A Training-Dynamics View of Catastrophic Overfitting: Understanding and Prevention

Anonymous Authors[1]

Abstract

We study the underlying mechanism of catastrophic overfitting, a phenomenon in which models overfit to weak adversarial examples and lose true robustness, from the perspective of training dynamics. Despite numerous studies investigating catastrophic overfitting, the fundamental cause of sudden robustness collapse remains poorly understood. In this study, we systematically analyze the dynamics of adversarial training and reveal that the rapid amplification of the *mixed Hessian* causes catastrophic overfitting. Based on this insight, we propose a novel KL divergence-based regularizer that stabilizes training dynamics and effectively prevents catastrophic overfitting. Remarkably, our method consistently matches or even surpasses the robustness of multi-step adversarial training, despite using single-step adversarial training. Furthermore, when combined with multi-step adversarial training, our regularizer yields additional robustness improvements, indicating that mixed Hessian stabilization is a general principle applicable beyond the single-step regime.

1. Introduction

Adversarial vulnerability (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2018) remains one of the most critical threats to the reliable deployment of deep neural networks. Adversarial attacks produce small, even imperceptible, perturbations to input data that cause models to predict incorrectly. To defend against such perturbations, Adversarial Training (AT) (Goodfellow et al., 2014; Madry et al., 2018) has emerged as one of the most effective defenses. By generating adversarial examples on-the-fly during training and optimizing model parameters on these examples, AT significantly improves robustness. However, the standard AT formulation proposed by Madry et al. (2018) requires solving a computationally expensive inner maximization problem.

To propose an efficient and effective AT, a line of research investigated efficient AT methods based on the *Fast Gradient Sign Method (FGSM)* (Wong et al., 2020; Shafahi et al., 2019). They investigated training a robust model on single-step adversarial examples. However, a critical failure mode, *"catastrophic overfitting"* emerges, robust accuracy against *Projected Gradient Descent (PGD)* (Madry et al., 2018) rapidly drops to 0%, whereas robust accuracy against FGSM (Goodfellow et al., 2014) strictly increases, indicating the model overfit to the weak FGSM adversaries constructed on-the-fly at each training step, rather than learning genuinely robust features.

Although numerous methods have been proposed to mitigate catastrophic overfitting (Andriushchenko and Flammarion, 2020; Kim et al., 2021; Li et al., 2022; de Jorge Aranda et al., 2022; Huang et al., 2023; Lin et al., 2023; Rocamora et al., 2024; Golgooni et al., 2021), the fundamental mechanism under its sudden emergence remains poorly understood. In particular, it is still unclear what drives the sudden collapse of robustness during training and how training dynamics trigger this collapse. This motivates us to investigate the following questions:

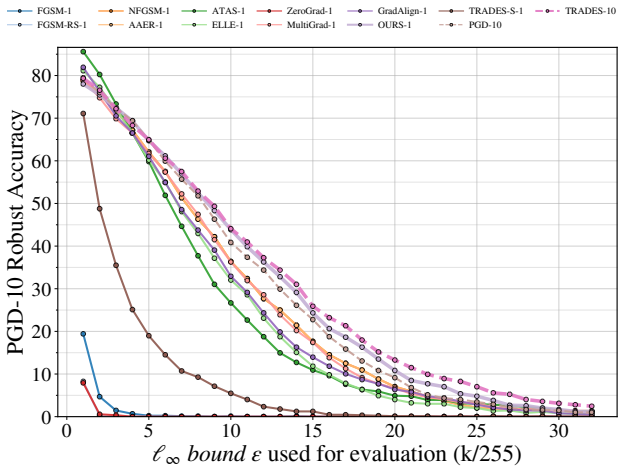
Which changes in training dynamics trigger catastrophic overfitting during single-step AT?

How can we efficiently prevent catastrophic overfitting and improve adversarial robustness?

