Exploring Sparsity and Smoothness of Arbitrary Lp Norms in Adversarial Attacks

Christof Duhme
Department of Computer Science
University of Münster
Münster, Germany

Florian Eilers
Department of Computer Science
University of Münster
Münster, Germany

Xiaoyi Jiang
Department of Computer Science
University of Münster
Münster, Germany

***Abstract*—Adversarial attacks against deep neural networks are commonly constructed under ℓ_p norm constraints, most often using $p = 1$, $p = 2$ or $p = \infty$, and potentially regularized for specific demands such as sparsity or smoothness. These choices are typically made without a systematic investigation of how the norm parameter p influences the structural and perceptual properties of adversarial perturbations. In this work, we study how the choice of p affects sparsity and smoothness of adversarial attacks generated under ℓ_p norm constraints for values of $p \in [1, 2]$. To enable a quantitative analysis, we adopt two established sparsity measures from the literature and introduce three smoothness measures. In particular, we propose a general framework for deriving smoothness measures based on smoothing operations and additionally introduce a smoothness measure based on first-order Taylor approximations. Using these measures, we conduct a comprehensive empirical evaluation across multiple real-world image datasets and a diverse set of model architectures, including both convolutional and transformer-based networks. We show that the choice of ℓ_1 or ℓ_2 is suboptimal in most cases and the optimal p value is dependent on the specific task. In our experiments, using ℓ_p norms with $p \in [1.3, 1.5]$ yields the best trade-off between sparse and smooth attacks. These findings highlight the importance of principled norm selection when designing and evaluating adversarial attacks.**

I. Introduction

Deep neural networks are successfully applied in many academic, corporate and public areas, including safety critical applications such as health care or autonomous driving. In these settings, adversarial attacks pose a serious security threat [1], [2], as small and often imperceptible perturbations to the input can lead to incorrect model predictions. Understanding the structure and properties of adversarial attacks is therefore essential for designing models that are robust against such threats.

The general objective of adversarial attacks is to modify an input in a way that is difficult for humans to perceive while causing a model to change its prediction. Adversarial perturbations are typically constrained to be small according to some norm. In addition to the overall magnitude of the perturbation, both sparsity and smoothness have been shown to play an important role in the perceptual quality of adversarial examples [3], [4].

Adversarial attacks are often framed as an optimization problem

$$\max_{||\delta||_p < \varepsilon} l(f(x+\delta), y) \tag{1}$$

where f denotes a neural network, x an input image with label y, l a loss function, and $\varepsilon > 0$ controls the attack strength.

Many works have investigated how to regularize this optimization problem to promote smooth [5] or sparse [6], [7] perturbations. Other works have focused on restricting the search space from $\{||\delta||_p < \varepsilon\}$ to a search space that only allows smooth [3], [8], [9] or sparse [4], [10] solutions. In these frameworks, the ℓ_p norm $||\cdot||_p$ is regularly defaulted to $p = 1$, $p = 2$ or $p = \infty$ without further investigation.

To the best of our knowledge, no work so far has explicitly investigated how the choice of p effects the sparsity and smoothness of the adversarial attack. In this work, we study the behavior of adversarial perturbations generated under ℓ_p norm constraints for values of $p \in [1, 2]$. Our analysis focuses on quantifying how sparsity and smoothness change as functions of p across different model architectures and datasets.

Sparsity in adversarial attacks is commonly understood as perturbing only a small number of pixels. Accordingly, many approaches rely on the ℓ_0 norm to enforce sparsity. We adopt a broader notion of sparsity, in which only a small number of pixels in the adversarial attack change significantly but tiny changes in more pixels are neglected, which is not covered by the ℓ_0 constraint. This is in line with [11] who present desirable properties for sparsity measures and evaluate them on commonly used measures. We choose the two sparsity measures investigated in their work that satisfy the most desirable properties.

To the best of our knowledge, no smoothness measures for adversarial attacks have been studied. We introduce a general framework for construction smoothness measures based on smoothing operations. In addition, we propose a smoothness measure derived from first-order Taylor approximations. These measures allow us to quantitatively assess the smoothness of adversarial perturbations generated under different ℓ_p norm constraints with $p \in [1, 2]$ in a white box scenario on different datasets and architectures.

To summarize, our contributions are as follows:

1) We propose a general framework for deriving smoothness measures from smoothing operations.
2) We introduce two smoothness measures based on this framework and additionally propose a smoothness measure based on first-order Taylor approximations.

3) We analyze sparsity and smoothness of adversarial attacks generated under arbitrary ℓ_p norm constraints with these three smoothness measures and two well-known sparsity measures from literature.
4) We show that neither of the standard choices $p = 1$ and $p = 2$ is optimal with regard to sparsity and smoothness, but the optimal choice p depends on both the model architecture and the dataset.

The work is structured as follows. Section II reviews related work on sparsity and smoothness in adversarial attacks. Section III provides methodological background and Section IV introduces the proposed smoothness measures. Section V describes the experimental setup, followed by the presentation and analysis of the experimental results in Section VI. Finally, Section VII concludes the paper.

II. Related Work

Ever since the introduction of adversarial examples [12], extensive research has explored attack and defense methods in computer vision, with recent surveys summarizing a decade of progress across vision tasks and reviewing domain-specific challenges in medical image analysis [13], [14]. In contrast to these broad overviews, we focus on how varying p in ℓ_p-constrained attacks affects the structural properties of perturbations, particularly sparsity and smoothness.

