A Appendix

A.1 Robust Models

The experimental setup described in this paper (Sect. 3.1) utilizes pre-trained baseline and robust models obtained from RobustBench Croce et al. (2021). The goal of RobustBench is to track the progress in adversarial robustness for ℓ_∞ and ℓ_2-norm attacks since these are the most studied settings in the literature. We summarize in Table 2 the models we employed in our paper. Each entry in the table includes the label reference from RobustBench, the short name we assigned to the model, and the corresponding clean and robust accuracy under the specific threat model. The robustness of these models is evaluated against an ensemble of white-box and black-box attacks, specifically AutoAttack. Complementary, we also include models trained to be robust against ℓ_1 sparse attacks, i.e., (Croce & Hein, 2021b) and (Jiang et al., 2023). Our experimental setup is designed to encompass a wide range of model architectures and defensive techniques, ensuring a comprehensive and thorough performance evaluation of the considered attacks.

Table 2: Summary of Robustbench Croce et al. (2021) models used in our experiments. For each model, we report its reference label in Robustbench Croce et al. (2021), its threat model, and corresponding clean and robust accuracy.

Dataset	Reference	med	Threat model	Clean accuracy %	Robust accuracy %
CIFAR10	Standard	C1 (Croce et al., 2021)	-	94.78	0
	Carmon2019Unlabeled	C2 (Carmon et al., 2019)	ℓ_∞	89.69	59.53
	Augustin2020Adversarial	C3 (Augustin et al., 2020)	ℓ_2	91.08	72.91
	Engstrom2019Robustness	C4 (Engstrom et al., 2019)	ℓ_∞- ℓ_2	87.03 - 90.83	49.25 - 69.24
	Gowal2020Uncovering	C5 (Gowal et al., 2021)	ℓ_2	90.90	74.50
	Chen2020Adversarial	C6 (Chen et al., 2020)	ℓ_∞	86.04	51.56
	Xu2023Exploring_WRN-28-10	C7 (Xu et al., 2023)	ℓ_∞	93.69	63.89
	Addepalli2022Efficient_RN18	C8 (Addepalli et al., 2022)	ℓ_∞	85.71	52.48
Imagenet	Standard_R18	I1 (He et al., 2015)	-	76.52	0
	Engstrom2019Robustness	I2 (Engstrom et al., 2019)	ℓ_∞	62.56	29.22
	Wong2020Fast	I3 (Wong et al., 2020)	ℓ_∞	55.62	26.24
	Salman2020Do_R18	I4 (Salman et al., 2020)	ℓ_∞	64.02	34.96
	Hendrycks2020Many	I5 (Hendrycks et al., 2021)	ℓ_∞	76.86	52.90
	Debenedetti2022Light_XCiT-S12	I6 (Debenedetti et al., 2023)	ℓ_∞	72.34	41.78

A.2 σ−zero Objective Function Visualization

In Figure 4, we depict the behavior of the loss terms of σ−zero when applied to the Imagenet data sample, specifically, the frog in Figure 1. When the sample is not adversarial, the attack algorithm increases the ℓ_0 norm, highlighted by the bumps in the orange curve, to find a valid adversarial $\boldsymbol{\delta}$. Conversely, when an adversarial example is found, the loss term is cropped to zero, and the algorithm focus solely on minimizing the ℓ_0 in $\boldsymbol{\delta}$.

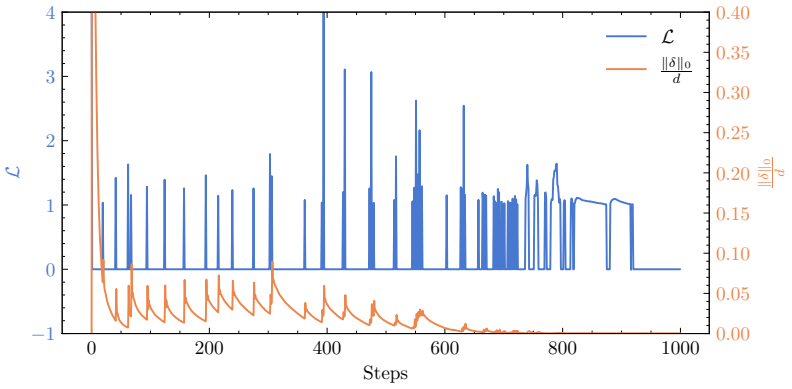

Figure 4: σ−zero loss terms during the optimization procedure.

