Adaptive Norm Selection Prevents Catastrophic Overfitting in Fast Adversarial Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a novel solution to Catastrophic Overfitting (CO) in fast adversarial training based solely on adaptive l^p norm selection. Unlike existing methods requiring noise injection, regularization, or gradient clipping, our approach dynamically adjusts training norms based on gradient concentration, preventing the vulnerability to multi-step attacks that plagues single-step methods.

We begin with the empirical observation that, with small perturbations, CO occurs predominantly under l^∞ rather than l^2 norms. Building on this observation, we formulate generalized l^p attacks as a fixed-point problem and develop l^p-FGSM to analyze the l^2-to-l^∞ transition. Our key discovery: CO arises when concentrated gradients—with information localized in few dimensions—meet aggressive norm constraints.

We quantify gradient concentration via Participation Ratio from quantum mechanics and entropy metrics, yielding an adaptive l^p-FGSM that dynamically adjusts the training norm based on gradient structure. Experiments show our method achieves robust performance without auxiliary regularization or noise injection, offering a principled solution to the CO problem.

1 Introduction

Deep neural networks have achieved remarkable success across computer vision, NLP, and speech recognition [1, 2, 3], yet remain vulnerable to adversarial perturbations—subtle input modifications that cause misclassifications [4, 5]. This vulnerability poses important challenges in safety-critical applications including autonomous vehicles [6], healthcare [7], and financial systems [8].

Among defense strategies, adversarial training—incorporating adversarially perturbed examples during training—has proven most effective [5, 9]. However, multi-step methods like Projected Gradient Descent (PGD) [9] impose significant computational costs that limit their applicability in large-scale settings. Fast single-step methods address this efficiency concern but suffer from Catastrophic Overfitting (CO), where models maintain single-step robustness while failing against multi-step attacks [10].

Several approaches have been developed to address CO. RS-FGSM [10] adds uniform random perturbations within the ϵ-ball before applying FGSM, though effectiveness diminishes with larger perturbation radii. GradAlign [11] enforces local linearity by aligning input gradients at clean and adversarial points through double backpropagation, improving robustness but doubling computational overhead. ZeroGrad [12] zeros out small gradient components below a dynamic threshold, preventing overfitting to low-magnitude noise directions with minimal extra cost. N-FGSM [13] removes gradient clipping and uses stronger noise, achieving 3× speedup over GradAlign while maintaining comparable robustness.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Recent work has explored CO from various perspectives. AAER [14] identifies "abnormal adversarial examples" where loss decreases during inner maximization and regularizes their occurrence. LAP [15] reveals that pseudo-robust shortcuts form in early network layers, applying adaptive weight perturbations that decrease from former to latter layers. SKG-FAT [16] addresses class imbalance through differentiated class weights and self-knowledge guided label relaxation, achieving 5× speedup over PGD-10. ELLE [17] approximates local linearity regularization without expensive double backpropagation, adapting regularization strength during training. FGSM-PCO [18] prevents inner optimization collapse by generating adversarial examples through adaptive fusion of current and historical perturbations.

While these methods have made important contributions, they typically require auxiliary techniques such as noise injection, regularization, double backpropagation, or architectural modifications. This observation motivates our investigation into whether CO can be addressed through more direct mechanisms.

Our work begins with an empirical observation: CO exhibits interesting norm-dependent behavior. For comparable perturbation amplitudes, l^∞-norm training shows pronounced CO while l^2-defense remains more stable, though with limited cross-norm robustness (Figure 1). This suggests that the choice of norm constraint may play a more fundamental role in CO than previously recognized.

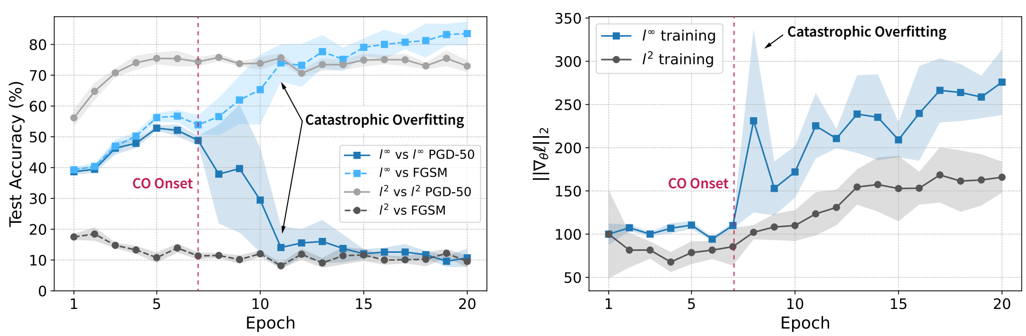


Figure 1: CO phenomena on CIFAR-10 [19] using WideResNet-28-10 [20]: **Left:** l^∞ training ($\epsilon = 8/255$) shows accuracy collapse against PGD-50 [9], while l^2 ($\epsilon = 32/255$) remains stable. **Right:** CO onset correlates with gradient norm increase in l^∞ training only.

Building on this observation, we move beyond traditional linear approximations underlying FGSM and adopt a local convexity hypothesis. This leads us to reformulate adversarial attack generation as a fixed-point problem, naturally yielding the l^p-FGSM family of attacks. Initial exploration reveals that higher p values ($p \geq 32$) delay but do not prevent CO, while lower values avoid CO at the cost of reduced robustness (Figure 2).

To understand this trade-off, we investigate gradient concentration as a potential mechanism underlying CO. We quantify this through the Participation Ratio (PR) [21, 22]—a measure from quantum mechanics that we adapt to adversarial training as $\mathtt{PR}_1$. Much like its predecessor PR, the adapted metric $\mathtt{PR}_1$ captures how many dimensions meaningfully contribute to gradient magnitude and most importantly connect naturally to the angular separation between l^2 and l^∞ bounded perturbations.

Our key insight is that catastrophic overfitting emerges precisely when concentrated gradients—with information localized in few dimensions—meet aggressive norm constraints. This concentration can be quantified through participation ratio metrics (detailed in Appendix M), allowing us to adaptively select norm constraints that prevent CO without sacrificing robustness. Based on this understanding, we develop adaptive l^p-FGSM that dynamically adjusts the training norm p based on gradient structure. When gradients concentrate (low PR), the method reduces p to maintain better alignment with natural l^2 geometry; when gradients distribute more uniformly, higher p values can enhance robustness.

This approach achieves competitive performance on standard benchmarks without requiring noise injection, regularization, or architectural changes. Unlike previous approaches that focus on loss landscapes or gradient alignment, our method directly addresses the gradient concentration phenomenon that precipitates catastrophic overfitting. By providing this connection between gradient geometry and CO, our work offers a complementary perspective suggesting that careful norm selection alone can serve as an effective tool for improving single-step adversarial training.

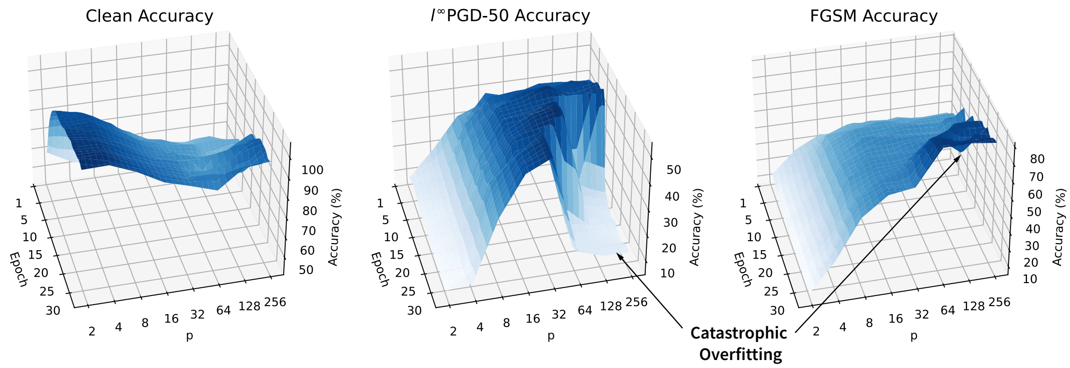


Figure 2: Impact of l^p norm choice on training dynamics and robustness for CIFAR-10 with WideResNet-28-10. The choice of p reveals a key trade-off: higher values ($p \geq 32$) initially show better robustness but become vulnerable to Catastrophic Overfitting (CO), evident in the l^∞ PGD-50 plot (second left). Lower p values prevent CO but with reduced adversarial robustness. Results shown for $\epsilon = 8/255$ over 30 epochs.

2 Preliminaries

We consider a classification function $c(x;\theta) : x \mapsto \mathbb{R}^C$ that maps input features x to output logits for classes in set C. The prediction probability $\pi_i(x;\theta)$ for label i is given by the softmax function: $\pi_i(x;\theta) = \exp(c_i(x;\theta))/\sum_j \exp(c_j(x;\theta))$, where $c_i(x;\theta)$ denotes the i-th logit and θ represents model parameters [23].

Adversarial robustness requires that the predicted class remains unchanged under bounded perturbations. Function c is robust to adversarial perturbations of magnitude ϵ at input x if the class with maximum probability for x retains the highest probability for $x + \delta$, where δ is any perturbation within the l^p ball of radius ϵ [4, 5]:

$$\operatorname*{argmax}_{i \in C} \pi_i(x+\delta;\theta) = \operatorname*{argmax}_{i \in C} \pi_i(x;\theta),\ \forall \delta \in B_p(\epsilon) \tag{1}$$

This work considers general l^p norms with $p \geq 2$, using $B(\epsilon)$ to denote $B_p(\epsilon)$ for simplicity.

Standard training employs Empirical Risk Minimization (ERM) [24] over dataset distribution $\mathcal{D}$:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x;y,\theta)] \tag{2}$$

where ℓ represents the loss function, typically cross-entropy $\ell(x;y,\theta) = -y^T \log(\pi(x;\theta))$, and y is the one-hot encoded label. While ERM achieves satisfactory performance on clean data, networks remain vulnerable to adversarial attacks [4, 5], with test accuracy dropping substantially under distributional shifts caused by adversarial perturbations.

Adversarial training [5, 9] addresses this vulnerability by incorporating adversarial examples during training, simulating potential distributional shifts to learn features robust to input perturbations:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta \in B(\epsilon)} \ell(x+\delta;y,\theta)\right] \tag{3}$$

The inner maximization $\max_{\delta\in B(\epsilon)} \ell(x+\delta;y,\theta)$ is typically approximated through gradient-based optimization. Projected Gradient Descent (PGD) [9] performs iterative updates:

$$\delta \leftarrow \Pi\left(\delta - \mu\nabla_x \ell(x+\delta;y,\theta)\right) \tag{4}$$

where projection operator Π ensures perturbations remain within bounds through scaling (l^2) or clipping (l^∞).

Multi-step methods like PGD incur significant computational costs. The Fast Gradient Sign Method (FGSM) [5] provides efficiency through first-order Taylor expansion $\ell(x_0+\delta) \approx \ell(x_0) + \delta^T \nabla_x \ell$, using gradient sign to solve the maximization problem:

$$\delta_{\text{FGSM}} = \underset{\delta \in B_\infty(\epsilon)}{\operatorname{argmax}} \left(\ell(x_0) + \delta^T \nabla_x \ell\right) = \epsilon \operatorname{sign}\left(\nabla_x \ell\right) \tag{5}$$

While FGSM efficiently solves the linearized maximization problem in Eq. (3) under l^∞ constraints, it suffers from Catastrophic Overfitting (CO). Wong et al. [10] proposed adding random noise $\eta \sim \mathcal{U}[-\epsilon, \epsilon]$ as remedy:

$$\delta_{\text{RS-FGSM}} = \Pi_{B_\infty(\epsilon)}\left(\eta + \epsilon \operatorname{sign}\left(\nabla_x \ell\left(x_0 + \eta\right)\right)\right) \tag{6}$$

Our work extends beyond first-order approximations by characterizing the inner maximization in Eq. (3) under general l^p constraints, leading to a fixed-point formulation.

3 Theoretical Framework

We develop a theoretical foundation that moves beyond the local linearity assumption underlying FGSM by adopting a local convexity framework. This perspective reveals that optimal perturbations reside on constraint boundaries, enabling our fixed-point formulation for general l^p norms and providing the mathematical foundation for preventing catastrophic overfitting through principled norm selection.[1] [2]

Under local convexity, optimal adversarial perturbations are guaranteed to lie on the boundary $\partial B_p(\epsilon)$, as any interior critical point must be a local minimum when the Hessian $\nabla_x^2 \ell$ is positive definite. We demonstrate that this condition emerges naturally during training through Hessian analysis and empirical validation (detailed in Appendix A). This enables controlled transitions between the catastrophic overfitting-resistant l^2 regime and the catastrophic overfitting-prone l^∞ regime.