Most attacks are based on approximately solving the optimization problem in (1) using gradient-based methods. But black-box attacks [15] and generative models have also been used to produce adversarial perturbations, including GAN-based approaches [16] and diffusion-based methods [17], [18]. In addition, adversarial attacks in the physical world have been studied in a variety of settings [19], [20].

a) Optimization-based attacks and evaluation: To solve (1) with gradient descent, Projected Gradient Descent (PGD) was introduced [21] and later enhanced to the parameter-free Auto-Projected Gradient Descent (APGD) [22]. Croce and Hein subsequently extended this line of work to ℓ_1-APGD [23], which explicitly respects image-domain constraints (e.g., $x + \delta \in [0, 1]^n$) and was designed to produce sparse perturbations. However, they did not verify this sparsity claim with independent measure. For attacks under more general ℓ_p constraints, Frank-Wolfe [24] based approaches provide an alternative to projected methods by avoiding explicit projections [25], but only focus on $p = \infty$ in the evaluation. Building on this direction, Boreiko et al. introduced Auto-Frank-Wolfe (AFW) as an adaptive Frank-Wolfe scheme to generate perturbations under arbitrary ℓ_p constraints for $p > 1$ [26]. Recent work explores optimization strategies for sparsity beyond classical norms. For example, Cinà et al. proposed gradient-based optimization of generalized zero-norm attacks evaluating relaxations of sparsity measures [27], illustrating that alternative norm relaxations yield improved empirical trade-offs between accuracy and perturbation size.

b) Norm choice and perceptual relevance: Despite the widespread use of pixel-space ℓ_p norms (most commonly $p \in \{1, 2, \infty\}$), the relationship between these norms and human perception is not straightforward. Sen et al. explicitly ask whether adversarial attacks should be evaluated using pixel p-norms and discuss alternative similarity measures [28]. Relatedly, perceptual threat models and attacks based on learned perceptual distances have been explored, for example via neural perceptual threat models [29], and by considering perturbation classes that can be perceptually realistic despite large pixel ℓ_p distances, such as spatial transformations [30]. However, recent meta-attack frameworks such as DAASH demonstrate that combining multiple 'ℓ_p' based attacks with perceptual distortion metrics (e.g., SSIM, LPIPS) yields perturbations with superior perceptual quality even when constrained by pixel norms [31].

c) Sparse adversarial attacks: Sparsity and smoothness are often cited as desirable properties for imperceptible adversarial perturbations, motivating methods that explicitly target these characteristics. To generate sparse attacks, Croce and Hein propose attacks minimizing ℓ_0-distance and variants that additionally control per-pixel changes to improve imperceptibility [4]. Other approaches factorize the perturbation into magnitude and pixel selection components [32], learn distortion maps indicating less perceptually sensitive regions [33], or consider extreme settings such as one-pixel attacks [6].

d) Smooth and frequency-structured attacks: To promote smooth perturbations, Zhang et al. integrate Laplacian smoothing into the attack pipeline [5]. Complementary approaches restrict the search space to low-frequency perturbations [8] or generate perturbations within function classes that yield visually structured changes, such as harmonic functions [3]. More recent approaches target structured sparsity: Lin et al. propose interpretable sparse attacks that minimize the number of perturbed pixels while preserving attack effectiveness [34], and Heshmati et al. introduce ATOS, a framework that generates structured, group-sparse adversarial perturbations that highlight class-relevant features [35].

e) Closest prior work: The most direct prior work to our study is Boreiko et al. [26], who generate sparse visual counterfactual explanations (VCEs) using an intermediate norm $\ell_{1.5}$ and argue qualitatively for a favorable trade-off between sparsity and smoothness. However, their choice of p is fixed and the effect of varying p is not quantified. In contrast, we systematically analyze how varying $p \in [1, 2]$ affects sparsity and smoothness using multiple quantitative measures, and we provide a method to identify the optimal p per model and dataset.

III. Background

We will give a short introduction to adversarial attacks with arbitrary ℓ_p norm constraints as well as sparsity measures.

A. ℓ_p Norms

As defined in (1), adversarial attacks are commonly formulated as constrained optimization problem, where the search space for the perturbation δ is given by the ℓ_p ball $B_\varepsilon^p = \{\delta \in \mathbb{R}^n | \ ||\delta||_p < \varepsilon\}$ where n denotes the image dimension

and $\varepsilon > 0$ controls the attack strength. While most prior work focuses on specific choices $p = 1$, $p = 2$ or $p = \infty$, p does not need to be restricted to these values. For $1 \leq p < \infty$, the ℓ_p norm in $\mathbb{R}^n$ is defined as:

$$||x||_p = \sum_{k=1}^{n} (|x_k|^p)^{\frac{1}{p}} \tag{2}$$

B. Adversarial Attacks for Arbitrary ℓ_p Norms

APGD [23] was introduced to eliminate the need for manual hyperparameter tuning in PGD [21]. While APGD efficiently solves the optimization problem in (1), it does not respect the image domain (usually $[0, 1]$) and relies on clipping after calculation of the attack. To address this limitation for sparse attacks, APGD was extended to ℓ_1-APGD [23], which explicitly respects image-domain constraints while solving the adversarial optimization problem for the ℓ_1 norm.

Both APGD and ℓ_1-APGD are restricted to specific norm choices and cannot be directly applied to arbitrary values of p. Therefore, AFW [26] was introduced as an adaptive version of the Frank-Wolfe algorithm [25] that can iteratively solve the optimization problem for arbitrary ℓ_p norms for $p > 1$, while explicitly respecting the image domain.

