seems to reach convergence more slowly than BB in ℓ_0 and ℓ_∞. In ℓ_2, the CW attack is the slowest to converge, due to the need of carefully tuning the weighting term, as described in Sect. 4.

On the CIFAR10 dataset (Figs. 7-8), our attack always rivals or outperforms the others, with the notable exception of DDN for the ℓ_2 norm, which sometimes finds smaller perturbations more quickly, as also shown in Table 3.

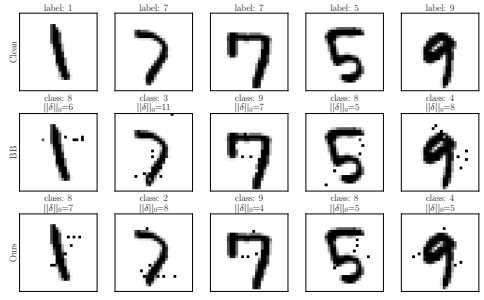


(a) Untargeted ℓ_∞ attacks against M2 [17].

(b) Untargeted ℓ_2 attacks against M2 [17].

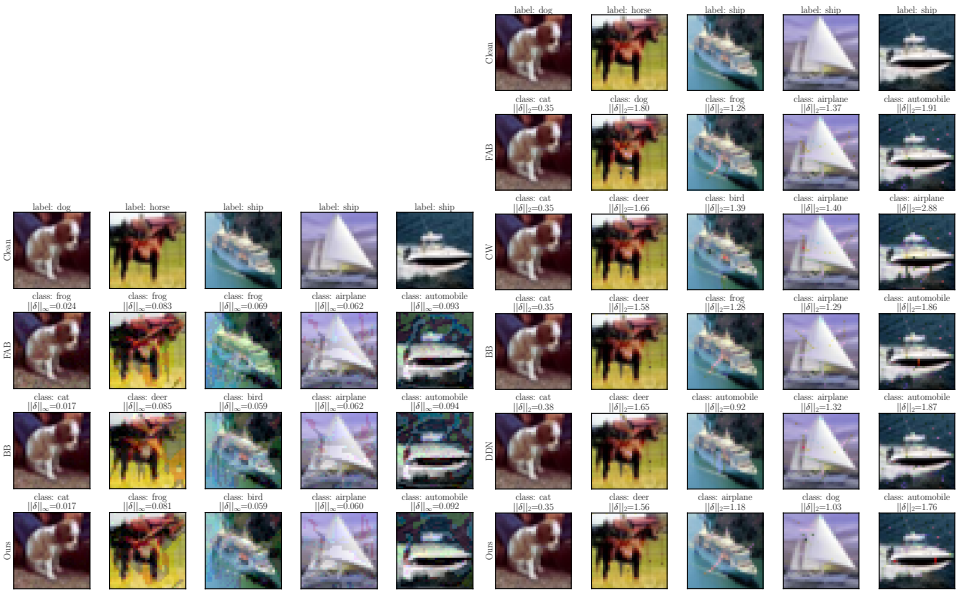


(c) Untargeted ℓ_1 attacks against M2 [17].



(d) Untargeted ℓ_0 attacks against M2 [17].

Figure 3: Adversarial examples on MNIST dataset.



(a) Untargeted ℓ_∞ attacks against C2 [7].

(b) Untargeted ℓ_2 attacks against C2 [7].

(c) Untargeted ℓ_1 attacks against C2 [7].

(d) Untargeted ℓ_0 attacks against C2 [7].

Figure 4: Adversarial examples on CIFAR10 dataset.