3.1 l^2 Norm-Bounded Adversarial Attacks

Given that optimal perturbations exist on the boundary under local convexity, we use Lagrange multipliers to reformulate the constrained maximization problem in Eq. (3) as an unconstrained optimization, leading to a fixed-point characterization.

Proposition 1. *For a training sample x_0 with non-null gradient, the optimal perturbation $\delta^\star$ within $B_2(\epsilon)$ exists and solves the fixed-point problem $\delta^\star = F(\delta^\star)$, where:*

$$F(\delta) = \epsilon \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_2} \tag{7}$$

F is Lipschitz continuous around its origin with constant $K = 2\epsilon \|\nabla_x^2 \ell\| / \|\nabla_x \ell(x_0)\|_2$:

$$\|F(\delta) - F(0)\| \leq K \|\delta\| \tag{8}$$

and the fixed-point problem converges if $K < 1$.

Proof. See Appendix B. □

Equation (7) defines a fixed-point iteration that approximates the optimal perturbation, as illustrated in Figure 3. The Lipschitz constant K connects to curvature control techniques: CURE [25] minimizes Hessian norms for robustness, while Srinivas et al. [26] introduced gradient norm division for scale-invariant curvature. Reducing K accelerates convergence of the inner maximization in Eq. (3).

Corollary (GradAlign Connection). *When $\nabla_x \ell(x_0)$ aligns with $\nabla_x \ell(x_0 + \epsilon \nabla_x \ell / \|\nabla_x \ell\|)$, the fixed-point converges instantly*[3]:

$$\frac{\nabla_x \ell(x_0 + \epsilon \nabla_x \ell / \|\nabla_x \ell\|)}{\|\nabla_x \ell(x_0 + \epsilon \nabla_x \ell / \|\nabla_x \ell\|)\|} = \frac{\nabla_x \ell}{\|\nabla_x \ell\|} \tag{9}$$

GradAlign [11] regularizes gradient alignment, effectively improving the initialization of our fixed-point algorithm, explaining its empirical success.

[1] If local convexity does not hold, the framework gracefully defaults to the standard local linearity approach.
[2] For one-step adversarial training, local linearity and convexity lead to identical outcomes.
[3] In this ideal case, the normalized gradient is already the fixed point.

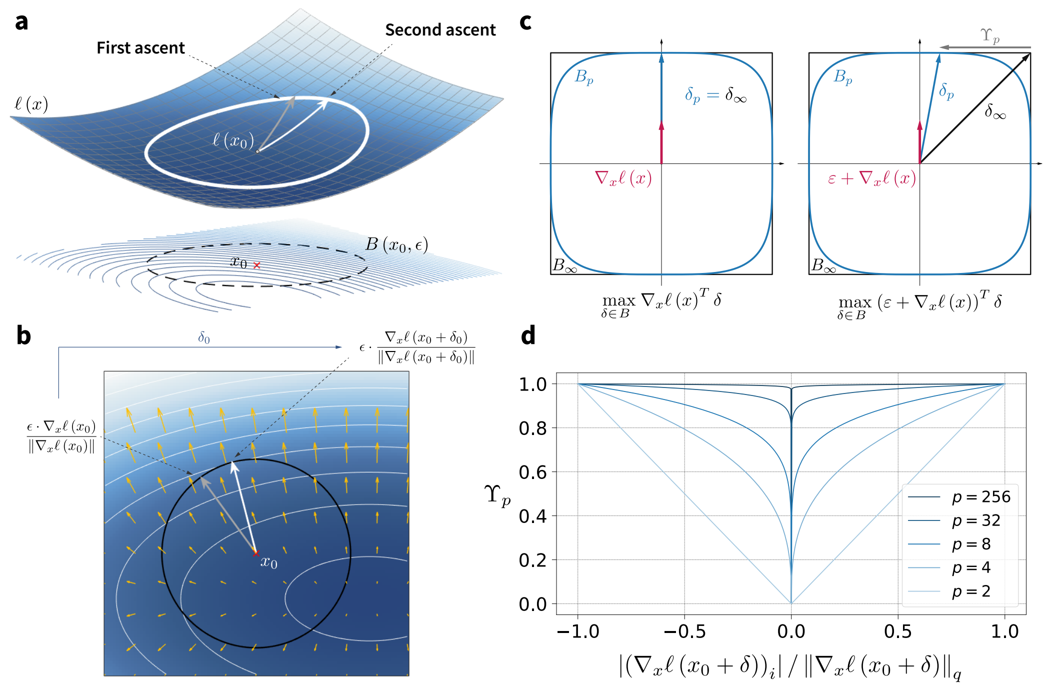


Figure 3: Geometric interpretation of l^p-FGSM framework. **(a,b)** Fixed-point algorithm iterations for optimal perturbation identification under l^2 constraint (Eq. 7). **(c)** Attack geometry under different l^p norms: *Left* - ideal scenario with aligned gradients; *Right* - effect of gradient noise showing l^∞ sensitivity versus l^p stability. **(d)** Transition function Υ_p variation across p values, demonstrating smooth high-pass filtering behavior.

3.2 l^p Norm-Bounded Adversarial Attacks

We extend the fixed-point framework to general l^p norms, which serve as smooth interpolations between l^2 and l^∞. This extension enables our approach to catastrophic overfitting through controlled norm transitions based on gradient structure.

Proposition 2. *For a training sample x_0 with non-null gradient under $B_p(\epsilon)$ constraint, the optimal perturbation $\delta^\star$ exists and solves the fixed-point equation $\delta^\star = F_p(\delta^\star)$, where:*

$$F_p(\delta) = \epsilon \, sign(\nabla_x \ell(x_0+\delta)) \left| \frac{\nabla_x \ell(x_0+\delta)}{\|\nabla_x \ell(x_0+\delta)\|_q} \right|^{q-1} \tag{10}$$

with l^q being the dual norm of l^p: $\frac{1}{p} + \frac{1}{q} = 1$. All operations are element-wise.

Proof. See Appendix C. □

Unified Attack Spectrum: Equation (10) provides a unified formulation spanning from l^2 to l^∞. For $p = q = 2$, we recover Eq. (7); as $p \to \infty$, we obtain $q = 1$ and recover FGSM. The transition between regimes is governed by:

$$\Upsilon_p(\delta) = \left| \frac{\nabla_x \ell(x_0+\delta)}{\|\nabla_x \ell(x_0+\delta)\|_q} \right|^{q-1} \tag{11}$$

This function acts as a smooth high-pass filter, approaching unity everywhere except near zero (Figure 3d). Unlike discontinuous thresholding in ZeroGrad [12], our approach provides smooth gradient filtering that preserves differentiability and training stability.

Convergence Analysis: For $p > 2$, global Lipschitz continuity fails due to the discontinuous sign function and concave power term $(q-1)$ near zero gradients. However, we ensure local Lipschitzness by maintaining gradients bounded away from zero:

$$\exists m > 0 : \forall i, \forall \delta \in \partial B_p(\epsilon), |\nabla_x \ell(x_0+\delta)_i| > m \tag{12}$$

This condition motivates our algorithmic design: adding constant ε to gradient components ensures both numerical stability and theoretical convergence guarantees. Under this modification, F_p becomes locally Lipschitz with constant $K(p, m)$ (detailed in Appendix D).

3.3 Gradient-Aware Adaptive Norm Selection

While fixed p values can balance robustness and stability, our preliminary analysis reveals fundamental limitations. As detailed in Appendix E, higher p values delay catastrophic overfitting but eventually succumb to it, while lower p values provide stability at the cost of reduced robustness. This fundamental trade-off varies significantly across datasets, with dataset complexity critically influencing optimal p selection, motivating our adaptive approach.

High-Dimensional Perturbation Analysis: The choice of norm becomes increasingly critical as dimensionality grows. In $\mathbb{R}^d$, perturbation amplitudes scale directly with dimension:[4]

$$\|\delta_2\|_2 = \epsilon, \;\; \|\delta_\infty\|_2 \overset{a.s.}{=} \epsilon\, d^{1/2}, \;\; \max \|\delta_p\|_2 = \epsilon\, d^{(1/2-1/p)} \tag{13}$$

These relationships, which appear in adversarial PAC-Bayes bounds [27], reveal that l^∞-bounded perturbations yield vectors dramatically distant from original samples as dimension increases. For CIFAR-10 ($d = 3,072$) and ImageNet ($d \sim 1.5 \times 10^5$), this effect becomes particularly significant.

Our key insight: reducing p effectively constrains the perturbation space from dimension d to an effective dimension d_e, where $d^{(1/2-1/p)} \sim d_e^{1/2}$. This suggests that measuring the intrinsic effective dimension of gradients can guide appropriate p selection.

Participation Ratio for Gradient Concentration: We adapt the Participation Ratio from quantum mechanics [21, 22], which quantifies electron localization, to measure gradient concentration:

$$\mathtt{PR}(x) = \frac{(\sum_i |x_i|^2)^2}{\sum_i |x_i|^4} = \left(\frac{\|x\|_2}{\|x\|_4}\right)^4 \tag{14}$$

For adversarial training, we substitute the standard ones vector with the gradient's sign vector, yielding:

$$\mathtt{PR}_1 = \left(\frac{\|\nabla_x \ell\|_1}{\|\nabla_x \ell\|_2}\right)^2 \tag{15}$$

This effective dimension varies between 1 and d for non-null vectors and naturally connects to the angular separation between δ_2 and δ_∞ attacks:

$$\cos(\theta_{2,\infty}) = \frac{\|\nabla_x \ell\|_1}{\|\nabla_x \ell\|_2 d^{1/2}} = \sqrt{\frac{\mathtt{PR}_1}{d}} \tag{16}$$

Figure 4 provides empirical validation of our theoretical framework. Both participation ratios drop sharply at CO onset, with corresponding increases in angular separation between l^2 and l^∞ perturbations. This confirms gradient concentration's role in triggering catastrophic behavior and validates our adaptive norm selection strategy.

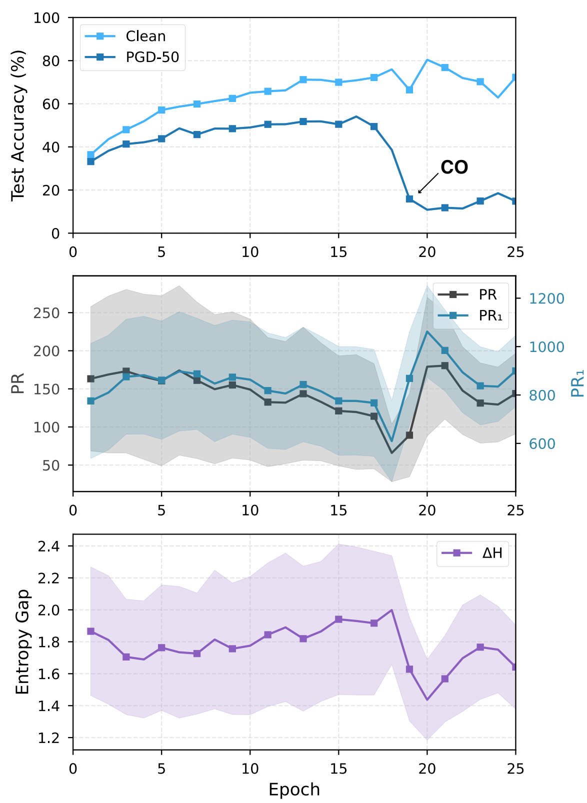


Figure 4: Evolution of Participation Ratios ($\mathtt{PR}$, $\mathtt{PR}_1$) and entropy gap during training. Sharp declines in these metrics precisely align with Catastrophic Overfitting (CO) onset, demonstrating how gradient concentration directly precedes and triggers adversarial vulnerability.

[4] For $p > 2$, maximum occurs when all components have equal amplitude.