Let $w \in \mathbb{R}^n,\ w = \nabla_x l(f(x), y)$ be the gradient. To avoid the necessary projection step as required to solve (1) by APGD Boreiko et al. then reformulate the optimization problem as:

$$\arg\max_{\delta \in \mathbb{R}^n} \langle w, \delta \rangle \text{ s.t. } ||\delta||_p \leq \varepsilon,\ x + \delta \in [0, 1]^n \tag{3}$$

The closed-form solution of this optimization step is given by

$$\delta_i^* = \min \left\{ \gamma_i, \left(\frac{|w_i|}{p\mu^*} \right) \right\} \text{sign} w_i \tag{4}$$

where

$$\gamma_i = \max\{-x_i \text{ sign}(w_i), (1 - x_i) \text{ sign}(w_i)\} \tag{5}$$

and $\mu^* > 0$, which can be computed in $O(n \log n)$ time.

The solution (3) provides a framework to calculate adversarial attacks for $p \geq 1$ while respecting the image constraints $[0, 1]$. We restrict our empirical analysis to $p \in [1, 2]$.

C. Sparsity Measures

Sparsity is commonly cited as a desirable property of adversarial perturbations, as perturbing only a small number of pixels is often associated with increased imperceptibility. It is often just measured by the ℓ_0 measure

$$\ell_0(x) = |\{x_i \neq 0\}| \tag{6}$$

We argue, however, that small perturbations are imperceptible anyways, thus a less strict notion of sparsity that accounts for the distribution of perturbation magnitudes is more informative in the setting of adversarial attacks. Following prior work [11], we adopt sparsity measures that characterize how energy is distributed across coefficients. In this case, the coefficients are the adversarial perturbations. Such measures have desirable properties originally studied in economic settings [36], [37], where sparsity is analogous to wealth concentration.

We choose the Gini Index and the Hoyer measure, since they satisfy the most important properties for fixed-size inputs such as images.

The Gini Index [38] is defined as

$$S_{\text{Gini}}(c) = 1 - 2 \sum_{k=1}^{N} \frac{c_k}{||c||_1} \left(\frac{N - k + \frac{1}{2}}{N} \right) \tag{7}$$

for ordered data $c_1 \leq c_2 \leq ... \leq c_N$. The Gini Index is a weighted sum that remains sensitive to small coefficients and is normalized with respect to the number of coefficients, making it suitable for comparisons across datasets.

The Hoyer measure [39] is defined as

$$S_{\text{Hoyer}}(c) = \left(\sqrt{N} - \frac{\sum_j c_j}{\sqrt{\sum_j c_j^2}} \right) \left(\sqrt{N} - 1 \right)^{-1}. \tag{8}$$

This is a normalized ratio between ℓ_1 and ℓ_2. It is zero if and only if all coefficients are equal, and one if and only if exactly one coefficient is non-zero.

IV. Smoothness Measures

To the best of our knowledge, there are no established quantitative smoothness measures specifically designed for adversarial perturbations. We therefore introduce two complementary approaches to measure smoothness. First, we propose a general framework that derives smoothness measures from smoothing operators. Second, we introduce a smoothness measure based on first-order Taylor approximations. Throughout this section, we interpret higher smoothness values as indicating smoother perturbations.

A. Smoothing Operation based Smoothness Measure

We propose a general framework for constructing smoothness measures based on smoothing operations. The central observation underlying this framework is that a smooth signal changes less under the application of a smoothing operator than a non-smooth signal.

Let I denote an image (or perturbation), and let C_α be a smoothing operator parameterized by a non-negative smoothing strength α, where larger values of α correspond to stronger smoothing. We denote by $C_\alpha(I)$ the result of applying the smoothing operator C_α to I. We define the C-smoothness T_C of an image I as a negative exponentially weighted average of the deviation between I and its smoothed version $C_\alpha(I)$:

$$T_C(I) = -\int_{\alpha \in \mathbb{R}^+} \exp(-\alpha)\ |C_\alpha(I) - I|\ d\alpha \tag{9}$$

The negative sign ensures that higher values of T_C correspond to smoother signals, in line with the sparsity measures introduced in Section III-C. This formulation relies on two properties commonly satisfied by smoothing operators.

First, convergence to the mean:

$$C_\alpha(I) \xrightarrow[\alpha \to \infty]{} \text{mean}(I) \tag{10}$$

Second, monotonicity of the deviation with respect to the smoothing strength:

$$|C_{\alpha_1}(I) - I| \geq |C_{\alpha_2}(I) - I| \quad \Leftrightarrow \quad \alpha_1 \geq \alpha_2 \tag{11}$$

The exponential weighting prevents large deviations at high smoothing strengths from dominating the measure, while still accounting for the full range of smoothing scales.

In this work, we instantiate this framework using two smoothing operators:

- Gaussian smoothing, where $\alpha = \sigma$ is the standard deviation of the Gaussian kernel, and
- Low-pass filtering, where α determines the cutoff frequency.

These choices allow us to capture smoothness both in the spatial domain and in the frequency domain.