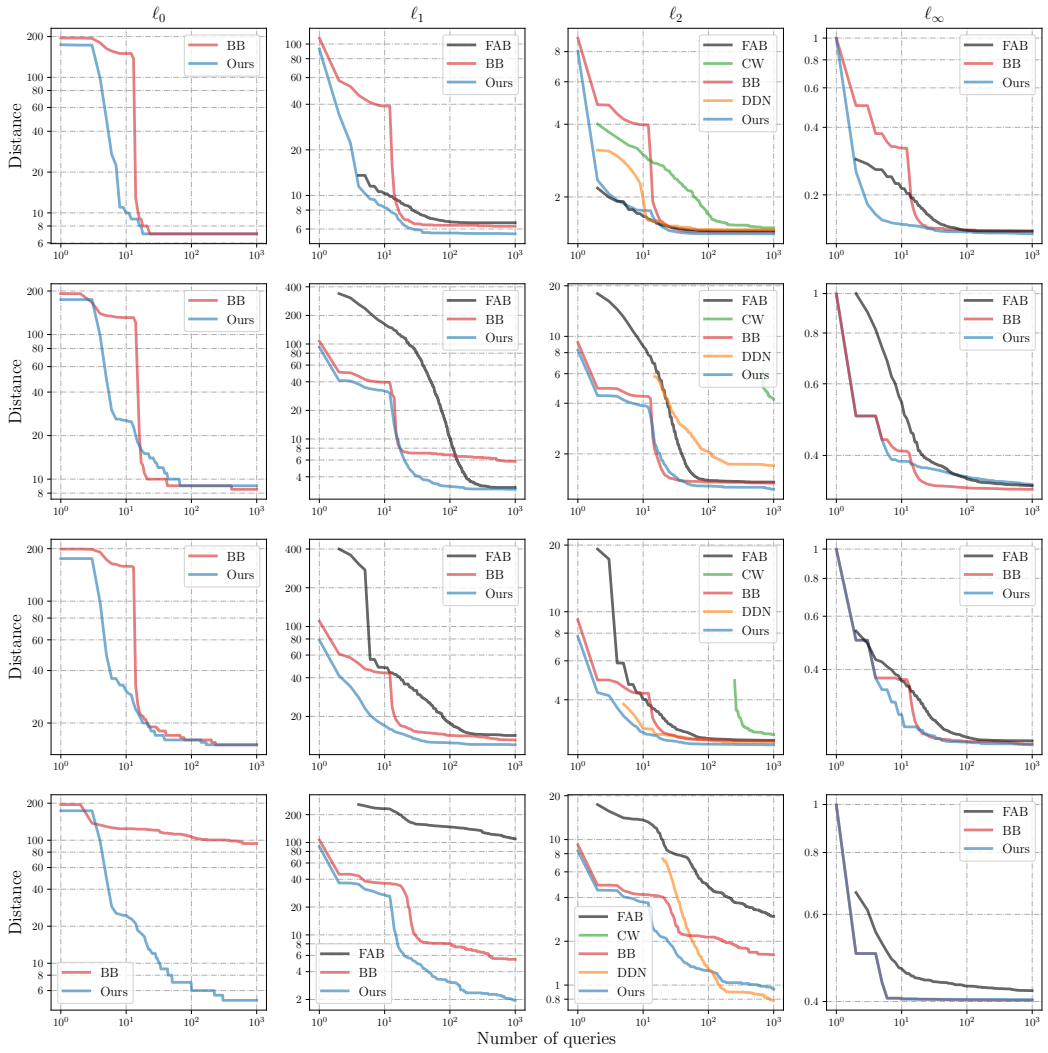


Figure 5: Query-distortion curves for untargeted (U) attacks on the M1, M2, M3, and M4 MNIST models.