Noise-Induced Alignment: Classical CO remedies involve noise injection [10, 13]. Our framework shows that noise increases $\mathtt{PR}_1$, enhancing alignment between l^∞ and l^2 attacks:

Lemma 1 (Noise-Induced Alignment). *For normalized gradient $g = \nabla_x \ell / \|\nabla_x \ell\|_2$ and additive zero-mean noise $\eta \sim \mathcal{U}[-M, M]^d$, there exists $\alpha > 0$ such that if $M < \alpha \|g\|_\infty$, then:*

$$\mathbb{E}\left[\frac{\|g+\eta\|_1}{\|g+\eta\|_2}\right] \geq \frac{\|g\|_1}{\|g\|_2} \tag{17}$$

Proof. See Appendix F. □

Monotonic Angular Relationships: We establish that norm reduction systematically improves angular alignment:

Lemma 2 (Monotonicity of Angular Separation). *For any non-null gradient $\nabla_x \ell$ and $p \geq 3$, the cosine between l^2 and l^p perturbations satisfies:*

$$\cos(\theta_{2,\infty}) \leq \cos(\theta_{2,p}) \quad \textit{where} \quad \cos(\theta_{2,p}) = \frac{\|\nabla_x \ell\|_q^q}{\|\nabla_x \ell\|_2 \|\nabla_x \ell\|_{2(q-1)}^{q-1}} \tag{18}$$

Proof. See Appendix G. □

Entropy-Based Norm Selection: Direct computation of optimal p from Eq. (18) proves challenging. For $q \in [1, 2]$ and moderate increases, first-order Taylor expansion provides computational efficiency (details in Appendix H):

$$\cos(\theta_{2,p}) = \sqrt{\frac{\mathtt{PR}_1}{d}}\,(1 + (q-1)\Delta H) + \mathcal{O}((q-1)^2) \tag{19}$$

where $\Delta H = H_m - H$ is the entropy gap between logarithmic mean entropy H_m and Shannon entropy H of normalized gradient components:

$$H = -\sum_{i=1}^{d} \rho_i \log(\rho_i), \quad H_m = -\log \prod_{i=1}^{d} (\rho_i)^{1/d}, \quad \rho_i = \frac{|\nabla_x \ell_i|}{\|\nabla_x \ell\|_1} \tag{20}$$

Setting a threshold τ below which cosine alignment should not drop, we derive:

$$q^* \geq 1 + \frac{(\tau\sqrt{d/\mathtt{PR}_1} - 1)}{\Delta H},\ \tau \in [0, 1] \tag{21}$$

This formula captures the interplay between gradient geometry and norm selection: when gradients concentrate (low $\mathtt{PR}_1$) and entropy gap decreases, q increases (lower p) to maintain alignment. For practical implementation:

$$\tau \equiv (1+\alpha)\cos(\theta_{2,\infty}) \equiv \cos((1-\beta)\theta_{2,\infty}) \tag{22}$$

These theoretical insights directly inform our algorithmic design. By dynamically adjusting p based on gradient concentration metrics $\mathtt{PR}_1$ and entropy gap, we maintain alignment with natural l^2 geometry when gradients concentrate (low PR_1) and increase p when gradients distribute uniformly (high PR_1). This adaptive approach prevents concentrated gradients from meeting aggressive norm constraints—precisely the condition triggering CO.

3.4 l^p-FGSM Algorithm

Our l^p-FGSM algorithm performs one fixed-point iteration ($\delta^{(1)} = F_p(\delta^{(0)})$) with zero initialization, maintaining computational efficiency while accessing the full spectrum of l^p attack geometries. The epsilon stabilization step serves dual purposes: ensuring numerical stability and satisfying the Lipschitz conditions in Eq. (12).

The adaptive norm selection mechanism automatically adjusts p based on gradient concentration statistics, enabling transitions between attack geometries as training progresses. When gradients concentrate (indicating potential CO onset), the algorithm reduces p to maintain alignment with natural l^2 geometry. When gradients distribute uniformly, higher p values enhance robustness.

This theoretical framework establishes that adaptive norm selection is mathematically sound, maintains convergence properties, and provides a principled solution to catastrophic overfitting without auxiliary techniques like noise injection or regularization.

Algorithm 1 l^p-FGSM

Input: Model θ, data x, labels y, loss ℓ, optimizer, attack amplitude ϵ, norm p (dual q)
repeat
  Sample minibatch (x_0, y_0)
  Compute gradient $g_x \leftarrow \nabla_{x_0}\ell(x_0, y_0)$
  Apply stability term: $\bar{g}_x \leftarrow \varepsilon + |g_x|$
  if adaptive **then** Update q via Eq. 21 using $\mathtt{PR}_1$, ΔH
  Compute attack $\delta_p \leftarrow \epsilon \cdot \text{sign}(g_x) \cdot |\bar{g}_x / \|\bar{g}_x\|_q|^{q-1}$
  Update θ with $\nabla_\theta \ell(x_0 + \delta_p, y_0)$ and optimizer
until Convergence criteria
Output: Robust model θ

4 Experiments and Results

Our l^p-FGSM approach provides computational efficiency over methods requiring double backpropagation, with overhead limited to gradient norm calculations. We evaluate our method on standard datasets, examine norm selection and gradient concentration relationships, and compare against state-of-the-art fast adversarial training methods.

4.1 Comparison with Benchmark Techniques

To rigorously evaluate the effectiveness of adaptive l^p-FGSM, we conducted comprehensive comparisons against several well-established fast adversarial training methods, including RS-FGSM [10], ZeroGrad [12], N-FGSM [13], and GradAlign [11]. This diverse subset, representing fundamentally different conceptual approaches to addressing CO, provides a robust basis for assessing the capacity of adaptive l^p norms to mitigate the phenomenon while maintaining adversarial robustness. For consistency and fair comparison, we used the recommended hyperparameters for each benchmark method as specified in their respective publications.

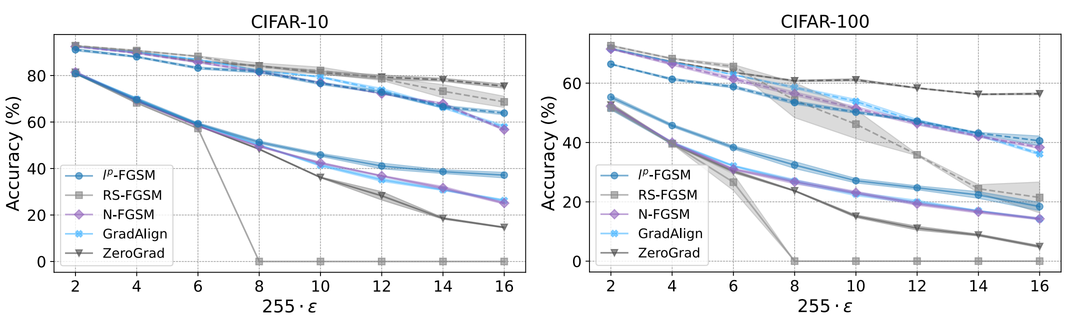


Figure 5: Performance benchmarking of adaptive l^p norm-based training against single-step and fast adversarial techniques using PGD-50-10, demonstrating the competitive efficacy of adaptive l^p-FGSM. Results were achieved with an SGD optimizer with a cosine learning rate schedule (30 epochs, minimum 0.001, maximum 0.2), weight decay of $5 \cdot 10^{-4}$, and a dropout rate of 0.1. For CIFAR-10, $\beta = 0.01$ was applied, while for CIFAR-100, $\beta = 0.1$ was used (Eq. 22). We switched from ADAM to SGD for these comparisons as it is the standard optimizer in adversarial training literature and facilitates direct comparison with published results.

Our empirical studies, summarized in Figure 5, demonstrate that adaptive l^p-FGSM not only meets but often surpasses the robustness benchmarks of leading fast methods [9, 25, 28, 11, 13]. This success hinges on the choice of the l^p norm, which enhances robustness against l^∞ attacks while resolving CO without requiring noise injection or expensive regularization. All components of l^p-FGSM (Alg. 1) are efficient to compute with minimal overhead, making the approach particularly attractive for large-scale applications where computational efficiency is a priority.

The performance advantage of our method is particularly pronounced at higher perturbation magnitudes ($\epsilon \geq 8/255$), where many competing approaches suffer from CO or significant robustness degradation. This innovative use of norm selection introduces a simple yet effective approach to fast adversarial training, offering a novel perspective to advance robust machine learning.

4.2 Experiments with ImageNet

To evaluate adaptive l^p-FGSM on high-resolution images representative of real-world applications, we conducted extensive experiments on ImageNet-1k [29], training a pre-trained ResNet-50 model with ADAM optimizer (lr=10^{-4}, batch size 128) for 15 epochs. We tested our method ($\beta = 0.1$, $\varepsilon = 10^{-12}$) against PGD-50 attacks across a range of perturbation magnitudes $\epsilon = (2, 4, 6)/255$ and compared with established methods including FGSM, RS-FGSM, and N-FGSM.

As shown in Table 4.2, while FGSM experiences catastrophic overfitting at $\epsilon = 6/255$ (evidenced by the near-zero adversarial accuracy), adaptive l^p-FGSM achieves superior adversarial robustness across all perturbation levels while maintaining competitive clean accuracy. The performance advantage is particularly significant at $\epsilon = 4/255$ and $\epsilon = 6/255$, where our method outperforms RS-FGSM by 3.23% and 3.30% in adversarial accuracy, respectively.

Table 1: Comparative Analysis of Robustness Against PGD-50-10 on ImageNet-1k. FGSM, RS-FGSM and N-FGSM results are from [13]. All methods utilize ImageNet-1k pre-trained weights and undergo 15 epochs of training. Results show clean accuracy (top) and PGD-50 accuracy (bottom).

ImageNet-1k ResNet-50			
Method	$\epsilon = 2/255$	$\epsilon = 4/255$	$\epsilon = 6/255$
FGSM	54.72%	48.50%	48.55%
	38.21%	25.86%	0.08%
RS-FGSM	**56.29%**	**50.81%**	47.67%
	36.86%	25.12%	16.49%
l^p-FGSM	53.18%	48.42%	**48.61%**
	37.94%	**28.35%**	**19.79%**
N-FGSM	54.39%	47.56%	47.70%
	38.07%	26.28%	17.12%

These results on ImageNet-1k demonstrate the scalability of our approach to large, complex datasets and its effectiveness in addressing CO in practical settings. The consistent performance advantages across different perturbation magnitudes highlight the robustness of the adaptive norm selection strategy in diverse scenarios, reinforcing the potential of l^p-FGSM as a general-purpose solution for fast adversarial training.

5 Conclusion

We presented adaptive l^p-FGSM, a principled approach to mitigating catastrophic overfitting in fast adversarial training. Our investigation began with the observed discrepancy between l^2 and l^∞ norms, motivating us to explore the full l^p spectrum between these extremes. This led to reformulating adversarial attack generation as a fixed-point problem, enabling efficient single-step methods while providing theoretical insights through Lipschitz continuity analysis. While our approach relies on local convexity assumptions, it gracefully defaults to local linearity when these assumptions do not hold.

Our key finding—that catastrophic overfitting emerges when concentrated gradients meet aggressive norm constraints—provides a unifying perspective on previous observations. By adapting the Participation Ratio from quantum mechanics to measure both gradient concentration and angular separation, we established a quantitative connection between gradient geometry and adversarial vulnerability. This insight led to dynamically adjusting the training norm p based on gradient structure. Although our method avoids double backpropagation, it still requires hyperparameters for angle constraints that warrant further optimization across different architectures and datasets.

This work contributes to understanding fast adversarial training by connecting gradient geometry to training dynamics through an information-theoretic lens. By establishing adaptive norm selection as a theoretically motivated approach, we hope to inspire further research into geometric perspectives on adversarial robustness. Our results suggest that careful consideration of gradient structure may be essential in developing efficient and robust training methods. By establishing a theoretical foundation for addressing catastrophic overfitting, our work contributes to the broader goal of developing reliable machine learning systems that maintain robustness guarantees even under computational constraints.

References

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[4] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *arXiv preprint arXiv:1312.6199*, 2013.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[6] Yue Wang, Esha Sarkar, Saif Eddin Jabari, and Michail Maniatakos. On the vulnerability of deep reinforcement learning to backdoor attacks in autonomous vehicles. In *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges*, pages 315–341. Springer, 2023.

[7] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[8] Micah Goldblum, Avi Schwarzschild, Ankit B Patel, and Tom Goldstein. Adversarial attacks on machine learning systems for high-frequency trading. *arXiv preprint arXiv:2002.09565*, 2020.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[10] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[11] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[12] Zeinab Golgooni, Mehrdad Saberi, Masih Eskandar, and Mohammad Hossein Rohban. Zerograd: Mitigating and explaining catastrophic overfitting in fgsm adversarial training. *arXiv preprint arXiv:2103.15476*, 2021.

[13] Pau de Jorge Aranda, Adel Bibi, Riccardo Volpi, Amartya Sanyal, Philip Torr, Grégory Rogez, and Puneet Dokania. Make some noise: Reliable and efficient single-step adversarial training. *Advances in Neural Information Processing Systems*, 35:12881–12893, 2022.