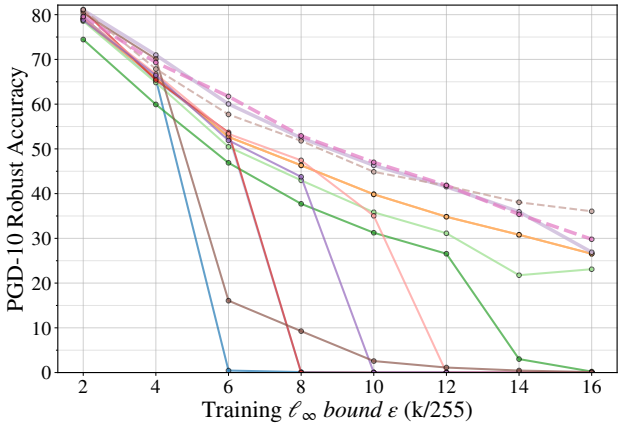
To investigate the underlying mechanism of catastrophic overfitting, we decompose several aspects of training dynamics. First, we analyze the *mixed Hessian*, which characterizes how input perturbations influence the parameter gradients. We show that catastrophic overfitting is fundamentally driven by the rapid amplification of the mixed Hessian spectral norm. This analysis departs from prior work that primarily examines input curvature (Andriushchenko and Flammarion, 2020; Rocamora et al., 2024), revealing that the *joint input-parameter* curvature, rather than input curvature alone, is the central quantity governing catastrophic

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) Robustness under increasing evaluation ϵ



(b) Robustness across diverse training ϵ

Figure 1. **PGD-10 robust accuracy and catastrophic overfitting across evaluation and training radii.** Solid lines denote single-step methods, while dashed lines denote multi-step methods. (a) We train each method with fixed perturbation radius ϵ as 8/255 and evaluate with various ϵ. The suffix $-n$ indicates the number of adversarial gradient steps used to generate perturbation. While several single-step methods suffer from catastrophic overfitting, our KL divergence-based regularizer effectively prevents catastrophic overfitting and consistently improves PGD robust accuracy over PGD-10 AT, while remaining comparable to TRADES, despite using single-step training. (b) We vary the training radius (from $2/255$ to $16/255$) and evaluate at the corresponding training radius. This setting directly tests whether each method remains stable across different training configurations; our method maintains strong robustness without catastrophic overfitting across the various radii.

overfitting.

Building on this insight, we introduce *three consistency views*—input curvature, loss-landscape sharpness, and joint input-parameter curvature—and show that a single joint KL divergence controls all three through its second-order expansion. We then propose a *KL divergence-based consistency regularizer* that stabilizes training dynamics by directly suppressing the joint curvature—and thereby the mixed Hessian—preventing catastrophic overfitting and improving robustness under strong adversarial attacks (see Figure 1). Notably, our regularizer also delivers additional robustness gains when applied on top of multi-step AT, suggesting that mixed Hessian stabilization captures a universal mechanism of robust training. We summarize our contributions as follows:

- We establish a theoretical connection between mixed Hessian amplification and parameter update misalignment, providing a principled explanation of catastrophic overfitting.
- We introduce a training dynamics perspective for catastrophic overfitting and present the first systematic decomposition of catastrophic overfitting dynamics, including mixed Hessian spectral growth, parameter update misalignment, and KL divergence-based curvature views.
- We propose a new regularizer that prevents catastrophic overfitting by stabilizing training dynamics and further improves robustness against various attacks, achieving performance comparable or even outperforming multi-step AT despite using single-step AT. Furthermore, our regularizer generalizes to the multi-step regime, providing additional robustness gains.

2. Related works

2.1. Adversarial robustness

Adversarial attack was first investigated by Szegedy et al. (2013) and was broadly investigated over a decade (Madry et al., 2018; Goodfellow et al., 2014; Lee et al., 2020; 2021; Schmidt et al., 2018; Tramer et al., 2018; Kim et al., 2020; Zhu et al., 2023; Li and Li). Despite this progress, training models that are reliably robust to adversarial perturbations remains an open problem. Among existing approaches, PGD-training (Madry et al., 2018) has emerged as one of the most reliable empirical directions. It is formulated as:

$$\min_{\theta} \left[\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(x+\delta, y; \theta) \right]. \tag{1}$$

However, solving the inner maximization problem is computationally expensive. This has motivated a line of work on more efficient AT methods (Shafahi et al., 2019; Wong et al., 2020). Although these methods substantially reduce training cost, they are known to suffer from *catastrophic overfitting*.

2.2. Catastrophic overfitting

Wong et al. (2020) observed that during single-step AT, robust accuracy against PGD (Madry et al., 2018) rapidly

drops to 0% over a few epochs, whereas accuracy against FGSM (Goodfellow et al., 2014) increases to nearly 100%. This phenomenon was termed *catastrophic overfitting*. To mitigate this issue, Wong et al. (2020) suggested adding a random step to FGSM and employing an early-stopping scheme. Similarly, de Jorge Aranda et al. (2022) also explored input randomization strategies. However, early-stopping does not fundamentally mitigate catastrophic overfitting. Moreover, Andriushchenko and Flammarion (2020) argued that the benefit of adding a random step mainly comes from reducing the effective perturbation magnitude.