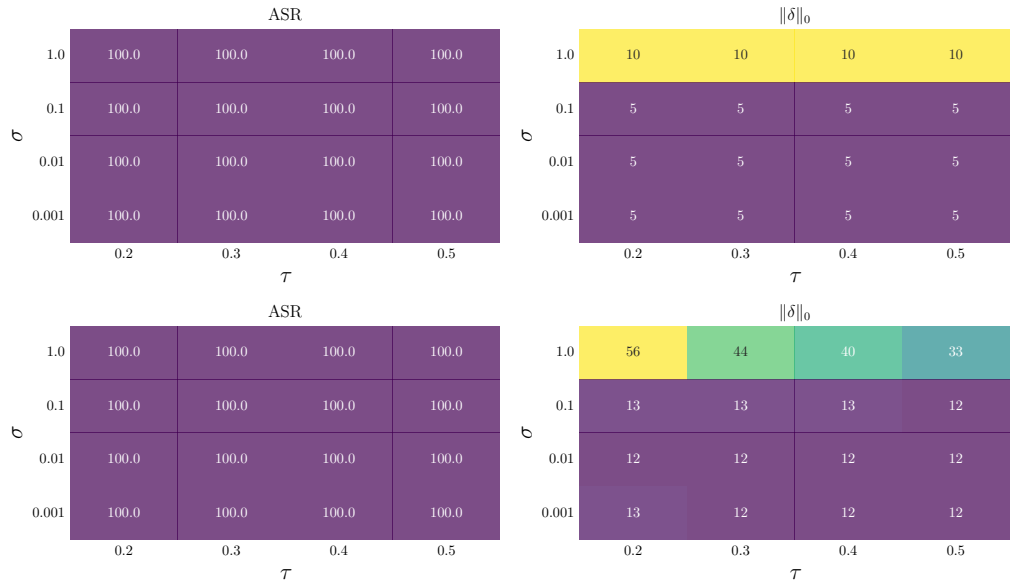

Figure 5: Ablation study on σ (y-axis) and τ (x-axis) for CIFAR10 C1, (top-row), CIFAR10 C8, (bottom-row). For each combination, we report the attack success rate (leftmost matrix) and the ℓ_0 norm on the output adversarial perturbation (rightmost matrix).

A.3 Ablation Study

To assess the strength and potential limitations of our proposed attack, we conducted an ablation study on its key hyperparameters. Specifically, we investigated the impact of varying two critical parameters, τ and σ. The parameter τ governs the tolerance threshold in Algorithm 1, which induces sparsity within the adversarial noise. Conversely, σ defines the approximation quality of $\hat{\ell}_0$ in Equation 4 compared to the actual ℓ_0 function. Our ablation study, depicted in Figure 5, involved two distinct models: C1 (top row), C8 (bottom row). We executed the attack on 1000 randomly selected samples from each dataset and recorded the Attack Success Rate (ASR) and the median ℓ_0 norm of the resulting adversarial perturbations. Remarkably, we observe a significant robustness of σ-zero with respect to these two hyperparameters, except for the extreme case of $\sigma = 1$. With regard to the τ parameter, it is observed that the choice of the initial value exerts negligible influence on the ultimate outcome, given that the parameter dynamically adapts throughout the optimization process. Concerning σ, as also noted in Cinà et al. (2022), its selection is not particularly challenging, especially when incorporating the sparsity projection operator. Overall, the ablation study revealed consistent trends across the models. In all cases, we identified a broad parameter configuration range where our attack maintained robustness, making hyperparameter optimization for the attacker a swift task. This robustness is further evidenced by the results presented in Table 1, where our attack consistently outperforms state-of-the-art attacks even with a shared hyperparameter configuration across all models.

A.4 Attack Comparison with 100 steps

In our experimental setup, we also explore scenarios where the attacker's access to queries is limited, thus reducing the number of iterations for the attack. To simulate this perspective, we replicate our experimental comparison involving σ-zero and state-of-the-art sparse attacks while restricting the number of steps to 100. The results are summarized in Table 3. Notably, compared to the results presented in Table 1, most competitive attacks undergo a decrease in their ASR, while σ-zero consistently maintains a 100% success rate. In conclusion, σ-zero remains a promising choice for crafting minimum ℓ_0-norm attacks against DNNs, even when the attacker has limited query resources.