[14] Runqi Lin, Chaojian Yu, and Tongliang Liu. Eliminating catastrophic overfitting via abnormal adversarial examples regularization. *Advances in Neural Information Processing Systems*, 36:67866–67885, 2023.

[15] Runqi Lin, Chaojian Yu, Bo Han, Hang Su, and Tongliang Liu. Layer-aware analysis of catastrophic overfitting: Revealing the pseudo-robust shortcut dependency. *arXiv preprint arXiv:2405.16262*, 2024.

[16] Chengze Jiang, Junkai Wang, Minjing Dong, Jie Gui, Xinli Shi, Yuan Cao, Yuan Yan Tang, and James Tin-Yau Kwok. Improving fast adversarial training via self-knowledge guidance. *IEEE Transactions on Information Forensics and Security*, 2025.

[17] Elias Abad Rocamora, Fanghui Liu, Grigorios G Chrysos, Pablo M Olmos, and Volkan Cevher. Efficient local linearity regularization to overcome catastrophic overfitting. *arXiv preprint arXiv:2401.11618*, 2024.

[18] Zhaoxin Wang, Handing Wang, Cong Tian, and Yaochu Jin. Preventing catastrophic overfitting in fast adversarial training: A bi-level optimization perspective. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *University of Toronto Technical Report*, 2009.

[20] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[21] Philip W Anderson. Absence of diffusion in certain random lattices. *Physical review*, 109(5):1492, 1958.

[22] Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman Lectures on Physics, Vol. III: Quantum Mechanics*. Addison-Wesley, 1965.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[24] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. *Advances in Neural Information Processing Systems*, 2018.

[26] Suraj Srinivas, Kyle Matoba, Himabindu Lakkaraju, and François Fleuret. Efficient training of low-curvature neural networks. *Advances in Neural Information Processing Systems*, 35:25951–25964, 2022.

[27] Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. Pac-bayesian spectrally-normalized bounds for adversarially robust generalization. *Advances in Neural Information Processing Systems*, 36:36305–36323, 2023.

[28] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[30] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

[31] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[34] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[35] Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8119–8127, 2021.

[36] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

A Local Convexity Analysis

In this appendix, we provide a detailed analysis of the local convexity framework that underlies our l^p-FGSM approach. We examine both the theoretical foundations and empirical evidence for local convexity emergence during adversarial training.

A.1 Theoretical Foundation of Local Convexity

While fast adversarial training traditionally relies on local linearity assumptions through first-order Taylor expansions, we examine a more general local convexity framework that emerges from analyzing the Hessian of the loss function with respect to inputs. When the Hessian $\nabla_x^2 \ell$ is positive definite, any critical point in the perturbation ball's interior must be a local minimum, forcing the maximum to occur on the boundary $\partial B_p(\epsilon)$—a useful property that enables efficient single-step methods.

The Hessian structure can be decomposed with respect to the output logits as:

$$\nabla_x^2 \ell = \left(\frac{\partial \pi}{\partial x_0}\right) \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial x_0}\right)^T + \frac{\partial^2 \pi}{\partial x_0^2} \frac{\partial \ell}{\partial \pi} \tag{23}$$

This decomposition reveals two distinct components:

Gauss-Newton Term: The first term $\left(\frac{\partial \pi}{\partial x_0}\right) \frac{\partial^2 \ell}{\partial \pi^2} \left(\frac{\partial \pi}{\partial x_0}\right)^T$ is positive semi-definite since $\frac{\partial^2 \ell}{\partial \pi^2}$ represents the Hessian of the cross-entropy loss with respect to predictions, which is always positive definite for proper probability distributions.

Error-Dependent Term: The second term $\frac{\partial^2 \pi}{\partial x_0^2} \frac{\partial \ell}{\partial \pi}$ involves the prediction errors $\frac{\partial \ell}{\partial \pi}$. As training progresses and the model's predictions improve, these error terms diminish, reducing the magnitude of the second term relative to the first.

A.2 Convergence to Local Convexity During Training

The natural emergence of local convexity during training can be understood through the evolution of the Hessian structure in Eq. (23). As the model learns to minimize the training loss, the prediction errors $\frac{\partial \ell}{\partial \pi}$ systematically decrease. This causes the potentially indefinite second term to diminish in magnitude relative to the positive semi-definite Gauss-Newton term, leading to an overall positive definite Hessian.

This convergence can be accelerated through architectural choices that control the second-order derivatives $\frac{\partial^2 \pi}{\partial x_0^2}$:

Activation Function Selection: Smooth activation functions like SELU [30] or GELU [31] have well-behaved second derivatives, leading to more stable convergence to local convexity compared to non-smooth activations.

Network Depth and Width: Deeper networks tend to develop local convexity more readily as the composition of smooth functions preserves convexity properties under appropriate conditions.

However, our empirical analysis demonstrates that even standard ReLU networks, despite their non-smooth activation functions, naturally develop local convexity through the training process, as visualized in Figure 6.

A.3 Empirical Evidence for Local Convexity

Figure 6 provides empirical validation of the local convexity emergence during training. The visualization shows the loss landscape around training points at different stages of the training process.

Early Training (Upper Panels): After one epoch, the loss landscapes exhibit irregular, non-convex characteristics with multiple local minima and saddle points. The landscapes are complex and do not satisfy the local convexity assumption.

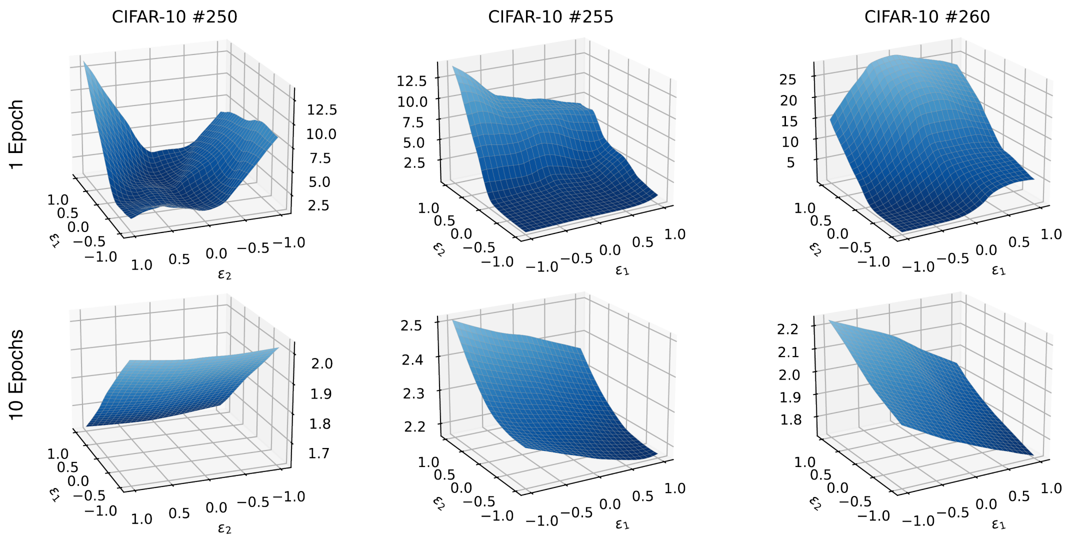


Figure 6: Empirical evidence for local convexity emergence during training on CIFAR-10. The upper panels display the loss landscape after one epoch of training, while the lower panels show the same landscape after ten epochs with l^p-FGSM training. Training points are positioned at $(0, 0)$; ε_1 and ε_2 are eigenvectors corresponding to the extreme eigenvalues of the input Hessian $\nabla_x^2 \ell$ for each sample. The progressive development of convex loss landscapes validates our theoretical framework and provides justification for boundary-focused adversarial search strategies.

Later Training (Lower Panels): After ten epochs of training, the same landscapes show clear convex structure around the training points. The loss increases monotonically as we move away from the training point in any direction within the neighborhood, confirming positive definiteness of the local Hessian.

This empirical observation has several important implications:

1. **Theoretical Validation:** The emergence of local convexity validates our theoretical framework and justifies the use of boundary-focused optimization strategies.

2. **Practical Robustness:** Even when local convexity does not hold initially, the framework can gracefully default to local linearity assumptions, ensuring robustness across different training phases.

3. **Efficient Optimization:** The development of local convexity enables more efficient single-step adversarial example generation, as the optimal perturbations are guaranteed to be on the constraint boundary.

B Appendix: Demonstration l^2 Optimal Attack

Proposition 1. *Consider a training sample x_0 with a non-null gradient. The optimal perturbation $\delta^\star$ within $B(\epsilon)$ exists and corresponds to the solution of a fixed-point problem $\delta^\star = F(\delta^\star)$, where*

$$F(\delta) = \epsilon \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_2} \tag{24}$$

The function F exhibits Lipschitzian behavior around its origin, satisfying:

$$\|F(\delta) - F(0)\| \leq 2\epsilon \frac{\|\nabla_x^2 \ell\|}{\|\nabla_x \ell(x_0)\|_2} \|\delta\| \tag{25}$$

The fixed-point problem converges if it is contractive:

$$K = 2\epsilon \frac{\|\nabla_x^2 \ell\|}{\|\nabla_x \ell(x_0)\|_2} < 1 \tag{26}$$

Proof. Assuming that the Hessian of the loss function, $\nabla_x^2 \ell$, is positive definite, any critical point in the interior would be a minimum. The implicitly assumed compactness guarantees the existence of the maximum on the boundary. The constrained maximization uses the Lagrangian:

$$\mathcal{L}(\delta, \lambda) = \ell(x_0 + \delta) - \frac{\lambda}{2}(\delta^T \delta - \epsilon^2) \tag{27}$$

The derivatives yield the following equations:

$$\{ \tfrac{\partial}{\partial \delta} \mathcal{L} = \nabla_x \ell(x_0 + \delta) - \lambda \delta = 0 \tfrac{\partial}{\partial \lambda} \mathcal{L} = -\tfrac{1}{2}(\delta^T \delta - \epsilon^2) = 0 \tag{28}$$

Since the maximum exists on the boundary, the constraint $\delta^T \delta = \epsilon^2$ is activated; hence the Lagrange multiplier λ is non-null. The gradient at $x_0 + \delta$ cannot be null (minimum otherwise), therefore $\|\nabla_x \ell(x_0 + \delta)\| > 0$.

Solving the two Lagrangian equations yields:

$$\delta = \pm \epsilon \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|} \tag{29}$$

Given the positive Hessian assumption, moving along the gradient (equivalent to choosing the positive sign) results in a greater change in the loss function ℓ. Consequently:

$$\delta = \epsilon \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|} \tag{30}$$

The maximum $\delta^\star$ is the solution to a fixed-point problem. The existence and uniqueness of the solution $\delta^\star$ is guaranteed if $F(\delta)$ is contractive, i.e., Lipschitz continuous with a Lipschitz constant $K < 1$.

To demonstrate this Lipschitz continuity, we consider:

$$\|F(\delta) - F(0)\| = \epsilon \Big\| \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|} - \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|} \Big\|$$

By introducing a cross term and using the triangle inequality:

$$\|F(\delta) - F(0)\| \leq \epsilon \Big\| \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|} - \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0)\|} \Big\| + \epsilon \Big\| \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0)\|} - \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|} \Big\|$$

The first term can be bounded:

$$\|F(\delta_1) - F(0)\| \leq \epsilon \frac{\|\nabla_x^2 \ell(x_0)\| \|\delta\|}{\|\nabla_x \ell(x_0)\|} + \epsilon \|\nabla_x \ell(x_0 + \delta)\| \Big| \frac{1}{\|\nabla_x \ell(x_0 + \delta)\|} - \frac{1}{\|\nabla_x \ell(x_0)\|} \Big|$$

After unifying the denominator:

$$\|F(\delta) - F(0)\| \leq \epsilon \frac{\|\nabla_x^2 \ell\| \|\delta\|}{\|\nabla_x \ell(x_0)\|} + \frac{\epsilon}{\|\nabla_x \ell(x_0)\|} \Big| \|\nabla_x \ell(x_0 + \delta)\| - \|\nabla_x \ell(x_0)\| \Big|$$

Using the triangle inequality again:

$$\Big| \|\nabla_x \ell(x_0 + \delta)\| - \|\nabla_x \ell(x_0)\| \Big| \leq \|\nabla_x \ell(x_0 + \delta) - \nabla_x \ell(x_0)\| \leq \|\nabla_x^2 \ell\| \|\delta\|$$

This leads to:

$$\|F(\delta) - F(0)\| \leq 2\epsilon \frac{\|\nabla_x^2 \ell(x_0)\| \|\delta\|}{\|\nabla_x \ell(x_0)\|} \tag{31}$$

The Lipschitz constant is:

$$K = 2\epsilon \cdot \frac{\|\nabla_x^2 \ell(x_0)\|}{\|\nabla_x \ell(x_0)\|} \tag{32}$$

Assuming $K < 1$, the fixed point problem converges. □

C Appendix: Demonstration l^p Optimal Attack

Proposition 2. *For a training sample x_0 exhibiting a non-null gradient and a constraint within $B_p(\epsilon)$, the optimal perturbation, denoted as $\delta^\star$, exists and corresponds to the solution of a fixed-point problem: $\delta^\star = F_p(\delta^\star)$. Specifically, we have:*

$$F_p(\delta) = \epsilon \text{sign}(\nabla_x \ell(x_0 + \delta)) \left| \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} \right|^{q-1} \tag{33}$$

where the l^q norm serves as the dual to l^p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$. The absolute value and multiplication operations are element-wise.