Therefore, many lines of research investigated to mitigate catastrophic overfitting (Andriushchenko and Flammarion, 2020; Kim et al., 2021; Li et al., 2022; de Jorge Aranda et al., 2022; Huang et al., 2023; Lin et al., 2023; Rocamora et al., 2024; Golgooni et al., 2021). Among them, Andriushchenko and Flammarion (2020) proposed *"GradAlign"*, which regularizes gradient similarity between pairs of nearby points in the input space. Related gradient-based regularization strategies that explicitly aim to preserve local linearity have also been explored in subsequent work (Rocamora et al., 2024). However, existing approaches primarily regularize input space gradients, whereas our analysis focuses on *joint input-parameter curvature* through the *mixed Hessian dynamics.*

Another line of research, Kim et al. (2021) observed that the worst-case adversarial examples do not lie at the maximum perturbation magnitude. Based on this observation, they proposed *"FGSM-CKPT"*, performs a line search over perturbation magnitudes for each adversarial example. Related adaptive-perturbation strategies have been further explored in (Huang et al., 2023). Despite many lines of research eager to mitigate catastrophic overfitting, the inherent mechanism underlying catastrophic overfitting remains unclear and has not yet been fully explained.

3. Background

We consider an image classification model f_θ that maps an input $x \in \mathbb{R}^{d_x}$ to logits $z = f_\theta(x) \in \mathbb{R}^{|\mathcal{C}|}$, where $\mathcal{C}$ denotes the set of classes. We denote the predictive class probabilities by $p_\theta(y|x) = \mathrm{softmax}(f_\theta(x))_y$. Let $\ell(z, y)$ denote the cross-entropy loss applied to the logits. The overall loss can be written as $\mathcal{L}(x, y; \theta) = -\log p_\theta(y|x) = \ell(f_\theta(x), y)$. The model is trained using Stochastic Gradient Descent (SGD). We denote the parameter update at step t as $\Delta_t = \theta_{t+1} - \theta_t$.

We define three Hessian as follows:

$$H_{x\theta} = \frac{\partial^2 \mathcal{L}}{\partial \theta\, \partial x},\ H_{xx} = \frac{\partial^2 \mathcal{L}}{\partial x^2},\ H_{\theta\theta} = \frac{\partial^2 \mathcal{L}}{\partial \theta^2}. \tag{2}$$

The mixed Hessian $H_{x\theta}$ measures how the parameter gradient changes with respect to the input. We denote the evaluation point explicitly via its arguments. Unless otherwise specified, $H_{x\theta}$ is evaluated at the FGSM-perturbed point $(x + \delta_t, y; \theta_t)$, and we abbreviate $H_{x\theta}(x + \delta_t, y; \theta_t)$ as $H_{x\theta}$ for brevity.

In FGSM-training, a single-step perturbation is constructed as $\delta_t = \epsilon \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(x, y; \theta_t))$, where the subscript t denotes the training step. The model parameters are then updated by minimizing the loss at the perturbed input $x + \delta_t$. Accordingly, the parameter update direction is given by the gradient

$$\nabla_{\theta_t} \mathcal{L}(x + \delta_t, y; \theta_t), \tag{3}$$

up to a learning rate factor. In contrast, PGD-training employs a multi-step adversarial perturbation δ_t^*. Starting from $\delta_t^{(0)} = 0$, PGD iteratively updates $\delta_t^{(k+1)} = \Pi_{\|\delta\|_\infty \le \epsilon}(\delta_t^{(k)} + \alpha \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(x + \delta_t^{(k)}, y; \theta_t)))$, and we denote the final perturbation after K steps as $\delta_t^* = \delta_t^{(K)}$. The corresponding parameter update direction during PGD-training is given by the gradient

$$\nabla_{\theta_t} \mathcal{L}(x + \delta_t^*, y; \theta_t). \tag{4}$$

We denote $\sigma_1(A)$ as the spectral norm (the largest singular value) of a matrix A. Let $A = \sum_i \sigma_i u_i v_i^\top$ be its singular value decomposition, where u_i, v_i denotes the left, right singular vectors, respectively. In particular, $v_1(A)$ denotes the top singular vector.