B. Taylor Approximation based Smoothness Measure

We propose a smoothness measure based on first-order Taylor approximations. This measure is motivated by the observation that smooth signals can be locally well approximated by their first-order Taylor expansion. Let I be a (possibly multi-dimensional) image function, let a be a reference point, and let $\langle \cdot, \cdot \rangle$ be the inner product. The first-order Taylor approximation I_a^T of I at a is given by:

$$I_a^T(x) = I(a) + \langle \nabla I(a), (x - a) \rangle \tag{12}$$

For discrete images, numerical gradient approximations can be used to compute $\nabla I(a)$. To evaluate how well this local approximation represents the image, we consider a neighborhood $\mathcal{X}(x)$ for each pixel x, such as the 4-neighborhood or 8-neighborhood. We then define an approximate image as:

$$I_{\text{approx}}(x) = \text{mean}_{a \in \mathcal{X}} I_a^T(x) \tag{13}$$

Since both numerical gradient approximations and neighborhood averaging can be expressed as convolution operations, I_{approx} can be efficiently computed using a single convolution with a precomputed kernel. If the first-order Taylor expansion provides a good local approximation, the difference $I_{\text{approx}} - I$ should be close to zero. We therefore define the Taylor-based smoothness measure with a distance measure D as:

$$T_\nabla(I) = -D(I_{\text{approx}} - I) \tag{14}$$

Depending on the desired properties, any distance function D can be applied here. In this work, we choose the ℓ_2 norm for D, but other norms or even the sparsity measures introduced in Section III-C are possible choices. As before, the negative sign ensures that larger values correspond to smoother signals.

V. Experimental Setup

This section describes the experimental setup used in our study. For reproducibility, the source code can be found at https://zivgitlab.uni-muenster.de/ag-pria/p-norms.

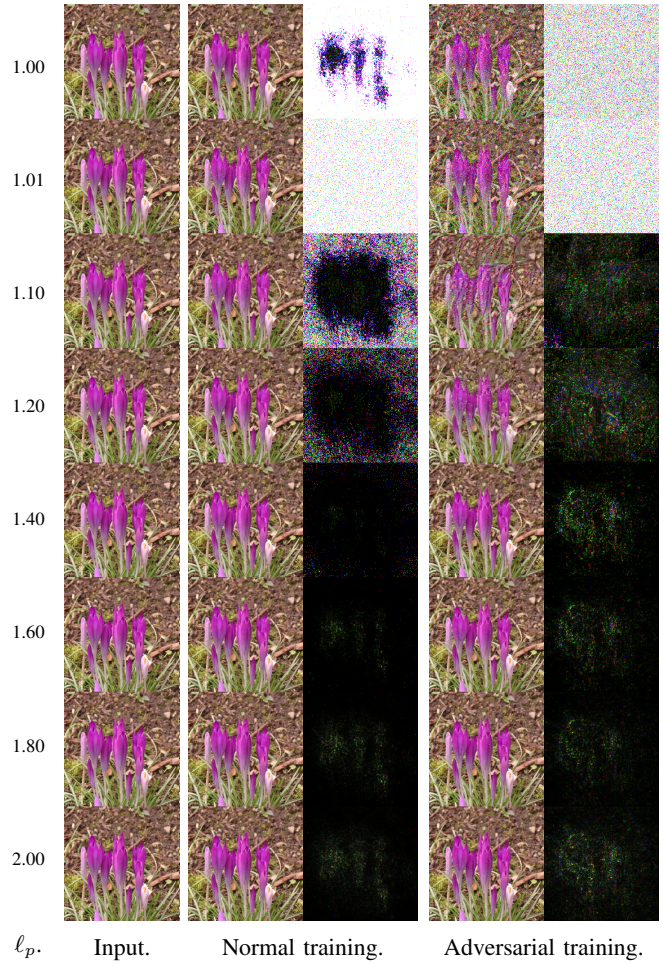
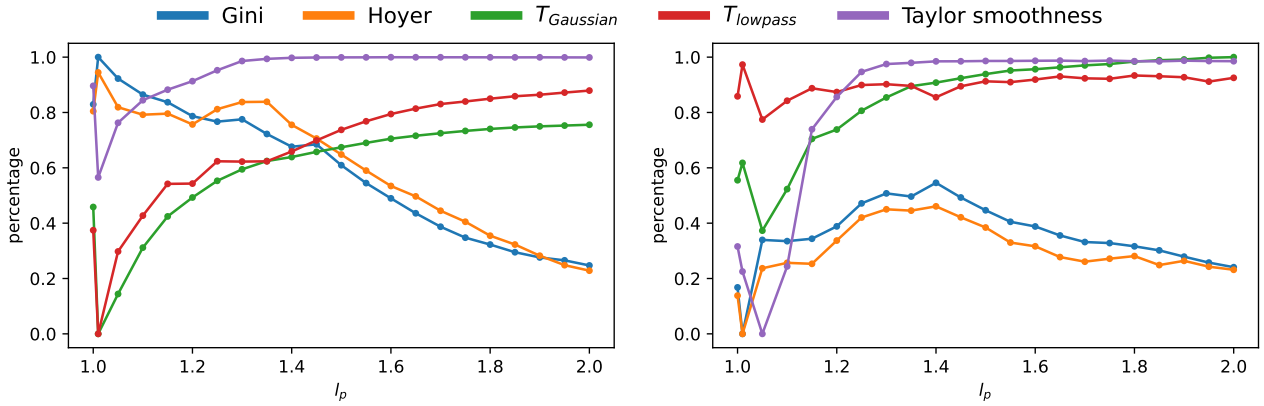


Fig. 1: Adversarially examples images for ResNet-18 and Flowers102. *Columns (left to right):* ℓ_p norm constraint, original image, adversarial image, and perturbation δ. *Rows (top to bottom):* increasing values of $p \in \{1.00, 1.01, 1.10, 1.20, 1.40, 1.60, 1.80, 2.00\}$. White pixels in δ denotes zero-valued entries.