Proof. Assuming the same hypotheses as in Appendix A, a maximum exists on the boundary of the B_p ball. We formulate the Lagrangian with the l^p equality constraint:

$$\mathcal{L}_p(\delta, \lambda) = \ell(x_0 + \delta) - \lambda(\|\delta\|_p - \epsilon) \tag{34}$$

The l^p norm is given by:

$$\|\delta\|_p = \left(\sum_i |\delta_i|^p \right)^{\frac{1}{p}} \tag{35}$$

Hence, its derivative is:

$$\frac{\partial}{\partial \delta} \|\delta\|_p = \text{sign}(\delta) \left(\frac{|\delta|}{\|\delta\|_p} \right)^{p-1} \tag{36}$$

The derivatives of the Lagrangian are:

$$\begin{cases} \frac{\partial}{\partial \delta} \mathcal{L}_p = \nabla_x \ell(x_0 + \delta) - \lambda \text{sign}(\delta) \left(\frac{|\delta|}{\|\delta\|_p} \right)^{p-1} = 0 \\ \frac{\partial}{\partial \lambda} \mathcal{L}_p = -(\|\delta\|_p - \epsilon) = 0 \end{cases} \tag{37}$$

Using the dual norm l^q defined with $\frac{1}{p} + \frac{1}{q} = 1 \rightarrow q = \frac{p}{p-1}$, we can characterize λ as:

$$\|\nabla_x \ell(x_0 + \delta)\|_q = \frac{|\lambda|}{\|\delta\|_p^{p-1}} (\|\delta\|_p^p)^{\frac{1}{q}} = |\lambda| \tag{38}$$

Substituting into the first derivative of the Lagrangian:

$$\nabla_x \ell(x_0 + \delta) = \pm \|\nabla_x \ell(x_0 + \delta)\|_q \ \mathrm{sign}(\delta) \left(\frac{|\delta|}{\|\delta\|_p} \right)^{p-1} \tag{39}$$

From this, δ and $\nabla_x \ell(x_0 + \delta)$ have the same sign up to a multiplicative coefficient (i.e., $\pm$):

$$\frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} = \pm \left| \frac{\delta}{\|\delta\|_p} \right|^{p-1} \mathrm{sign}(\delta) \tag{40}$$

Extracting δ and using $\|\delta\|_p = \epsilon$ yields:

$$\delta = \pm \epsilon \mathrm{sign}(\nabla_x \ell(x_0 + \delta)) \times \left(\frac{|\nabla_x \ell(x_0 + \delta)|}{\|\nabla_x \ell(x_0 + \delta)\|_q} \right)^{\frac{1}{p-1}}$$

The solution with the negative sign would yield a locally decreasing loss function, so we take the positive solution. The Lagrange multiplier for maximization is positive:

$$\lambda = \|\nabla_x \ell(x_0 + \delta)\|_q \tag{41}$$

Using $p = \frac{q}{q-1} \to p - 1 = \frac{1}{q-1}$, we get the final result:

$$\delta = \epsilon \mathrm{sign}(\nabla_x \ell(x_0 + \delta)) \times \left(\frac{|\nabla_x \ell(x_0 + \delta)|}{\|\nabla_x \ell(x_0 + \delta)\|_q} \right)^{q-1}$$

$\square$

D Appendix: Lipschitzness of the l^p Fixed-Point Problem

We assume: $\exists\, m > 0 : \forall \delta \in \partial B_p(\epsilon), |\nabla_\theta \ell(x_0 + \delta)_i| > m$, and proceed to demonstrate Lipschitzness of the function $F_p(\delta)$ verifying the fixed point, defined as:

$$F_p(\delta) = \epsilon \ \mathrm{sign}(\nabla_x \ell(x_0 + \delta)) \left| \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} \right|^{q-1} \tag{42}$$

The sign function can be circumvented by using "one power" of the absolute value of the gradient:

$$F_p(\delta) = \epsilon \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} \times \left(\frac{|\nabla_x \ell(x_0 + \delta)|}{\|\nabla_x \ell(x_0 + \delta)\|_q} \right)^{q-2} \tag{43}$$

The term $q - 2$ is negative, which is permissible since we assumed a lower limit m for gradient values. Our objective is to prove that $F_p(\delta)$ is Lipschitz continuous around $\delta = 0$.

First, let's define:

$$f_q(\delta) = \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} \tag{44}$$

We have:

$$F_p(\delta) = \epsilon f_q(\delta) |f_q(\delta)|^{q-2} \tag{45}$$

Similar to Appendix A, by introducing a cross term we can show that f and $|f|$ are Lipschitz continuous, with a constant K_f such that:

$$|f_q(\delta) - f_q(0)| \leq K_f \|\delta\| \tag{46}$$

The same steps are applied as follows:

$$\||f_q(\delta)| - |f_q(0)|\| \leq \|f_q(\delta) - f_q(0)\| \leq \left\| \frac{\nabla_x \ell(x_0 + \delta)}{\|\nabla_x \ell(x_0 + \delta)\|_q} - \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|_q} \right\| \tag{47}$$

By further manipulation and using the triangle inequality:

$$\||f_q(\delta)| - |f_q(0)|\| \leq \left\| \frac{\nabla_x \ell(x_0+\delta)}{\|\nabla_x \ell(x_0+\delta)\|_q} - \frac{\nabla_x \ell(x_0+\delta)}{\|\nabla_x \ell(x_0)\|_q} \right\| + \left\| \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|_q} - \frac{\nabla_x \ell(x_0+\delta)}{\|\nabla_x \ell(x_0)\|_q} \right\| \tag{48}$$

This leads to:

$$\||f_q(\delta)| - |f_q(0)|\| \leq \left(1 + \frac{\|\nabla_x \ell(x_0+\delta)\|}{\|\nabla_x \ell(x_0+\delta)\|_q}\right) \times \frac{\|\nabla_x^2 \ell(x_0)\|}{\|\nabla_x \ell(x_0)\|_q} \|\delta\| \tag{49}$$

In a finite-dimensional vector space, all norms are equivalent:

$$\exists C \geq 0,\ \frac{\|\nabla_x \ell(x_0+\delta)\|}{\|\nabla_x \ell(x_0+\delta)\|_q} \leq C \tag{50}$$

Next, examining $|x|^{q-2}$ on the interval $[m, +\infty)$ with $q-2$ negative:

$$\forall (x,y) \in [m, +\infty),\ ||x|^{q-2} - |y|^{q-2}| \leq (2-q) m^{q-3} |x-y| \tag{51}$$

Using these results for the local Lipschitz continuity of F_p:

$$\frac{1}{\epsilon} \|F_p(\delta) - F_p(0)\| = \|f_q(\delta)|f_q(\delta)|^{q-2} - f_q(0)|f_q(0)|^{q-2}\| \tag{52}$$

Through a series of bounds:

$$\begin{aligned} \frac{1}{\epsilon} \|F_p(\delta) - F_p(0)\| &\leq \|f_q(\delta)|f_q(\delta)|^{q-2} - f_q(\delta)|f_q(0)|^{q-2}\| \\ &\quad + \|f_q(\delta)|f_q(0)|^{q-2} - f_q(0)|f_q(0)|^{q-2}\| \end{aligned} \tag{53}$$

Further simplifying:

$$\begin{aligned} \frac{1}{\epsilon} \|F_p(\delta) - F_p(0)\| &\leq \frac{\|\nabla_x \ell(x_0+\delta)\|}{\|\nabla_x \ell(x_0+\delta)\|_q} \times (2-q) m^{q-3} |f_q(\delta) - f_q(0)| \\ &\quad + \left| \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|_q} \right|^{q-2} \times \|f_q(\delta) - f_q(0)\| \end{aligned} \tag{54}$$

This yields:

$$\|F_p(\delta) - F_p(0)\| \leq K(p,m)\epsilon \times \frac{\|\nabla_x^2 \ell(x_0)\|}{\|\nabla_x \ell(x_0)\|_q} \|\delta\| \tag{55}$$

where:

$$K(p,m) = \left(C(2-q) m^{q-3} + \left(\frac{m}{\|\nabla_x \ell(x_0)\|_q}\right)^{q-2}\right)(1+C) \tag{56}$$

E Preliminary Validation of Fixed l^p Norms

To understand the fundamental limitations of fixed p values and motivate our adaptive approach, we conducted systematic evaluation of l^p-FGSM across different norm values on standard datasets. This preliminary analysis reveals the inherent trade-offs that necessitate adaptive norm selection. All experiments were conducted on a single NVIDIA A100 GPU.

E.1 Experimental Setup

We evaluate fixed l^p-FGSM following the framework of Wong et al. [10] using PGD-50 attacks on CIFAR-10, CIFAR-100 [19], and SVHN [32]. Experiments use PreactResNet18 [33] for SVHN and WideResNet28-10 [20] for CIFAR datasets, with results averaged over five seeds for reliability.

This validation deliberately excludes enhancements like weight decay, dropout, or noise injection to isolate the effects of norm selection and provide a clear baseline for understanding the impact of the l^p norm parameter. All experiments use perturbation radius $\epsilon = 8/255$ for both training and evaluation attacks.

E.2 Key Findings: The Fixed p Dilemma

Figure 7 presents comprehensive results across all three datasets, revealing several critical insights about the fundamental limitations of fixed norm approaches.

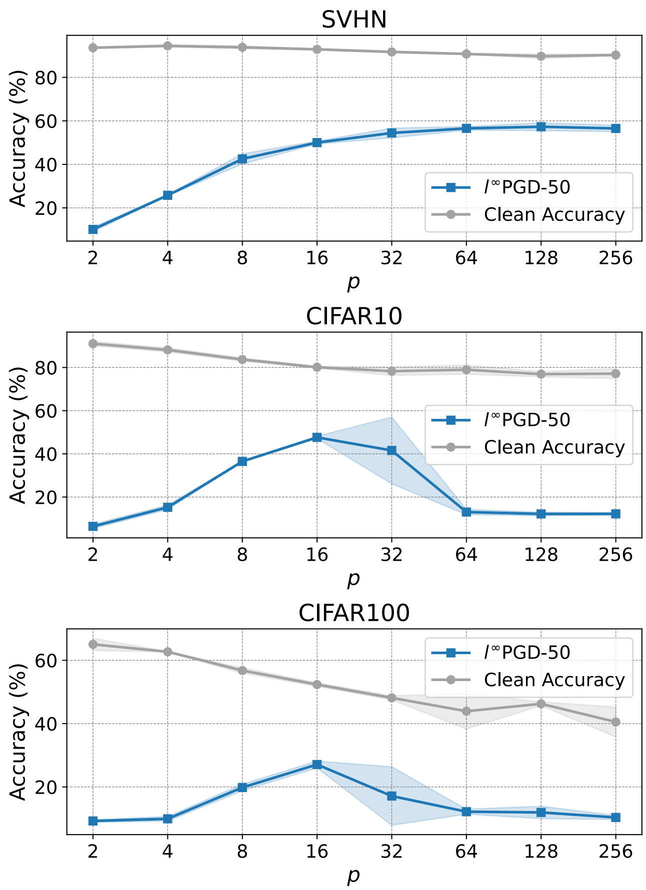


Figure 7: Detailed analysis of clean and adversarial accuracy across CIFAR-10, CIFAR-100, and SVHN datasets with $\epsilon = 8/255$ for different p values. The results demonstrate the fundamental limitations of fixed norm approaches: CIFAR-10 shows optimal performance at intermediate $p \approx 16 - 32$ before CO onset, SVHN exhibits remarkable resilience to CO even at higher p values, while CIFAR-100 displays heightened sensitivity to norm selection with narrow optimal ranges. These dataset-dependent behaviors highlight the critical need for adaptive norm selection.