4. Training dynamics and mixed Hessian

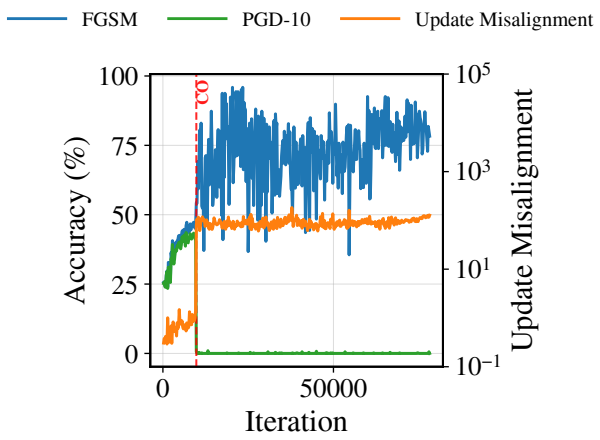
We first systematically analyze the training dynamics that drive catastrophic overfitting during single-step AT. Starting from a first-order analysis of the parameter update directions during FGSM/PGD-training, we show that the misalignment between these directions is governed by the mixed Hessian $H_{x\theta}$, and that the rapid spectral amplification of $H_{x\theta}$ is the primary driver of catastrophic overfitting.

4.1. Parameter update direction misalignment

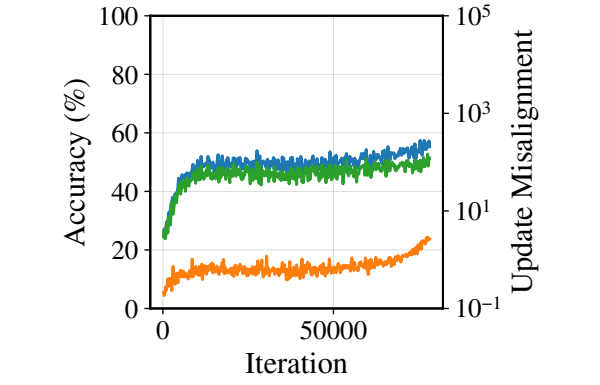
First, we investigate the parameter update direction misalignment. This misalignment between FGSM/PGD-training induced by Equations (3) and (4), is directly governed by the mixed Hessian $H_{x\theta}$:

$$\nabla_\theta \mathcal{L}(x + \delta_t, y; \theta_t) - \nabla_\theta \mathcal{L}(x + \delta_t^*, y; \theta_t) \ \approx\ H_{x\theta}(\delta_t - \delta_t^*). \tag{5}$$

Assume that PGD-training produces a worst-case adversarial perturbation δ^*. Robust optimization relies on alignment between the parameter update directions induced by FGSM and PGD perturbations. When this alignment collapses, **parameter updates are driven toward FGSM-specific adversarial directions instead of worst-case adversarial directions, leading to catastrophic overfitting.**



(a) FGSM-training



(b) PGD-training

Figure 2. **Parameter update misalignment leads to catastrophic overfitting.** FGSM/PGD-accuracy (blue/green line), and parameter update misalignment (orange line) during (a) FGSM and (b) PGD-training. Catastrophic overfitting emerges at iteration 9800 in (a), which is indicated as a red dashed vertical line (denoted as CO in the figure). Catastrophic overfitting coincides *with* parameter update misalignment.

Equation (5) reveals that this update direction misalignment is governed by two factors: (i) the perturbation misalignment $(\delta_t - \delta_t^*)$ and (ii) the mixed Hessian operator $H_{x\theta}$, which amplifies this misalignment. While prior works (Andriushchenko and Flammarion, 2020; Rocamora et al., 2024) primarily investigates the perturbation misalignment via input-space analysis—essentially proxies for the input Hessian H_{xx}—**our analysis identifies a fundamentally different object—the *mixed Hessian* $H_{x\theta}$**—and explicitly separates and measures both perturbation mismatch and its amplification, revealing their coupled role in catastrophic overfitting.

Figure 2 illustrates the training dynamics of FGSM/PGD-training. At the onset of catastrophic overfitting (Figure 2 (a) iteration 9800), the parameter update misalignment sharply increases. In contrast, the parameter update directions stay well-aligned during PGD-training. These observations realize the mechanism derived in Equation (5): **update misalignment leads to catastrophic overfitting.**