Table 3: For each attack, we report the corresponding ASR, median $\|\boldsymbol{\delta}\|_0$, sample-level average execution time and the number of queries, and the maximum VRAM consumed during the execution. * in VFGA indicates that the usage of a smaller batch size, to fit its execution in memory, may have led to a slight overestimation of the execution time.

		Performance				Computational Effort				Performance				Computational Effort		
Attack	Model	ASR(%)	$\text{ASR}(\%)_{10}$	$\text{ASR}(\%)_{50}$	$\|\boldsymbol{\delta}\|_0$	t(s)	q (x1000)	VRAM	Model	ASR(%)	$\text{ASR}(\%)_{10}$	$\text{ASR}(\%)_{50}$	$\|\boldsymbol{\delta}\|_0$	t(s)	q (x1000)	VRAM
MNIST																
EAD	M1	100.0	1.11	46.65	52.0	0.09	1.14	0.05	M2	100.0	1.2	35.57	61.0	0.07	0.99	0.05
VFGA		95.71	9.57	82.56	27.0	0.04	0.74	0.21		92.32	1.81	39.28	57.0	0.04	1.3	0.21
PDPGD		100.0	0.98	0.98	359.0	0.02	0.2	0.04		95.02	0.52	0.52	254.0	0.02	0.2	0.04
BB		100.0	12.8	98.0	20.0	0.13	1.19	0.05		87.87	26.53	83.0	18.0	0.12	1.69	0.05
FMN		88.93	7.22	83.09	30.0	0.01	0.2	0.04		14.81	4.02	14.03	∞	0.01	0.2	0.04
σ-zero		100.0	**12.46**	**98.55**	**21.0**	0.03	0.2	0.05		100.0	**38.3**	**99.88**	**13.0**	0.03	0.2	0.05
CIFAR10																
EAD	C1	100.0	6.82	19.09	146.0	0.26	0.77	1.47	C5	100.0	14.35	32.94	83.0	1.59	0.82	9.92
VFGA		98.99	49.14	93.46	**11.0**	0.16	0.38	11.96		87.75	27.64	67.1	29.0	3.6*	0.86	> 40
PDPGD		100.0	5.23	5.23	3057.0	0.06	0.2	1.31		99.75	11.26	11.26	2814.0	0.32	0.2	8.86
BB		100.0	**53.48**	97.55	**10.0**	0.59	0.95	1.47		17.92	13.18	16.88	∞	2.67	1.95	9.93
FMN		98.86	**62.85**	97.72	**8.0**	0.05	0.2	1.31		72.34	27.16	61.49	33.0	0.31	0.2	8.86
σ-zero		100.0	42.54	**99.11**	12.0	0.08	0.2	1.47		100.0	**35.09**	**85.95**	**18.0**	0.44	0.2	9.92
EAD	C2	100.0	12.74	28.74	100.0	0.27	0.8	1.47	C6	100.0	16.95	29.59	128.0	0.7	0.66	5.4
VFGA		93.69	28.99	75.38	24.0	0.22	0.72	11.71		97.08	34.27	82.07	20.0	4.24*	0.61	> 40
PDPGD		99.39	10.31	10.31	2421.0	0.06	0.2	1.32		54.16	13.96	13.96	3072.0	0.2	0.2	5.12
BB		14.97	11.58	14.29	∞	0.44	1.95	1.47		84.46	35.89	78.14	17.0	1.68	1.67	5.39
FMN		80.68	28.06	69.36	27.0	0.05	0.2	1.31		87.1	32.98	76.59	22.0	0.19	0.2	5.12