The results demonstrate striking dataset-dependent optimal ranges that expose the inadequacy of any universal fixed p approach. CIFAR-10 achieves optimal performance at intermediate p values around 16-32, demonstrating a clear sweet spot before catastrophic overfitting occurs. In contrast, SVHN exhibits remarkable resilience to CO even at higher p values, suggesting that simpler datasets can tolerate more aggressive norm constraints for extended periods. CIFAR-100 shows heightened sensitivity to norm selection with narrow optimal ranges, indicating that complex datasets require more conservative and careful norm tuning.

Across all datasets, we observe a universal trade-off pattern that reveals the inherent limitations of fixed approaches. Lower p values ($p \leq 4$) provide excellent stability against catastrophic overfitting but at the cost of significantly reduced adversarial robustness. Higher p values ($p \geq 64$) initially improve robustness but eventually lead to catastrophic overfitting, with the onset timing varying dramatically by dataset complexity. Intermediate p values offer the best balance but require careful tuning that is fundamentally dataset-dependent.

The relationship between dataset complexity and optimal norm selection proves particularly striking. Simple datasets like SVHN tolerate aggressive norms for longer periods, while complex datasets like CIFAR-100 require more conservative norm choices from the outset. This suggests that gradient structure varies significantly across problem domains, with complexity directly influencing the rate at which gradient concentration occurs during training.

E.3 Fundamental Limitations of Fixed Norm Approaches

These results expose several fundamental limitations of any fixed p approach that render such methods inadequate for general-purpose adversarial training. The lack of generalizability is perhaps most concerning: no single p value works optimally across all datasets, with configurations that succeed for SVHN failing dramatically for CIFAR-100. This dataset dependency makes fixed approaches impractical for real-world deployment where diverse data characteristics are encountered.

The static nature of fixed values conflicts directly with the dynamic nature of adversarial training. Gradient structure evolves throughout training, with early phases potentially benefiting from higher p values while later stages require lower values to prevent catastrophic overfitting. Fixed approaches cannot adapt to these changing conditions, forcing suboptimal compromises throughout the training process.

Even the best fixed p value for each dataset represents a compromise that sacrifices either robustness or stability. The narrow optimal ranges, particularly evident in CIFAR-100, make fixed approaches highly sensitive to hyperparameter selection and prone to overfitting validation performance. This sensitivity creates practical deployment challenges where slight dataset variations can push performance outside optimal ranges.

E.4 Theoretical Alignment and Motivation for Adaptive Approaches

These empirical observations align perfectly with our gradient concentration hypothesis and provide strong motivation for adaptive norm selection. Complex datasets like CIFAR-100 likely exhibit more concentrated gradients earlier in training, requiring conservative norm choices to prevent early catastrophic overfitting. Simple datasets like SVHN maintain more distributed gradients longer, tolerating aggressive norms without immediate vulnerability. Intermediate complexity datasets like CIFAR-10 require dynamic adaptation as gradient structure evolves throughout training.

The clear dataset dependency and fundamental trade-offs exposed in these experiments provide compelling evidence that fixed norm approaches are inherently limited. An effective solution must automatically adapt to different dataset characteristics without manual tuning, respond to changing gradient structure throughout training, base norm selection on measurable gradient properties that predict catastrophic overfitting onset, and maintain computational efficiency comparable to fixed approaches.

This preliminary analysis establishes the empirical foundation for our theoretical framework and demonstrates why gradient-aware adaptive norm selection is not merely beneficial but necessary for robust fast adversarial training across diverse problem domains. The development of our adaptive l^p-FGSM framework detailed in the main paper directly addresses these limitations through principled gradient concentration measurement and automatic norm adaptation.

F Appendix: Proof of Noise-Induced Alignment

Lemma 1 (Noise-Induced Alignment). *For $g \in \mathbb{R}^d$ nonzero and $\eta \sim \mathcal{U}[-M, M]^d$, $\exists \alpha > 0$ such that if $M < \alpha \|g\|_\infty$:*

$$\mathbb{E}\left[\frac{\|g+\eta\|_1}{\|g+\eta\|_2}\right] \geq \frac{\|g\|_1}{\|g\|_2} \tag{57}$$

Proof. Let $S_+ = \{i : |g_i| > M\}$ and $S_- = \{i : |g_i| \leq M\}$ partition coordinates.

For $i \in S_+$:

$$\sum_{i \in S_+} |g_i + \eta_i| \geq \sum_{i \in S_+} (|g_i| - M) \tag{58}$$

For $i \in S_-$, direct calculation yields:

$$\mathbb{E}[|g_i + \eta_i|] = \frac{1}{2M} \int_{-M}^{M} |g_i + \eta| \, d\eta = \frac{(g_i + M)^2 + (g_i - M)^2}{4M} = \frac{g_i^2 + M^2}{2M} \tag{59}$$

Thus for the l^1 norm:

$$\mathbb{E}[\|g + \eta\|_1] \geq \sum_{i \in S_+} (|g_i| - M) + \sum_{i \in S_-} \frac{g_i^2 + M^2}{2M} \tag{60}$$

For the l^2 norm, using $\mathbb{E}[\eta_i^2] = \frac{M^2}{3}$ and independence:

$$\mathbb{E}[\|g + \eta\|_2^2] = \sum_{i=1}^{d} \left(g_i^2 + \frac{M^2}{3} \right) \tag{61}$$

By Jensen's inequality applied to the concave function $f(x) = \sqrt{x}$:

$$\begin{aligned} \mathbb{E}[\|g + \eta\|_2] &= \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} (g_i + \eta_i)^2} \right] \\ &\leq \sqrt{\mathbb{E}\left[\sum_{i=1}^{d} (g_i + \eta_i)^2 \right]} \\ &= \sqrt{\sum_{i=1}^{d} \left(g_i^2 + \frac{M^2}{3} \right)} \end{aligned} \tag{62}$$

Let $\mathcal{E}$ be the event where:

$$\|g + \eta\|_2 \leq \sqrt{\sum_{i=1}^{d} \left(g_i^2 + \frac{M^2}{2} \right)} \tag{63}$$

Then:

$$\mathbb{E}\left[\frac{\|g + \eta\|_1}{\|g + \eta\|_2} \right] \geq \mathbb{P}(\mathcal{E}) \cdot \frac{\sum_{i \in S_+} (|g_i| - M) + \sum_{i \in S_-} \frac{g_i^2 + M^2}{2M}}{\sqrt{\sum_{i=1}^{d} \left(g_i^2 + \frac{M^2}{2} \right)}} \tag{64}$$

For $M < \alpha \|g\|_\infty$ with α sufficiently small:

- $\mathbb{P}(\mathcal{E})$ approaches 1
- The gain in S_- terms ($\frac{g_i^2 + M^2}{2M} > |g_i|$) exceeds the loss in S_+ terms
- The denominator remains close to $\|g\|_2$

Therefore, the ratio exceeds $\frac{\|g\|_1}{\|g\|_2}$. □

G Appendix: Proof of Monotonicity of Angular Separation

Lemma 2 (Monotonicity of Angular Separation). *For any gradient $\nabla_x \ell$ and $2 \leq p \leq \infty$, the cosine similarity between l_2 and l_p perturbations satisfies:*

$$\cos(\theta_{2,p}) \geq \cos(\theta_{2,\infty}) = \sqrt{\frac{\mathtt{PR}_1}{d}} \tag{65}$$

Proof. **Step 1: Express $\cos(\theta_{2,p})$ in normalized form.**

Let $q = \frac{p}{p-1}$ be the dual exponent of p; hence $2 \leq p \leq \infty$ implies $1 \leq q \leq 2$. Recall that:

$$\delta_p = \epsilon \operatorname{sign}(\nabla_x \ell(x_0)) \left| \frac{\nabla_x \ell(x_0)}{\|\nabla_x \ell(x_0)\|_q} \right|^{q-1} \tag{66}$$

$$\delta_\infty = \epsilon \operatorname{sign}(\nabla_x \ell(x_0)) \tag{67}$$

Then:

$$\cos(\theta_{2,p}) = \frac{\langle \delta_2, \delta_p \rangle}{\|\delta_2\|_2 \|\delta_p\|_2} \tag{68}$$

which yields:

$$\cos(\theta_{2,p}) = \frac{\|\nabla_x \ell\|_q^q}{\|\nabla_x \ell\|_2 \, \|\nabla_x \ell\|_{2(q-1)}^{q-1}} \tag{69}$$

We introduce the normalized vector:

$$g = \frac{\nabla_x \ell}{\|\nabla_x \ell\|_2} \tag{70}$$

Then $\|g\|_2 = 1$, and each coordinate of g satisfies $|g_i| \leq 1$. Using g, we can rewrite:

$$\|\nabla_x \ell\|_q = \|\nabla_x \ell\|_2 \left\| \frac{\nabla_x \ell}{\|\nabla_x \ell\|_2} \right\|_q = \|\nabla_x \ell\|_2 \, \|g\|_q \tag{71}$$

Hence:

$$\|\nabla_x \ell\|_q^q = \|\nabla_x \ell\|_2^q \, \|g\|_q^q \tag{72}$$

Similarly, we have:

$$\|\nabla_x \ell\|_{2(q-1)}^{q-1} = \|\nabla_x \ell\|_2^{q-1} \, \|g\|_{2(q-1)}^{q-1} \tag{73}$$

So:

$$\cos(\theta_{2,p}) = \frac{\|\nabla_x \ell\|_2^q \, \|g\|_q^q}{\|\nabla_x \ell\|_2 \, \|\nabla_x \ell\|_2^{q-1} \, \|g\|_{2(q-1)}^{q-1}} \tag{74}$$

$$= \frac{\|g\|_q^q}{\|g\|_{2(q-1)}^{q-1}} \tag{75}$$

Step 2: Show that $\|g\|_q^q \geq \|g\|_1$ and $\|g\|_{2(q-1)}^{q-1} \leq \|g\|_2^{q-1}$.

Since $\|g\|_2 = 1$, all coordinates $|g_i| \leq 1$. For $q \in [1, 2]$, raising each $|g_i|$ from exponent 1 up to q increases the value coordinate-wise, hence:

$$|g_i|^q \geq |g_i|^1 \Rightarrow \|g\|_q^q = \sum_i |g_i|^q \geq \sum_i |g_i|^1 = \|g\|_1 \tag{76}$$

For $q \leq 1.5$ (equivalent to $p \geq 3$), then $2(q-1) \leq 1$. In that regime, raising $|g_i|$ to a power below 1 makes sums larger. For $0 < \epsilon \leq 2(q-1)$:

$$|g_i|^{2(q-1)} \leq |g_i|^\epsilon \tag{77}$$

$$\Rightarrow \|g\|_{2(q-1)}^{2(q-1)} \leq \sum_i |g_i|^\epsilon = \|g\|_\epsilon^\epsilon \tag{78}$$

As $\epsilon \to 0$, $\|g\|_\epsilon^\epsilon \to d$, and we get:

$$\|g\|_{2(q-1)}^{q-1} \leq \sqrt{d} \tag{79}$$

Step 3: Combine results.

Using the factorization from Step 1 and the inequalities from Step 2:

$$\cos(\theta_{2,p}) = \frac{\|g\|_q^q}{\|g\|_{2(q-1)}^{q-1}} \tag{80}$$

$$\geq \frac{\|g\|_1}{\sqrt{d}} = \cos(\theta_{2,\infty}) \tag{81}$$

$\square$

H Appendix: Taylor Expansion of Cosine Similarity

Proposition 3. *For $q = 1 + \epsilon$ with small ϵ and normalized gradient components $\pi_i = \frac{|\nabla_x \ell_i|}{\|\nabla_x \ell\|_1}$, the cosine similarity between l^2 and l^p perturbations admits the following first-order expansion:*

$$\cos(\theta_{2,p}) = \sqrt{\frac{\mathtt{PR}_1}{d}} \left(1 + \epsilon(H_m - H)\right) + O(\epsilon^2) \tag{82}$$

where $\mathtt{PR}_1 = \left(\frac{\|\nabla_x \ell\|_1}{\|\nabla_x \ell\|_2}\right)^2$ is the participation ratio, H is the Shannon entropy, and H_m is the logarithmic mean entropy.