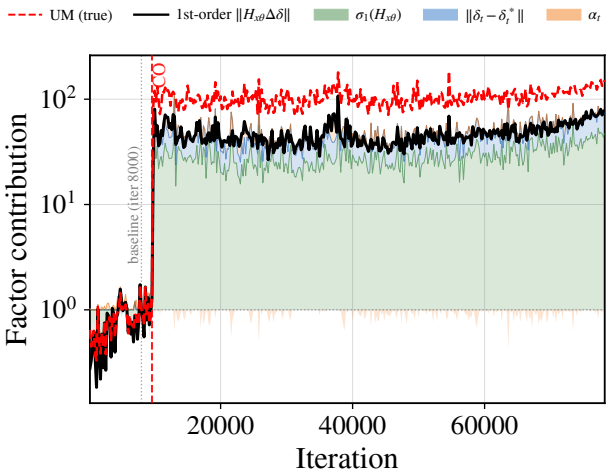


Figure 3. **Decomposition of catastrophic overfitting.** The y-axis shows the multiplicative ratio of each quantity relative to a baseline iteration just before catastrophic overfitting (gray dotted vertical line). By the multiplicative decomposition in Equation (7), the first-order term factorizes into $\sigma_1(H_{x\theta})$, $\|\delta_t - \delta_t^*\|$, and α_t, shown as stacked log-area bands. Among the three factors, $\sigma_1(H_{x\theta})$ amplifies by orders of magnitude near catastrophic overfitting, substantially exceeding the contributions of the other two factors, identifying it as the dominant factor of update misalignment.

4.2. Decomposition of catastrophic overfitting mechanism

In Equation (5), we show that the update misalignment is governed by the mixed Hessian $H_{x\theta}$ and the perturbation mismatch $\delta_t - \delta_t^*$. To identify the *dominant* factor driving catastrophic overfitting, we expand the update misalignment into its constituent components and empirically track each one during training. Taking norms of both sides of Equation (5) and applying the Taylor expansion explicitly:

$$\underbrace{\|\nabla_\theta \mathcal{L}(x+\delta_t) - \nabla_\theta \mathcal{L}(x+\delta_t^*)\|}_{\text{update misalignment (UM)}} = \underbrace{\|H_{x\theta}(\delta_t - \delta_t^*)\|}_{\text{first-order term}} + O(\|\delta_t - \delta_t^*\|^2), \tag{6}$$

where the first-order term admits the multiplicative decomposition

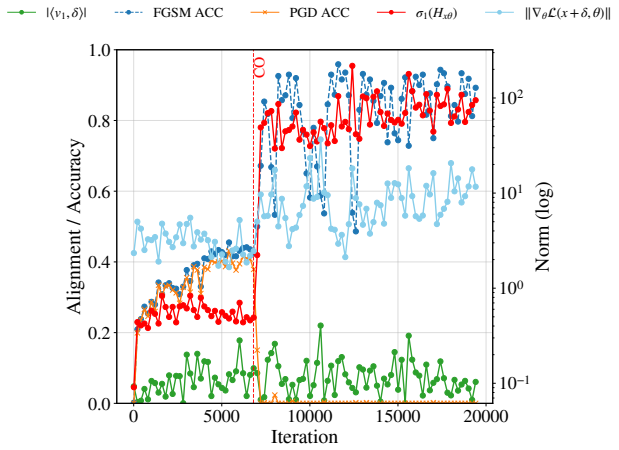
$$\|H_{x\theta}(\delta_t - \delta_t^*)\| = \alpha_t \cdot \sigma_1(H_{x\theta}) \cdot \|\delta_t - \delta_t^*\|, \tag{7}$$

with $\alpha_t \in [0, 1]$ denoting the alignment between $\delta_t - \delta_t^*$ and the top right singular vector of $H_{x\theta}$. The update misalignment thus decomposes into three multiplicative factors—$\sigma_1(H_{x\theta})$, $\|\delta_t - \delta_t^*\|$, and α_t—plus a higher-order residual.

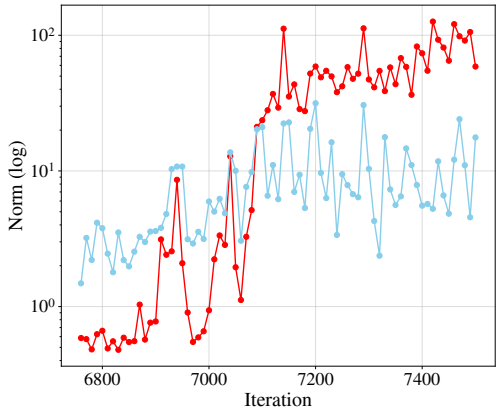
Figure 3 visualizes the resulting decomposition: the first-order term closely tracks the true update misalignment,

confirming the validity of the Taylor expansion. Among the multiplicative factors, $\sigma_1(H_{x\theta})$ amplifies by orders of magnitude near the onset of catastrophic overfitting, while $\|\delta_t - \delta_t^*\|$ and α_t grow at substantially smaller rates. The decomposition therefore identifies $\sigma_1(H_{x\theta})$ as the dominant factor of update misalignment, and consequently of catastrophic overfitting. We conclude that the **rapid spectral amplification of the mixed Hessian is the central driver of catastrophic overfitting** in single-step AT.