A. Models and Datasets

We evaluate our approach on a diverse set of convolutional and transformer-based image classification models. As convolutional architectures, we use ResNet [40] variants ResNet-18, ResNet-50 and ResNet-101, VGG [41] variants VGG-16 and VGG-19. As transformer-based architectures, we use ViT-B/16 and ViT-B/32 from ViT [42] and Swin-T v2 and Swin-S v2 from the Swin Transformer [43]. The latter are comparable in computational complexity to ResNet-50 and ResNet-101, respectively, allowing for a meaningful comparison between convolutional and transformer-based models.

All models are evaluated on RGB image classification datasets with varying image resolutions and data distributions: CIFAR-10, CIFAR-100 [44], Flowers102 [45] and German Traffic Sign Recognition Benchmark (GTSRB) [46].



(a) Attacks against normally trained models.

(b) Attacks against adversarially trained models.

Fig. 2: Mean sparsity and smoothness as a function of p averaged over all models and datasets. The curves are normalized independently per measure, model and dataset to the interval $[0, 1]$ to enable comparison across measures. Sparsity measures are defined in Section III-C and smoothness measures in Section IV.

B. Training Protocol

All models are pretrained on ImageNet-1k [47] and subsequently fine-tuned on the target datasets. Convolutional models are trained for 50 epochs, while transformer-based models are trained for 20 epochs. We use the cross-entropy loss and the Adam optimizer, with learning rates of 10^{-3} for convolutional models and 5×10^{-5} for transformer-based models. For adversarial training, 75% of each batch consists of adversarial examples. For each adversarial sample, the value of p is sampled uniformly from the interval $p \in [1, 2]$. Adversarial training on the pretrained models was performed for 20 epochs. All experiments are performed on a server with AMD EPYC Milan 7543P CPU, 256 GB RAM and 2x NVIDIA A40 48 GB GPU. In total ∼2,000 GPU hours are used.

For the adversarial attacks, we use $p \in [1, 2]$ with step size 0.05. As ℓ_1-APGD has a closed-form solution and AFW is an iterative algorithm, we also use $p = 1.01$ to approximate the ℓ_1 norm with AFW, since it is only defined for $p > 1$ and even smaller p result in numerical instability.

C. Choice of Attack Magnitude ε

The choice of the attack magnitude ε is often made arbitrarily in the literature and typically justified through qualitative inspection of adversarial examples. This makes quantitative comparisons across models, datasets, and norm choices difficult. We value a consistent choice of ε over visually good looking adversarial attacks to decrease variability and increase comparability between ℓ_p norms, models and datasets.

To ensure comparability, we adopt a performance-based strategy to determine ε for each model, dataset, and norm and denote it by ε_p. We do this as follows:

1) Test the performance of each model on each dataset.
2) For every combination of model & dataset, set the target accuracy as $\frac{1}{3}$ of the test accuracy.
3) For each model & dataset choose for every ℓ_p norm the ε_p, so that the test accuracy degrades to that target accuracy (but not further).

This procedure ensures that all attacks achieve comparable effectiveness, allowing us to isolate the influence of the norm choice on sparsity and smoothness.

D. Calculating Optimal p

Our goal is to identify values of p for which adversarial attacks simultaneously maximize sparsity and smoothness. Let $\mathcal{M}_S$ denote the set of all sparsity measures, $\mathcal{M}_T$ the set of all smoothness measures, and $\mathcal{M} = \mathcal{M}_S \cup \mathcal{M}_T$ their union and thus the set of all measures we are interested in. Each measure $m \in \mathcal{M}$ is normalized to the interval $[0, 1]$. We first compute the largest value $\beta_{\text{opt}} \geq 0$ such that there exists at least one value of p for which all measures exceed β_{opt}:

$$\beta_{\text{opt}} = \max\{\beta \geq 0 \mid \bigcap_{m \in \mathcal{M}} \{p : m(p) \geq \beta\} \neq \emptyset\} \tag{15}$$

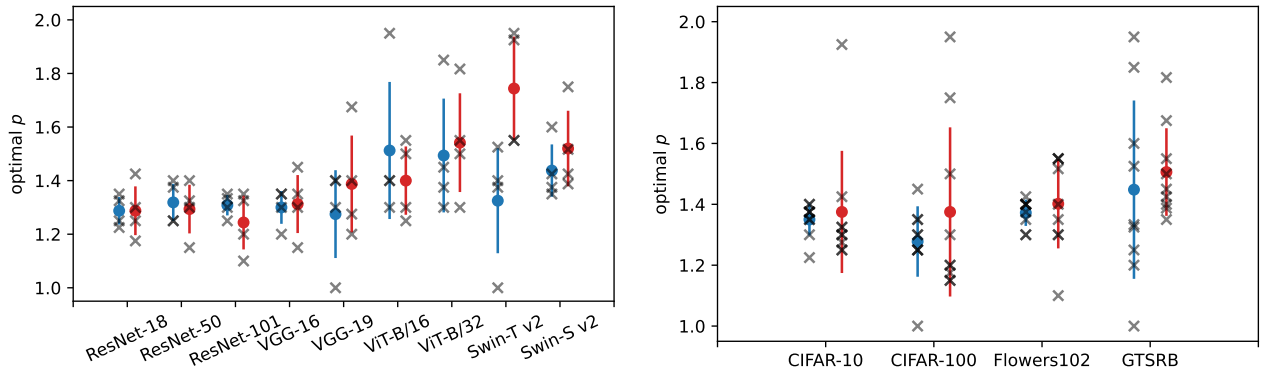
Using β_{opt}, we define the set of optimal norm parameters as:

$$\mathcal{M}_{p_{\text{opt}}} = \bigcap_{m \in \mathcal{M}} \{p : m(p) \geq \beta_{\text{opt}}\} \tag{16}$$

This construction guarantees that $\mathcal{M}_{\text{opt}}$ is non-empty and yields either a single optimal value or a set of equally optimal values of p. At the same time, β_{opt} provides a quantitative measure of how sparse and smooth the attacks are for the given model and dataset combination.