Proof. Starting with the cosine similarity for $q = 1 + \epsilon$:

$$\cos(\theta_{2,p}) = \frac{\|\nabla_x \ell\|_q^q}{\|\nabla_x \ell\|_2 \|\nabla_x \ell\|_{2(q-1)}^{q-1}} \tag{83}$$

The numerator expands directly as:

$$\|\nabla_x \ell\|_q^q = \sum_i |\nabla_x \ell_i|^{1+\epsilon} = \|\nabla_x \ell\|_1 \Big(1 + \epsilon \sum_i \frac{|\nabla_x \ell_i|}{\|\nabla_x \ell\|_1} \times \log |\nabla_x \ell_i| + O(\epsilon^2)\Big) \tag{84}$$

For the denominator term $\|\nabla_x \ell\|_{2\epsilon}^{\epsilon}$:

$$\|\nabla_x \ell\|_{2\epsilon}^{\epsilon} = \Big(1 + 2\epsilon \sum_i \frac{\log |\nabla_x \ell_i|}{d} + O(\epsilon^2)\Big)^{\frac{1}{2}} = 1 + \epsilon \sum_i \frac{\log |\nabla_x \ell_i|}{d} + O(\epsilon^2) \tag{85}$$

Combining terms with normalized gradient components π_i:

$$\cos(\theta_{2,p}) = \frac{\|\nabla_x \ell\|_1}{\|\nabla_x \ell\|_2 \sqrt{d}} \Big(1 + \epsilon\big(\sum_i \pi_i \log |\nabla_x \ell_i| - \sum_i \frac{\log |\nabla_x \ell_i|}{d}\big)\Big) + O(\epsilon^2) \tag{86}$$

The sums relate to entropy measures through:

$$\sum_i \pi_i \log |\nabla_x \ell_i| = -H + \log \|\nabla_x \ell\|_1 \tag{87}$$

$$\sum_i \frac{\log |\nabla_x \ell_i|}{d} = -H_m + \log \|\nabla_x \ell\|_1 \tag{88}$$

where:

$$H = -\sum_i \pi_i \log(\pi_i) \tag{89}$$

$$H_m = -\log \prod_{i=1}^{d} (\pi_i)^{\frac{1}{d}} \tag{90}$$

Therefore:

$$\cos(\theta_{2,p}) = \sqrt{\frac{\mathtt{PR}_1}{d}} \left(1 + \epsilon(H_m - H)\right) + O(\epsilon^2) \tag{91}$$

The entropy gap $\Delta H = H_m - H$ is always positive by Jensen's inequality. $\square$

I Appendix: AutoAttack Results

To ensure a comprehensive assessment, we have also included robust accuracy results evaluated with AutoAttack (AA) [34]. We present the clean (top) and robust (bottom) accuracies (3 seeds) for CIFAR-10 using WRN-28-8, evaluated with AA. The pattern observed is consistent with the results from PGD-50, showing a common trend.

Table 2: CIFAR-10 (WRN-28-8) Clean and AutoAttack Accuracy Evaluation. Results are averaged over multiple seeds. Clean accuracy (top) and AutoAttack accuracy (bottom).

CIFAR-10 WRN-28-10 AutoAttack				
$255 \cdot \epsilon$	FGSM	RS-FGSM	N-FGSM	l^p-FGSM
2	**90.81**% $\pm$ 0.07	90.64% $\pm$ 0.12	89.27% $\pm$ 0.21	89.02% $\pm$ 0.41
	74.72% $\pm$ 0.37	71.47% $\pm$ 0.44	73.14% $\pm$ 0.68	**76.14**% $\pm$ 0.62
4	**87.86**% $\pm$ 0.23	86.58% $\pm$ 0.22	86.34% $\pm$ 0.36	85.71% $\pm$ 0.53
	61.58% $\pm$ 0.12	54.85% $\pm$ 0.16	59.81% $\pm$ 0.27	**62.12**% $\pm$ 0.42
8	**84.89**% $\pm$ 1.20	80.14% $\pm$ 0.88	74.73% $\pm$ 0.46	79.81% $\pm$ 0.57
	0.00% $\pm$ 0.00	35.77% $\pm$ 0.24	41.65% $\pm$ 0.45	**42.43**% $\pm$ 0.58
12	**80.23**% $\pm$ 0.63	61.65% $\pm$ 1.32	62.56% $\pm$ 0.73	71.12% $\pm$ 0.38
	0.00% $\pm$ 0.00	0.00% $\pm$ 0.00	30.17% $\pm$ 1.16	**32.13**% $\pm$ 0.71
16	**74.61**% $\pm$ 0.19	69.20% $\pm$ 0.15	52.89% $\pm$ 0.27	58.43% $\pm$ 0.48
	0.00% $\pm$ 0.00	0.00% $\pm$ 0.00	22.50% $\pm$ 0.89	**25.89**% $\pm$ 0.59

The comparison encompasses standard FGSM [5], RS-FGSM [10], N-FGSM with (k=2) [13], and our proposed adaptive l^p-FGSM ($\beta = 0.01$). The experiments reveal a characteristic pattern of Catastrophic Overfitting (CO) across various perturbation magnitudes (ϵ) for FGSM and RS-FGSM. During CO, models maintain high clean accuracy while their robust accuracy against adversarial attacks deteriorates to near zero.

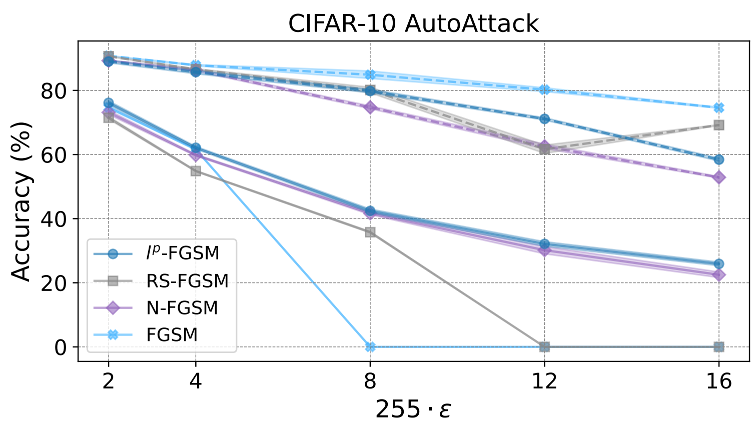


Figure 8: Comparative evaluation using AutoAttack on CIFAR-10 with WideResNet-28-10 across different perturbation magnitudes. Results demonstrate consistent robustness assessment between PGD-50 and AutoAttack [34], validating the reliability of our evaluation methodology.

The strong agreement between PGD-50 and AutoAttack results strengthens our evaluation methodology, as AutoAttack combines multiple complementary attack strategies [34, 11]. This comprehensive assessment validates our findings regarding the effectiveness of norm selection in preventing CO.

J Appendix: Long-Term Training Evaluation

To rigorously assess the durability and stability of the l^p-FGSM method under prolonged training conditions, we conducted an extended training experiment spanning 200 epochs. This experiment utilized the CIFAR-10 dataset with adversarial perturbation norms set at $\epsilon = 8/255$ and $\epsilon = 16/255$, using ADAM optimizer with a learning rate of 0.001.

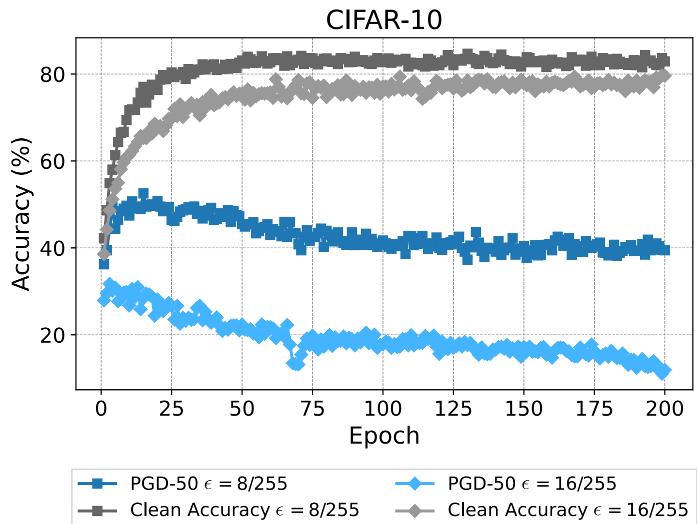


Figure 9: Extended training performance of l^p-FGSM on CIFAR-10. While Catastrophic Overfitting (CO) was not observed, the experiment highlights the occurrence of robust overfitting over a prolonged training period.

The results of this long-term training provide insightful observations. Crucially, no instances of Catastrophic Overfitting (CO) were detected throughout the training process, underscoring the robustness of the l^p-FGSM approach. However, a slight decrease in robustness, i.e., robust overfitting, occurs. This occurrence warrants early stopping and cyclical learning rates to offset this phenomenon.

K Appendix: l^p-FGSM Results Tables

Table 3: Comparative Analysis of Fast Adversarial Training Methods on SVHN Dataset

SVHN PreAct-18 PGD-50-10					
$\epsilon \cdot 255$	l^p-FGSM	RS-FGSM	N-FGSM	GradAlign	ZeroGrad
2	94.20% ±0.52	**96.16%** ±0.13	96.04% ±0.24	96.01% ±0.25	96.08% ±0.22
	86.22% ±0.22	86.17% ±0.17	86.46% ±0.12	86.44% ±0.15	86.47% ±0.17
4	94.16% ±0.64	**95.07%** ±0.08	94.56% ±0.18	94.57% ±0.24	94.83% ±0.19
	77.86% ±0.75	71.25% ±0.43	72.54% ±0.21	72.18% ±0.22	71.64% ±0.24
6	92.26% ±0.65	**95.16%** ±0.48	92.27% ±0.36	92.55% ±0.26	93.52% ±0.24
	64.12% ±1.27	0.00% ±0.00	58.44% ±0.18	57.36% ±0.27	51.77% ±0.58
8	91.06% ±0.69	**94.48%** ±0.18	89.59% ±0.48	90.16% ±0.36	92.43% ±1.33
	56.72% ±0.74	0.00% ±0.00	45.64% ±0.21	43.88% ±0.16	35.96% ±2.78
10	90.76% ±1.21	**93.82%** ±0.28	86.78% ±0.88	87.26% ±0.73	90.36% ±0.33
	45.46% ±1.04	0.00% ±0.00	33.98% ±0.48	32.88% ±0.36	21.36% ±0.37
12	90.02% ±0.38	**92.72%** ±0.56	81.49% ±1.66	84.12% ±0.44	88.11% ±0.47
	36.88% ±1.09	0.00% ±0.00	26.17% ±0.88	23.64% ±0.42	14.16% ±0.38

Table 4: Comparative Analysis of Fast Adversarial Training Methods on CIFAR-10 Dataset

CIFAR-10 WRN-28-10 PGD-50-10					
$\epsilon \cdot 255$	l^p-FGSM	RS-FGSM	N-FGSM	GradAlign	ZeroGrad
2	91.12% ±0.52	**92.86%** ±0.14	92.49% ±0.14	92.54% ±0.13	92.62% ±0.16
	80.84% ±0.25	**80.91%** ±0.14	81.42% ±0.34	81.32% ±0.43	81.41% ±0.32
4	88.07% ±0.34	**90.74%** ±0.23	89.64% ±0.23	89.93% ±0.34	90.21% ±0.22
	69.62% ±0.84	68.24% ±0.19	69.10% ±0.27	69.80% ±0.48	69.21% ±0.21
6	83.23% ±0.46	**88.25%** ±0.22	85.74% ±0.32	86.94% ±0.16	86.11% ±0.45
	59.24% ±0.51	57.24% ±0.19	58.26% ±0.18	59.14% ±0.16	58.44% ±0.19
8	81.67% ±0.61	83.61% ±1.77	81.64% ±0.35	82.16% ±0.21	**84.16%** ±0.21
	51.31% ±0.59	0.00% ±0.00	49.51% ±0.27	50.12% ±0.17	48.32% ±0.21
10	76.61% ±0.58	**82.17%** ±1.48	76.94% ±0.12	79.42% ±0.28	81.29% ±0.73
	45.87% ±0.68	0.00% ±0.00	42.39% ±0.39	41.42% ±0.52	36.18% ±0.19
12	72.84% ±0.54	78.64% ±0.74	72.18% ±0.17	73.72% ±0.82	**79.33%** ±0.92
	41.09% ±1.24	0.00% ±0.00	36.82% ±0.27	35.16% ±0.77	28.26% ±1.81
14	66.58% ±0.63	73.27% ±2.84	67.86% ±0.46	66.41% ±0.52	**78.18%** ±0.66
	38.65% ±0.81	0.00% ±0.00	31.68% ±0.68	30.85% ±0.34	18.56% ±0.35
16	63.84% ±0.76	68.68% ±2.43	56.75% ±0.44	57.88% ±0.74	**75.43%** ±0.89
	37.16% ±1.22	0.00% ±0.00	25.11% ±0.43	26.24% ±0.43	14.66% ±0.22