σ-zero		100.0	**32.58**	**86.2**	**18.0**	0.07	0.2	1.47		100.0	**37.99**	**90.93**	**16.0**	0.24	0.2	5.39
EAD	C3	100.0	9.14	10.67	451.0	0.31	0.71	1.89	C7	100.0	9.23	21.61	162.0	0.31	0.8	2.15
VFGA		91.64	21.7	66.55	33.0	0.34	0.87	16.53		75.79	22.76	56.58	39.0	1.34*	1.06	> 40
PDPGD		75.31	8.92	8.92	3052.0	0.09	0.2	1.8		96.2	6.31	6.31	2773.0	0.07	0.2	2.0
BB		57.05	19.46	53.53	40.0	0.59	1.9	1.89		100.0	**35.43**	**90.74**	**16.0**	0.64	1.07	2.16
FMN		71.3	20.37	62.41	36.0	0.09	0.2	1.8		68.88	23.21	59.69	35.0	0.06	0.2	2.0
σ-zero		100.0	**23.24**	**85.46**	**24.0**	0.11	0.2	1.89		100.0	**32.42**	**89.12**	**18.0**	0.08	0.2	2.15
EAD	C4	100.0	9.38	10.56	434.0	0.4	0.9	1.89	C8	100.0	15.76	26.17	144.5	0.14	0.79	0.41
VFGA		99.16	30.44	90.13	19.0	0.27	0.52	16.53		94.16	29.57	74.22	26.0	0.13	0.73	3.07
PDPGD		99.9	9.17	9.17	2709.0	0.09	0.2	1.8		90.95	14.29	14.29	3057.0	0.04	0.2	0.36
BB		32.83	16.28	32.41	∞	0.49	1.93	1.89		100.0	**35.98**	**89.39**	**17.0**	0.43	1.13	0.41
FMN		87.2	26.63	79.48	24.0	0.09	0.2	1.8		80.68	29.47	69.77	27.0	0.03	0.2	0.36
σ-zero		100.0	**29.45**	**92.77**	**19.0**	0.11	0.2	1.89		100.0	**32.49**	**85.19**	**19.0**	0.04	0.2	0.41
Imagenet																
EAD	I1	100.0	34.4	35.7	501.5	0.2	0.71	1.21	I4	100.0	56.1	60.7	0.0	0.2	0.77	1.21
VFGA		85.3	47.0	**72.4**	**14.0**	1.74*	0.7	> 40		83.9	61.9	75.4	1.0	1.49*	0.59	> 40
FMN		68.1	**47.6**	64.9	14.0	0.06	0.2	1.14		77.2	63.2	75.1	0.0	0.06	0.2	1.14
σ-zero		100.0	42.9	70.4	20.0	0.08	0.2	1.21		100.0	**69.4**	**88.0**	0.0	0.07	0.2	1.2
EAD	I2	100.0	44.5	50.2	48.0	0.48	0.75	4.36	I5	100.0	26.6	27.4	1105.0	0.4	0.62	4.35
VFGA		72.9	49.2	63.3	12.0	4.34*	0.86	> 40		74.0	**36.5**	**58.9**	**30.0**	4.5*	0.96	> 40
FMN		62.4	50.4	60.0	10.0	0.13	0.2	4.25		53.0	36.0	50.3	46.0	0.13	0.2	4.25
σ-zero		100.0	**54.8**	**79.9**	**6.0**	0.15	0.2	4.35		100.0	**32.3**	**57.1**	**40.0**	0.15	0.2	4.35
EAD	I3	100.0	54.8	60.5	0.0	0.47	0.74	4.35	I6	99.9	32.2	32.9	818.5	1.14	0.72	5.67
VFGA		83.0	62.2	76.0	1.0	3.47*	0.59	> 40		57.4	35.4	**46.9**	67.0	13.35*	1.2	> 40
FMN		74.2	63.3	71.8	0.0	0.13	0.2	4.25		47.5	**35.7**	45.2	∞	0.34	0.2	5.54
σ-zero		100.0	**69.9**	**86.8**	0.0	0.15	0.2	4.35		100.0	**36.9**	**53.3**	**40.0**	0.41	0.2	5.68