E. α Schedules for Smoothing Operation based Smoothness Measures

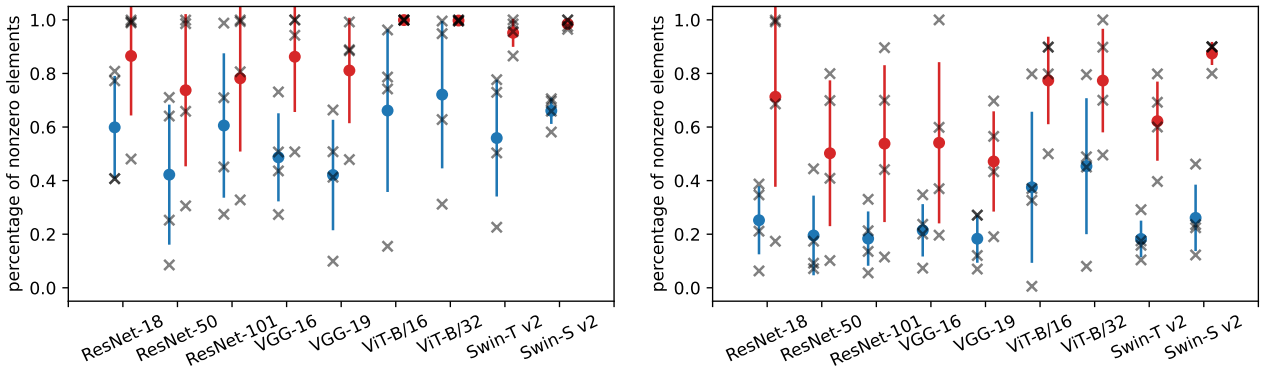
For the smoothness operation based smoothness measures T_{C_α} we use the following α schedules. For Gaussian smoothing we set $\sigma_i^2 = \alpha_i^2$ and linearly sample $\alpha_i \in [1, 10]$ with step size 1. For the low pass filter, we use the same $\alpha_i \in [1, 10]$ and set the cutoff frequencies c_i as $c_i = (10 - \alpha_i + 1)^2$ linearly scaled to half the image size. This ensures big cutoff steps for small α_i meaning high frequencies with lower impact on



(a) Optimal p per model aggregated over datasets.

(b) Optimal p per dataset aggregated over models.

Fig. 3: Optimal values of p maximizing sparsity and smoothness. Dots indicate the mean optimal value and error bars the standard deviation; individual runs are shown as crosses. Blue denotes attacks against normally trained models and red denotes attacks against adversarially trained models.



(a) Attacks generated using ℓ_1-APGD.

(b) Attacks generated using $\ell_{1.01}$-AFW.

Fig. 4: ℓ_0 sparsity of adversarial perturbations δ normalized by image size, shown as percentage of non-zero pixels. Mean and std given by dot and error bar, individual data points are represented as crosses. Blue denotes attacks against normally trained models and red denotes attacks against adversarially trained models.

smoothing get discarded quickly. Also, the smaller steps for higher α_i ensure a good sampling of the stronger smoothed images with high impact α_i.

VI. Results

In this section, we present an extensive empirical evaluation of the influence of the choice of the ℓ_p norm on the sparsity and smoothness of adversarial attacks. Unless stated otherwise, all reported results are averaged over multiple models and datasets.

A. Sparsity and Smoothness as a Function of p

We now analyze how sparsity and smoothness evolve as a function of the norm parameter $p \in [1, 2]$. Across all models and datasets, we observe that smoothness increases with increasing p, while sparsity generally decreases as shown in Fig. 2. However, neither trend is linear, and sparsity in particular exhibits non-monotonic behavior. Most notably, both sparsity measures have a peak around $p = 1.3$ and for adversarial training even attain their maximum. This indicates that sparsity is not maximized at $p = 1$, despite the common association between the ℓ_1 norm and sparse solutions. Especially for low p and after adversarial training, all measures do not increase/decrease monotonously. It stands out that between $p = 1$ and $p = 1.01$ the behavior of both sparsity and smoothness measures is strongly influenced by the attack algorithm changing from ℓ_1-APGD to AFW as the attack algorithm. Although both are based on APGD, the closed-form solution of ℓ_1-APGD leads to noticeably less sparse and smoother perturbations than the AFW-generated approximations. To further analyze this phenomenon, we conducted a separate study on these two versions (Section VI-C). The measures on the AFW-generated adversarial attacks converge more evenly for $p \to 1$ after the normal training than after adversarial training.

B. Optimal p across Models and Datasets

When aggregating optimal values of p across datasets for each model (see Fig. 3a), convolutional architectures exhibit a tight clustering of optimal values around $p = 1.3$ This behavior

is consistent across both normally trained and adversarially trained models. In contrast, transformer-based architectures show greater variability. Their optimal values of p are typically higher, often in the range $p \in [1.4, 1.5]$ with larger standard deviations. An exception is the Swin-T v2 model which has an average optimal p of 1.3 for normal training and 1.7 for adversarial training.

When aggregating across models for each dataset in Fig. 3b, we observe lower optimal p and standard deviation for the normal training than after adversarial training. Only the GTSRB has a higher standard deviation for normal training. The outliers with high p are from ViT-B/16 and ViT-B/32.