4.3. Why Hessian grows?



(a) Spectral norm dynamics



(b) Zoom around catastrophic overfitting

Figure 4. **Spectral norm and gradient dynamics during FGSM training.** (a) Training dynamics of $\sigma_1(H_{x\theta})$, singular vector alignment, and parameter gradient norm. (b) Zoom around the catastrophic overfitting point, showing that the amplification of $\sigma_1(H_{x\theta})$ precedes and coincides with the growth of the parameter gradient.

Figure 3 establishes that the amplification of $\sigma_1(H_{x\theta})$ is the dominant driver of update misalignment, and hence of catastrophic overfitting. A natural follow-up question is *why* $\sigma_1(H_{x\theta})$ amplifies during single-step AT. A full theoretical characterization is left as future work; here, we present a conjecture supported by empirical observations and a first-order analysis.

Conjecture: a progressive positive feedback between σ_1 and the parameter gradient. A first-order analysis of the mixed Hessian dynamics suggests

$$H_{x\theta}(\cdot;\theta_{t+1}) \approx H_{x\theta}(\cdot;\theta_t) - \eta\, \nabla_\theta H_{x\theta}(\cdot;\theta_t)\, \nabla_\theta \mathcal{L}(\cdot;\theta_t), \tag{8}$$

which yields the singular value update

$$\Delta\sigma_i \approx -\eta\, u_i^\top (\nabla_\theta H_{x\theta}(\,\cdot\,;\theta_t)\, \nabla_\theta \mathcal{L}(x+\delta_t, y; \theta_t)) v_i. \tag{9}$$

Furthermore, expanding the parameter gradient at the perturbed input to first order:

$$\nabla_\theta \mathcal{L}(x+\delta_t, y; \theta_t) \approx \nabla_\theta \mathcal{L}(x, y; \theta_t) + H_{x\theta}(x, y; \theta_t)\, \delta_t, \tag{10}$$

where $H_{x\theta}(x, y; \theta_t)\, \delta_t = \sum_i \sigma_i \langle v_i, \delta_t \rangle u_i$. Together, Equations (9) and (10) are consistent with a progressive positive feedback: an amplification of σ_1 enlarges the parameter gradient via Equation (10), which in turn further amplifies σ_1 via Equation (9).

Figure 4 provides empirical support: during FGSM training, $\sigma_1(H_{x\theta})$ amplifies sharply just prior to catastrophic overfitting, the alignment $\langle v_1, \delta_t \rangle$ remains relatively stable, and the parameter gradient norm grows alongside σ_1. These observations are consistent with the conjectured feedback loop. We therefore present this as a hypothesis rather than a derived result, and defer a rigorous analysis of σ_1 dynamics to future work.

Scope and future direction. Our analysis here characterizes *why* the mixed Hessian spectral norm amplifies during single-step (FGSM) AT, and how this amplification trigger catastrophic overfitting. A complementary question—why and how multi-step AT keeps $\sigma_1(H_{x\theta})$ controlled, as observed in Figure 2 (b)—remains beyond the scope of this work. Empirically, multi-step training stabilizes both the $\sigma_1(H_{x\theta})$ and parameter update alignment without explicit curvature regularization, which we hypothesize arises from the iterative recomputation of input gradients along the loss surface: each PGD step probes regions where the input gradient varies rapidly and thereby implicitly penalizes excessive mixed curvature. This is consistent with our discussion in Section 5.3. A formal characterization of this implicit stabilization mechanism is an interesting direction for future work. In this paper, we focus on how to achieve a comparable stabilization effect in the *single-step* regime through an explicit KL-based regularizer, which we propose in Section 5.

5. Stabilizing training dynamics through KL divergence consistency

In the previous section, we show that catastrophic overfitting is driven by the rapid amplification of the mixed

Hessian, and the perturbation misalignment. This suggests that preventing catastrophic overfitting requires controlling curvature not only in the input space, but also in the parameter space and their interaction. In this section, we introduce a KL divergence-based consistency regularizer that simultaneously suppresses three curvature components: (i) the mixed Hessian, which we identified in the previous section as the main driver of catastrophic overfitting, (ii) the input Hessian, which characterizes local linearity of the model in the input space, and (iii) the parameter Hessian, which reflects the sharpness of the loss-landscape.