Table 5: Comparative Analysis of Fast Adversarial Training Methods on CIFAR-100 Dataset

CIFAR-100 WRN-28-10 PGD-50-10					
$\epsilon \cdot 255$	l^p-FGSM	RS-FGSM	N-FGSM	GradAlign	ZeroGrad
2	66.42% ±0.15	**72.62%** ±0.24	71.52% ±0.14	71.61% ±0.23	71.64% ±0.22
	55.29% ±0.64	51.62% ±0.56	52.24% ±0.35	51.51% ±0.48	52.63% ±0.64
4	61.32% ±0.34	**68.27%** ±0.21	66.51% ±0.48	67.09% ±0.19	67.21% ±0.18
	45.73% ±0.46	39.56% ±0.14	39.96% ±0.31	39.81% ±0.48	39.61% ±0.32
6	58.79% ±0.45	**65.62%** ±0.66	61.42% ±0.63	62.86% ±0.10	63.65% ±0.12
	38.33% ±0.54	26.61% ±2.79	30.99% ±0.27	32.11% ±0.24	30.28% ±0.51
8	53.46% ±0.58	54.28% ±5.92	56.42% ±0.65	58.55% ±0.41	**60.78%** ±0.24
	32.41% ±1.18	0.00% ±0.00	26.71% ±0.68	26.97% ±0.61	23.72% ±0.16
10	50.23% ±0.42	46.18% ±4.88	51.51% ±0.61	53.85% ±0.73	**61.11%** ±0.39
	27.12% ±0.76	0.00% ±0.00	23.11% ±0.49	22.64% ±0.61	15.15% ±0.45
12	47.23% ±0.28	35.86% ±0.27	46.42% ±0.56	46.94% ±0.86	**58.36%** ±0.15
	24.74% ±0.67	0.00% ±0.00	19.32% ±0.51	19.94% ±0.65	11.12% ±0.66
14	43.18% ±0.25	24.42% ±1.38	42.14% ±0.36	42.63% ±0.50	**56.24%** ±0.16
	22.32% ±1.13	0.00% ±0.00	16.62% ±0.44	16.96% ±0.14	8.81% ±0.34
16	40.56% ±1.64	21.47% ±5.21	38.37% ±0.48	36.17% ±0.45	**56.42%** ±0.29
	18.41% ±1.42	0.00% ±0.00	14.29% ±0.38	14.23% ±0.26	4.92% ±0.38

L Appendix: Effects of ε-Softening and Noise Injection

We investigate two key components of our l^p-FGSM framework: the ε-softening term from Algorithm 1 and the integration of random noise.

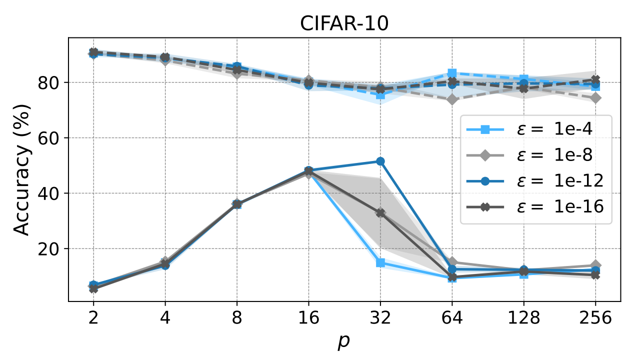
The ε-softening term, introduced to maintain Lipschitz continuity in our fixed-point formulation, helps numerical stability by avoiding zero division. Furthermore, there is a contrast with ZeroGrad [12] that nullifies small gradient components, while our softening ensures gradients maintain minimal non-zero values.

The theoretical motivation behind ε-softening stems from the observation that the fixed-point mapping's contractiveness is particularly sensitive near zero-gradient regions. By introducing a small, non-zero floor to gradient magnitudes, we maintain the desirable theoretical properties of our fixed-point formulation while improving numerical stability [11, 35].

For noise integration, following [10], we can employ a dual-purpose strategy where noise can either serve as input augmentation or initialization for perturbation crafting:

$$\begin{cases} x_0 \leftarrow x_0 + \eta, \; \eta \sim \mathcal{U}[-\epsilon, \epsilon] \\ \delta_0 \leftarrow \Pi_{\partial B_p(\epsilon)}(\eta) \end{cases} \tag{92}$$

These two noise placement approaches can be used independently. The random initialization at boundary $\partial B_p(\epsilon)$ particularly helps when gradient information is near zero. Our implementation differs from previous approaches in two key aspects: first, we project the noise onto the l^p ball boundary rather than using uniform sampling, and second, we reuse the same noise vector for both input augmentation and initialization, reducing computational overhead [36].



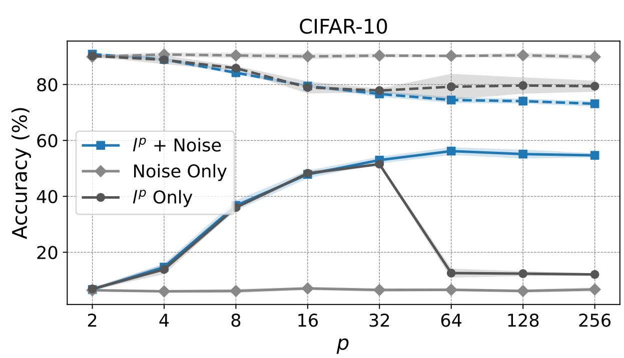


Figure 10: Analysis of ε-softening and noise effects on CIFAR-10 using WideResNet-28-10 against PGD-50 ($\epsilon = 8/255$). Top: Effect of ε-softening on clean (dashed) and adversarial (solid) accuracy for various p values. Optimal ε enhances stability against CO. Bottom: Synergistic effects of noise injection showing improved robustness against CO and enhanced overall accuracy. The results demonstrate that both components contribute significantly to preventing catastrophic overfitting while maintaining competitive performance.

Even though the main paper does not use any noise, the synergistic relationship between ε-softening and noise injection becomes apparent in their complementary effects on training stability. While ε-softening provides consistent gradient behavior, noise injection helps explore the loss landscape more effectively [34]. This combination proves particularly effective in preventing the gradient collapse often associated with CO [11].

Our extensive experiments on CIFAR-10 with WideResNet-28-10 (Figure 10) demonstrate that both components contribute meaningfully to the algorithm's performance. The ε-softening exhibits an optimal range where it enhances stability without compromising accuracy, while noise injection provides complementary benefits in preventing CO and improving overall robustness.

Notably, we observe that the combination of these techniques allows for more aggressive training schedules than previously possible [10, 37], achieving faster convergence while maintaining robustness. These findings suggest promising directions for future research in stabilizing adversarial training in conjunction with our adaptive l^p-FGSM.

M Appendix: Entropy Gap and $\mathtt{PR}_1$ for l^∞ vs l^p

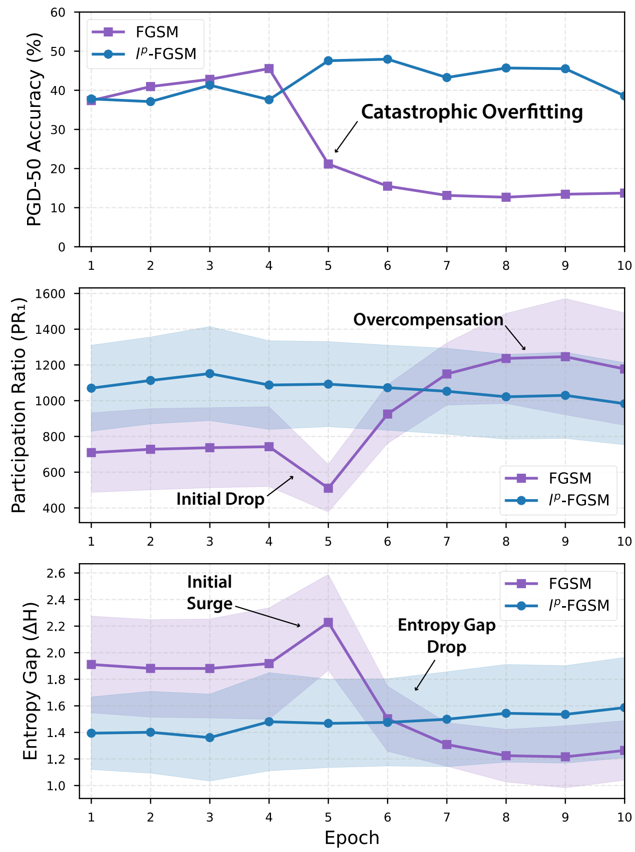


Figure 11: Evolution of Participation Ratios ($\mathtt{PR}_1$) and entropy gap during training with and without l^p-FGSM. Sharp patterns in these metrics align with the onset of Catastrophic Overfitting (CO), highlighting the link between gradient concentration and adversarial vulnerability. Same experimental setting as Figure 4.

Our preliminary analysis suggests that gradient concentration metrics (Participation Ratio and entropy gap) exhibit notable changes that appear to coincide with the onset of Catastrophic Overfitting. As shown in Figure 11, these metrics display an interesting pattern that warrants further investigation: a moderate increase, followed by a drop, and then what appears to be a compensatory response. While more extensive experimentation is needed to fully validate these observations, the pattern is consistent across multiple experimental runs.

The adaptation of Participation Ratio ($\mathtt{PR}$) from quantum mechanics [21, 22] to the adversarial training context as $\mathtt{PR}_1$ represents a novel approach to quantifying gradient behavior. In quantum systems, PR measures the effective number of states occupied by an electron; similarly, our $\mathtt{PR}_1$ aims to capture the effective dimensionality of gradient information. The entropy gap metric offers a complementary perspective, potentially providing insights into how information is distributed across gradient dimensions.

The observed pattern—initial increase, decline, and subsequent adjustment—may offer preliminary insights into the dynamics preceding CO. This behavior could potentially reflect the model's changing gradient geometry as it negotiates the complex loss landscape during adversarial training. The initial increase in both $\mathtt{PR}_1$ and entropy gap might suggest a temporary distribution of gradient information before concentration occurs.

By leveraging these metrics during training, our adaptive norm selection approach aims to detect potential instabilities and adjust accordingly. While our current results are promising, we acknowledge that the full relationship between these information-theoretic measures and adversarial robustness requires deeper exploration.

These initial findings provide support for our theoretical framework connecting gradient geometry to norm selection, suggesting that the l^p-FGSM approach may effectively mitigate CO without requiring additional techniques like gradient alignment or noise injection. Future work could explore these connections more thoroughly, potentially yielding broader insights into neural network behavior under adversarial constraints.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our main claim that adaptive l^p-FGSM prevents catastrophic overfitting through dynamic norm selection based on gradient concentration, which is supported by our theoretical analysis and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations including the local convexity assumption (Appendix A), convergence conditions (Section 3), and the need for broader evaluation across architectures in future work (conclusion).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results (Propositions 1-2, Lemmas 1-2) include complete assumptions and proofs in Appendices B-H, with the local convexity assumption clearly stated in Section 3 and Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed experimental settings including architectures, optimizers, learning rates, perturbation radii, and hyperparameters (Figure 5 caption, Section 4, Appendices). Algorithm 1 provides the complete implementation details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: While we plan to release our implementation publicly upon acceptance, we are unable to attach code with the current submission. Our algorithm is fully specified in Algorithm 1 with all hyperparameters detailed in Section 4 and appendices. All datasets used (CIFAR-10/100, SVHN, ImageNet) are standard publicly available benchmarks. The paper provides sufficient mathematical detail and pseudocode to enable implementation by other researchers.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are provided including optimizers (SGD/ADAM), learning rates, architectures, perturbation radii, and dataset-specific hyperparameters in Section 4 and throughout the appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All experimental results in Tables 1-5 and throughout appendices report mean and standard deviation over multiple seeds (3-5 seeds as specified).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on a single NVIDIA A100 GPU. Training times range from approximately 2-4 hours for CIFAR experiments (30 epochs) to 8-10 hours for ImageNet experiments (15 epochs).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work on adversarial robustness aims to improve the security and reliability of machine learning systems, conforming to ethical guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Adversarial robustness research has positive impacts for security-critical applications (mentioned in introduction) but could potentially be misused to develop stronger attacks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper focuses on a training algorithm for adversarial robustness and does not release pre-trained models or new datasets that could be misused.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (CIFAR-10/100, SVHN, ImageNet) and architectures (WideResNet, ResNet) are properly cited with references to original papers.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new datasets or release pre-trained models; we only propose a new training algorithm.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve human subjects research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were used only for writing and editing assistance, not as part of the core methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.