A.5 Comparison with Maximum-confidence Attacks

In our experimental comparisons presented in both Table 1 and Table 3, we include the Sparse-RS and PGD-ℓ_0 attacks introduced by (Croce et al., 2022) and (Croce & Hein, 2019). These attacks have been designed to generate sparse adversarial perturbations given a fixed budget k. Specifically, in their threat model, the attacker imposes a maximum limit on the number of perturbed features, and the attack then outputs the adversarial example that minimizes the model's confidence in predicting the true label of the sample. However, since the fixed-budget threat model differs from the minimum-norm scenario we consider in this paper, as we do not assume a maximum budget, we have developed a wrapper around Sparse-RS and PGD-ℓ_0 to ensure a fair comparison. Similarly to (Rony et al., 2021b), we employ a *sample-wise binary* search strategy to determine the smallest budget, denoted as k, that must be provided as input to them to achieve an Attack Success Rate (ASR) equal to 100% on the data. At each iteration of the binary search, if the attack is successful, we halve the value of k, asking Sparse-RS and PGD-ℓ_0 to perturb fewer features. Conversely, if the sample is not adversarial, we double the value of k. Thus, we enable the attacker to find the value of k that results in the

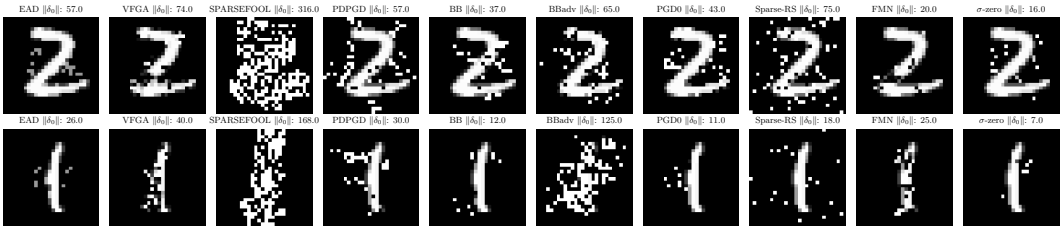

Figure 6: Randomly chosen adversarial examples from MNIST M2.

minimum ℓ_0 norm for each sample. Finally, the comparison tables showcase the best result with the minimum norm perturbation in the median across the iterations.

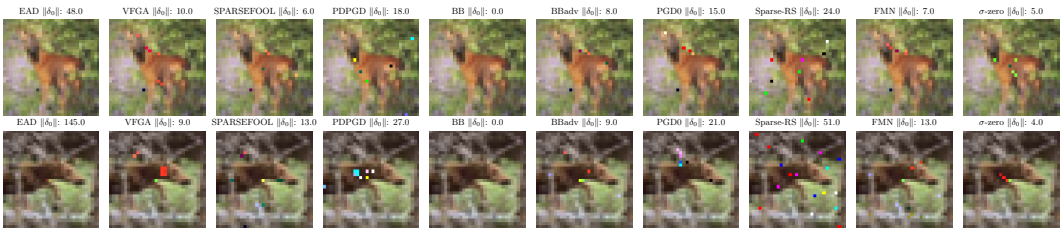

Figure 7: Randomly chosen adversarial examples from CIFAR10 C2.

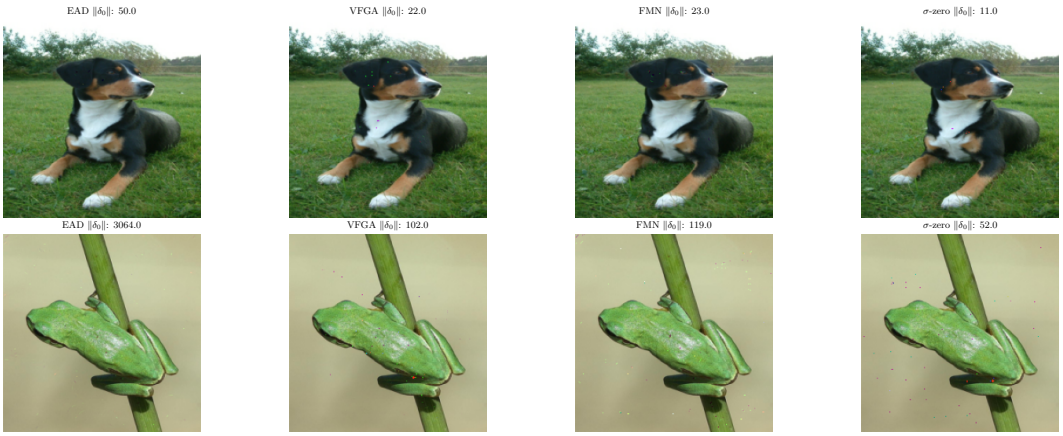

Figure 8: Randomly chosen adversarial examples from Imagenet I1.