5.1. KL divergence as a joint curvature regularizer

Define the predictive distribution $p_0(y) = p_\theta(y|x)$, $p_{\eta_x}(y) = p_\theta(y|x+\eta_x)$, $q_{\eta_\theta}(y) = p_{\theta+\eta_\theta}(y|x)$. We consider the KL divergence consistency term

$$\mathrm{KL}(p_{\eta_x} \,\|\, q_{\eta_\theta}) = \sum_y p_{\eta_x}(y) \log \frac{p_{\eta_x}(y)}{q_{\eta_\theta}(y)}. \tag{11}$$

Theorem 5.1 (KL divergence controls joint curvature)**.** *Assume that the predictive distribution $p_\theta(y|x)$ is positive and twice continuously differentiable. For sufficiently small (η_x, η_θ),*

$$\mathrm{KL}(p_{\eta_x} \,\|\, q_{\eta_\theta}) = \frac{1}{2}\eta^\top F(x,\theta)\eta + o(\|\eta\|^2), \quad \eta := [\eta_x; \eta_\theta]. \tag{12}$$

where

$$F(x,\theta) = \begin{bmatrix} F_{xx} & F_{x\theta} \\ F_{\theta x} & F_{\theta\theta} \end{bmatrix} \tag{13}$$

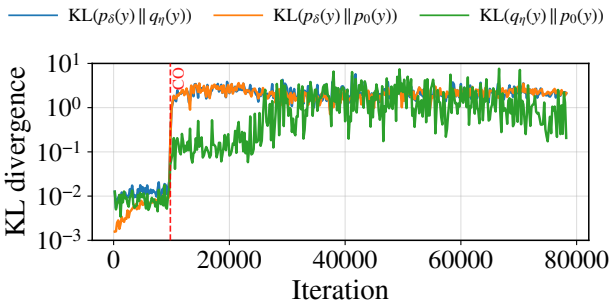
is the joint Fisher information matrix of the predictive distribution with respect to (x,θ). Under cross-entropy loss, each block equals the expected loss Hessian block:

$$F_{xx} = \mathbb{E}[\nabla^2_{xx}\mathcal{L}], \quad F_{\theta\theta} = \mathbb{E}[\nabla^2_{\theta\theta}\mathcal{L}], \quad F_{x\theta} = \mathbb{E}[\nabla^2_{x\theta}\mathcal{L}]. \tag{14}$$

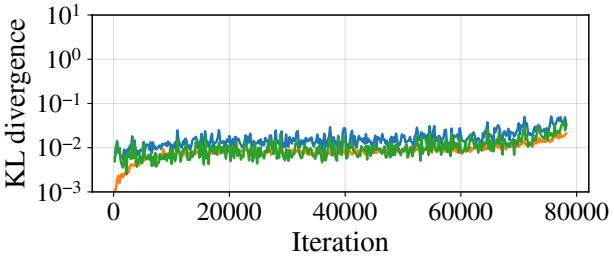
The proof is deferred to Appendix A. This result shows that minimizing a single KL divergence consistency term locally suppresses three curvature simultaneously.

5.2. Three consistency views and training dynamics

Building on Theorem 5.1, the KL divergence formulation in Equation (11) admits a natural decomposition into three consistency views. Specifically, the input consistency $\mathrm{KL}(p_{\eta_x}(y) \,\|\, p_0(y))$ has second-order form $\frac{1}{2}\eta_x^\top F_{xx}\eta_x$ and therefore penalizes input curvature, while the parameter consistency $\mathrm{KL}(q_{\eta_\theta}(y) \,\|\, p_0(y))$ yields $\frac{1}{2}\eta_\theta^\top F_{\theta\theta}\eta_\theta$ and corresponds to loss-landscape sharpness control. Finally, the mixed consistency $\mathrm{KL}(p_{\eta_x}(y) \,\|\, q_{\eta_\theta}(y))$ produces the full quadratic form containing the mixed block $F_{x\theta}$, which governs joint input-parameter curvature.



(a) FGSM-training



(b) PGD-training

Figure 5. **KL divergence dynamics during FGSM/PGD-training.** The input, parameter, and joint KL divergence terms—corresponding to input curvature, loss-landscape sharpness, and joint input-parameter curvature—grow rapidly near catastrophic overfitting under FGSM-training, while remaining comparatively stable under PGD-training.