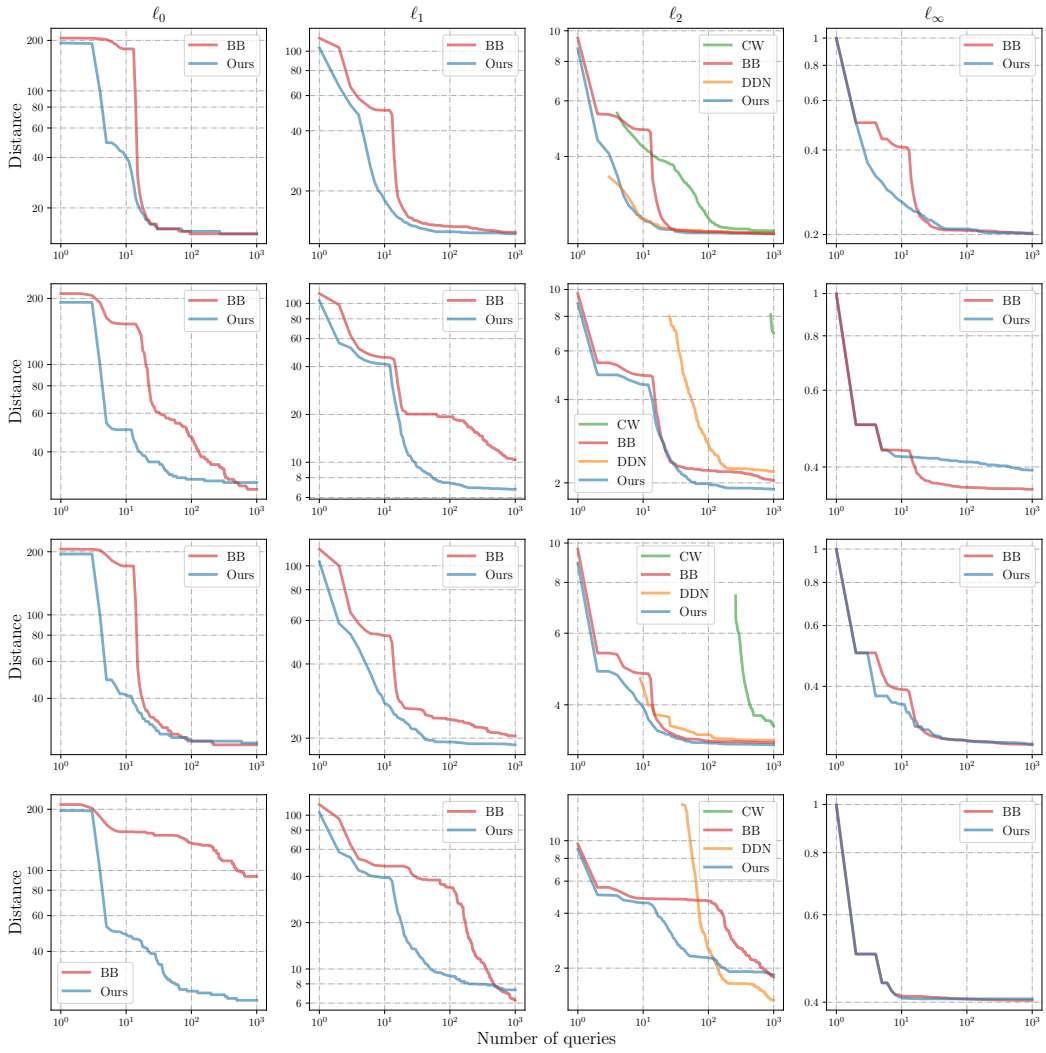


Figure 6: Query-distortion curves for targeted (T) attacks on the M1, M2, M3 and M4 MNIST models.