The transformer-based models, especially the ViT, are inherently more robust against adversarial attacks with small p. Aside from the higher optimal p, Fig. 4 shows that the adversarial attacks for p near 1 do use a high percentage of all available pixels. For the adversarial trained models, almost all of the image has to be attacked.

C. Comparison of ℓ_1-APGD and $\ell_{1.01}$-AFW

Since Fig. 2 shows that adversarial attacks generated with ℓ_1-APGD are more sparse than those generated with AFW using $p = 1.01$, we further analyze the sparsity behavior of these two attack methods using the stricter ℓ_0 measure across all evaluated models. As shown in Fig. 4, ℓ_1-APGD produces less sparse adversarial attacks than AFW for most models, regardless of whether the models are normally trained or adversarially trained. Note that for ℓ_0 low scores mean a more sparse result. This effect is consistent across datasets and architectures. In addition, Fig. 4 shows that adversarial training generally leads to less sparse attacks for both methods, indicating that adversarial training increases robustness particularly against sparse perturbations.

Qualitative inspection of the adversarial examples in Fig. 1 further supports these findings. Before adversarial training, ℓ_1-APGD generates perturbations that are more concentrated on salient object regions than those generated by AFW, even for values of p close to 1. However, this distinction largely vanishes after adversarial training.

Overall, these results indicate that the observed differences in sparsity between ℓ_1-APGD and AFW are not solely due to the choice of the norm parameter p, but are also influenced by the specific optimization algorithm used to generate the adversarial attacks.

D. Discussion

As expected, sparsity increases and smoothness decreases for $p \to 1$. Notably however, this behavior is neither linear nor monotonic over the interval $p \in [1, 2]$. In particular, sparsity does not continuously increase toward $p = 1$, but instead reaches a maximum around $p = 1.3$. The main reason for this is the different behaviors of ℓ_1-APGD compared to the approximative $\ell_1.01$-AFW, as further analyzed in Section VI-C. As for smoothness measures, all evaluated smoothness measures indicate that smoothness plateaus at around $p \approx 1.4$. Increasing p further does not lead to smoother attacks, but does lead to less sparse attacks. In particular, the interval $p \in [1.4, 1.8]$ shows the steepest decrease in sparsity. This behavior indicates the existence of a 'sweet spot' for the choice of p within the interval $[1.3, 1.5]$, which is consistent with the optimal values observed in Fig. 3.

Fig. 3 further shows that adversarial training leads to higher variance and generally larger optimal values of p, particularly for transformer-based models. In addition, Fig. 4 indicates that adversarial attacks with values of p close to 1 require modifying a large fraction of the available pixels after adversarial training. These observations suggest that adversarial training is particularly effective at improving robustness against sparse adversarial attacks.

VII. Conclusion

In this work, we investigated the effect of the choice of the ℓ_p norm on the properties of adversarial attacks. In particular, we analyzed how sparsity and smoothness behave when adversarial attacks are constructed under ℓ_p norm restrictions for $p \in [1, 2]$. To quantify sparsity, we employed two established sparsity measures that are well suited for fixed-size inputs such as images. To quantify smoothness, we introduced a general framework for deriving smoothness measures based on smoothing operations and proposed two instantiations of this framework. In addition, we introduced a smoothness measure based on first-order Taylor approximations.

Using these measures, we conducted a comprehensive empirical evaluation across multiple datasets and a diverse set of model architectures, including both convolutional and transformer-based networks. Our results show that both sparsity and smoothness are strongly influenced by the choice of the norm parameter p. In particular, we find that the commonly used choices $p = 1$ and $p = 2$ are generally suboptimal when sparsity and smoothness are considered jointly. Instead, intermediate values of p consistently yield adversarial perturbations with more favorable trade-offs between these two properties.

Moreover, we show that the optimal choice of p depends on both the model architecture and the data distribution. This dependency is especially pronounced for transformer-based models, which exhibit higher optimal values of p and greater variability compared to convolutional architectures which tend to agree on an optimal value of $p \approx 1.3$.

Finally, adversarial training substantially reduces the effectiveness of sparse attacks, forcing successful perturbations to modify a larger fraction of the input.

Overall, our findings show that the commonly used practice of selecting the ℓ_p norm without further justification is suboptimal. Careful selection of the norm parameter p enables better control over sparsity and smoothness and leads to more informative and comparable adversarial evaluations. We believe that these insights provide a useful foundation for future research on perceptually grounded adversarial attacks.

Acknowledgment

ChatGPT 5.2 was used exclusively for sentence-level reformulation, but not for text generation.

REFERENCES

[1] Q. Zhou, M. Zuley, Y. Guo, L. Yang, B. Nair, A. Vargo, S. Ghannam, D. Arefan, and S. Wu, "A machine and human reader study on AI diagnosis model safety under attacks of adversarial images," *Nature Communications*, vol. 12, no. 1, p. 7281, 2021.

[2] J. Zhang, Y. Lou, J. Wang, K. Wu, K. Lu, and X. Jia, "Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles," *IEEE Internet Things Journal*, vol. 9, no. 5, pp. 3443–3456, 2022.

[3] W. Heng, S. Zhou, and T. Jiang, "Harmonic adversarial attack method," *CoRR*, vol. abs/1807.10590, 2018.

[4] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4724–4732.

[5] H. Zhang, Y. Avrithis, T. Furon, and L. Amsaleg, "Smooth adversarial examples," *EURASIP Journal on Information Security*, vol. 2020, no. 1, pp. 1–12, 2020.