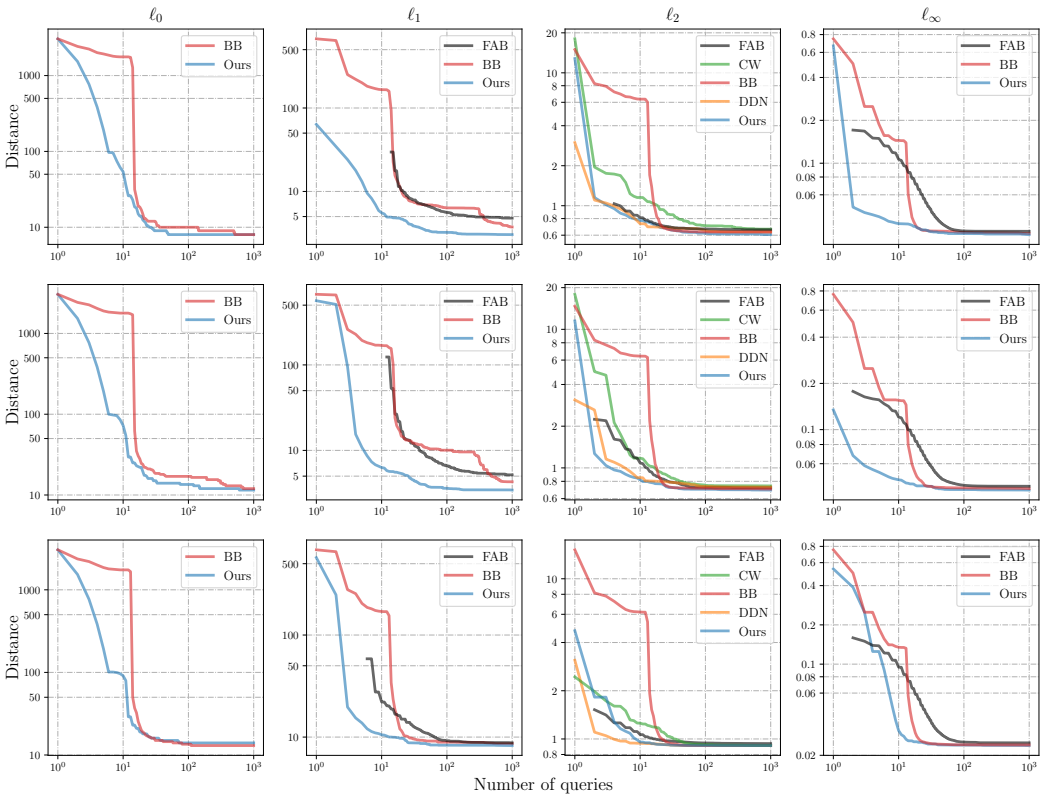


Figure 7: Query-distortion curves for untargeted (*U*) attacks on the C1 (*top*), C2 (*middle*), and C3 (*bottom*) CIFAR10 models.

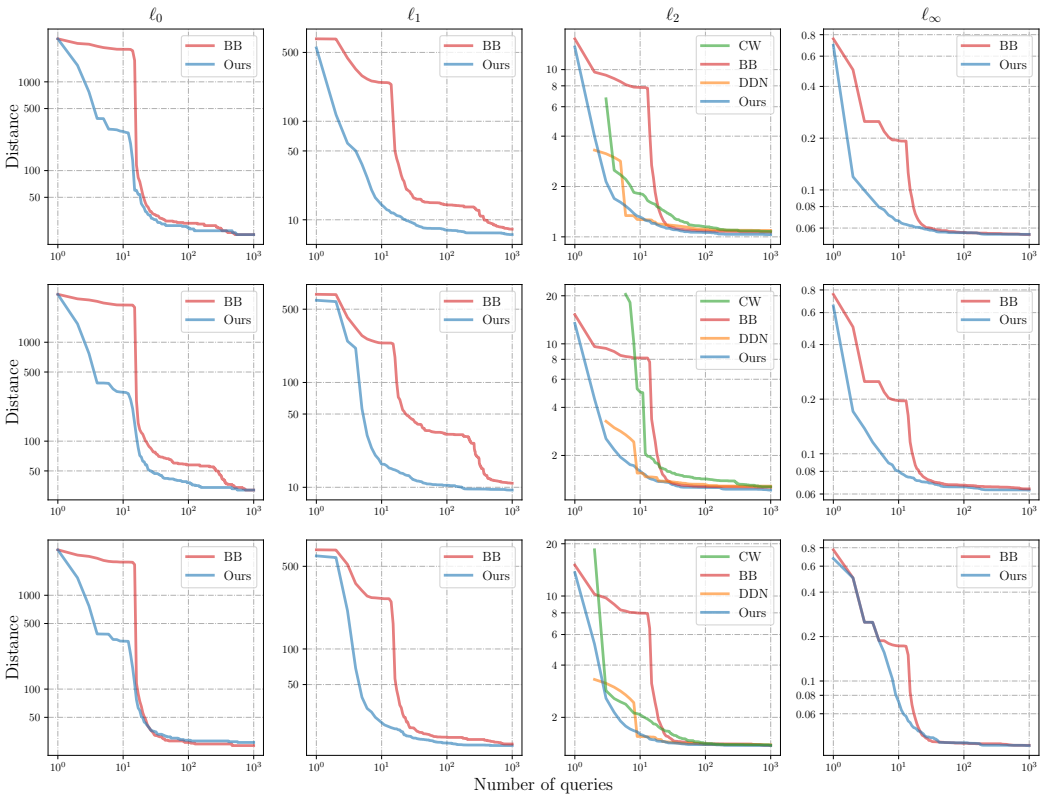


Figure 8: Query-distortion curves for targeted (T) attacks on the C1 (*top*), C2 (*middle*), and C3 (*bottom*) CIFAR10 models.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are discussed in Sect.5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] The potential negative societal impacts are discussed in the end of Sect. 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The experiments clearly describe our evaluation protocol, and the code will be attached in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify all hyperparameters and discuss how we selected them in Sect. 3.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We specify the hardware used in Sect. 3, and provide an analysis of computational and time requirements in Table 3 and Table 2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We acknowledged the projects used for the experiments in the code as dependencies, and linked the project URL of the additional resources.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]