[6] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[7] S. Sadiku, M. Wagner, and S. Pokutta, "Group-wise sparse and explainable adversarial attacks," *CoRR*, vol. abs/2311.17434, 2023.

[8] C. Guo, J. S. Frank, and K. Q. Weinberger, "Low frequency adversarial perturbation," in *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, 2019, pp. 1127–1137.

[9] R. Liu, W. Zhou, S. Wu, J. Zhao, and K. Lam, "SSTA: salient spatially transformed attack," *CoRR*, vol. abs/2312.07258, 2023.

[10] J. Liu, B. Lu, M. Xiong, T. Zhang, and H. Xiong, "Low frequency sparse adversarial attack," *Computers & Security*, vol. 132, p. 103379, 2023.

[11] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations*, 2014.

[13] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Survey on adversarial attack and defense for medical image analysis: Methods and challenges," *ACM Computing Surveys*, vol. 57, no. 3, pp. 79:1–79:38, 2025.

[14] C. Zhang, L. Zhou, X. Xu, J. Wu, and Z. Liu, "Adversarial attacks of vision tasks in the past 10 years: A survey," *ACM Computing Surveys*, vol. 58, no. 2, pp. 52:1–52:42, 2026.

[15] H. Wang, C.-C. Chang, C.-S. Lu, C. Leckie, and I. Echizen, "Greedypixel: Fine-grained black-box adversarial attack via greedy algorithm," *arXiv preprint arXiv:2501.14230*, 2025.

[16] Z. He, W. Wang, J. Dong, and T. Tan, "Transferable sparse adversarial attack," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 963–14 972.

[17] H. Xue, A. Araujo, B. Hu, and Y. Chen, "Diffusion-based adversarial sample generation for improved stealthiness and controllability," *CoRR*, vol. abs/2305.16494, 2023.

[18] X. Chen, X. Gao, J. Zhao, K. Ye, and C. Xu, "Advdiffuser: Natural adversarial example synthesis with diffusion models," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4539–4549.

[19] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence Safety and Security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[20] H. Ren, T. Huang, and H. Yan, "Adversarial examples: attacks and defenses in the physical world," *International Journal of Machine Learning and Cybernetics*, pp. 1–12, 2021.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[22] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of 37th International Conference on Machine Learning*, 2020, pp. 2206–2216.

[23] ——, "Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers," in *International Conference on Machine Learning*, 2021, pp. 2201–2211.

[24] M. Frank, P. Wolfe *et al.*, "An algorithm for quadratic programming," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[25] J. Chen, D. Zhou, J. Yi, and Q. Gu, "A Frank-Wolfe framework for efficient and effective adversarial attacks," in *Proceedings of AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3486–3494.

[26] V. Boreiko, M. Augustin, F. Croce, P. Berens, and M. Hein, "Sparse visual counterfactual explanations in image space," in *DAGM German Conference on Pattern Recognition*. Springer, 2022, pp. 133–148.

[27] A. E. Cinà, F. Villani, M. Pintor, L. Schönherr, B. Biggio, and M. Pelillo, "σ-zero: Gradient-based optimization of ℓ_0-norm adversarial examples," in *The Thirteenth International Conference on Learning Representations*, 2025.

[28] A. Sen, X. Zhu, L. Marshall, and R. Nowak, "Should adversarial attacks use pixel p-norm?" *arXiv preprint arXiv:1906.02439*, 2019.

[29] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual adversarial robustness: Defense against unseen threat models," in *International Conference on Learning Representations*, 2020.

[30] C. Xiao, J. Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[31] A. A. N. Nafi, H. Rahaman, Z. Haider, T. Mahfuz, F. Suya, S. Bhunia, and P. Chakraborty, "Dash: A meta-attack framework for synthesizing effective and stealthy adversarial examples," *arXiv preprint arXiv:2508.13309*, 2025.

[32] Y. Fan, B. Wu, T. Li, Y. Zhang, M. Li, Z. Li, and Y. Yang, "Sparse adversarial attack via perturbation factorization," in *16th European Conference on Computer Vision*. Springer, 2020, pp. 35–50.

[33] X. Dong, D. Chen, J. Bao, C. Qin, L. Yuan, W. Zhang, N. Yu, and D. Chen, "GreedyFool: Distortion-aware sparse adversarial attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 226–11 236, 2020.

[34] F. Lin, J. Lou, H. Wang, B. Jalaian, and X. Yuan, "Towards interpretable adversarial examples via sparse adversarial attack," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2025, pp. 92–110.

[35] A. Heshmati, S. S. Roudi, S. Amini, S. Ghaemmaghami, and F. Marvasti, "A versatile framework for designing group-sparse adversarial attacks," *arXiv preprint arXiv:2510.16637*, 2025.

[36] H. Dalton, "The measurement of the inequality of incomes," *The Economic Journal*, vol. 30, no. 119, pp. 348–361, 1920.

[37] S. Rickard and M. Fallon, "The gini index of speech," in *Proceedings of the 38th Conference on Information Science and Systems (CISS'04)*, 2004.

[38] C. Gini, "Measurement of inequality of incomes," *Economic Journal*, vol. 31, no. 121, pp. 124–126, 1921.

[39] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, no. 9, pp. 1457–1469, 2004.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations*, 2021.

[43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[44] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[45] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *6th Indian conference on Computer Vision, Graphics & Image Orocessing*. IEEE, 2008, pp. 722–729.

[46] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark," in *International Joint Conference on Neural Networks*, 2013, pp. 1–8.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.