We next examine how these KL divergence consistency views evolve during FGSM/PGD-training. Empirically, we observe that:

- input consistency increases as local input curvature grows
- parameter consistency increases with progressive loss sharpening
- mixed consistency amplifies sharply near catastrophic overfitting

Figure 5 shows the evolution of the three KL divergence consistency terms during training. All KL divergence measures increase during FGSM/PGD-training, which is consistent with progressive loss sharpening (Cohen et al., 2021; Jastrzębski et al., 2019) and input curvature growth (Andriushchenko and Flammarion, 2020). However, under FGSM-training, these quantities exhibit a rapid amplification near catastrophic overfitting, whereas PGD-training shows a much more stable increase. This behavior aligns with our previous analysis of mixed Hessian dynamics and suggests that controlling KL divergence-based consistency can help prevent curvature amplification.

5.3. Method: KL divergence-based curvature regularization

By Theorem 5.1, the mixed Hessian $H_{x\theta}$—identified in Section 4 as the primary driver of catastrophic overfitting—can be suppressed in expectation by minimizing the KL divergence $\mathrm{KL}(p_{\eta_x} \| q_{\eta_\theta})$. We therefore augment single-step adversarial training with a KL divergence consistency regularizer

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda \, \mathrm{KL}\big(p_{\eta_x}(y) \,\|\, q_{\eta_\theta}(y)\big), \tag{15}$$

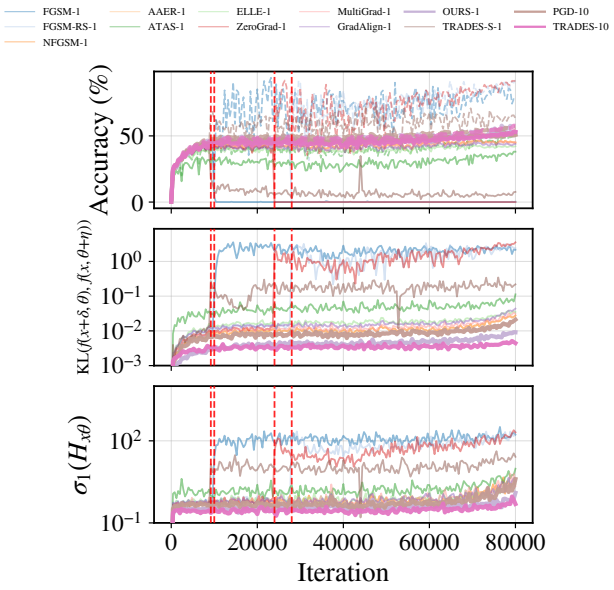
where η_x and η_θ are small perturbations sampled in the input and parameter space, respectively. Unlike prior works primarily control input sensitivity (Andriushchenko and Flammarion, 2020; Rocamora et al., 2024), this regularizer directly targets joint input-parameter curvature, which we identified as a key driver of catastrophic overfitting.

Difference between TRADES (Zhang et al., 2019). Our objective is related in form to TRADES:

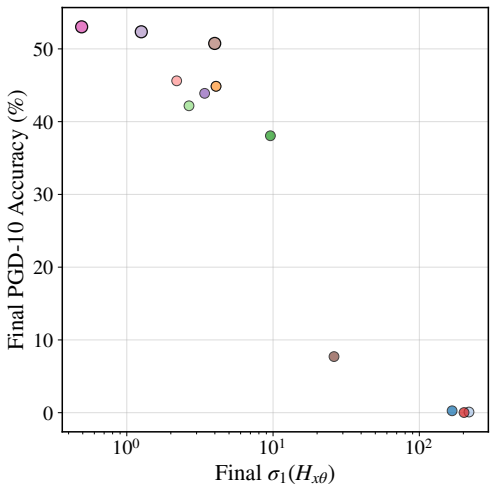
$$\min_{\theta} \mathbb{E}[\phi(p_0(y)) + \max_{\|\delta\|_\infty \le \epsilon} \phi(p_0(y), p_{\eta_x}(y))], \tag{16}$$

where ϕ is a non-negative loss function. However, TRADES enforces consistency under input perturbations only and does not explicitly regularize mixed consistency. Its success in preventing catastrophic overfitting comes from the multi-step inner maximization: when we replace it with a single-step perturbation (**TRADES-S**), catastrophic overfitting still occurs (Figure 6a, around iteration 9200), with amplification of $\sigma_1(H_{x\theta})$ and mixed consistency. This contrast shows that input-only consistency is insufficient in the single-step regime; without explicitly controlling joint input-parameter curvature, $\sigma_1(H_{x\theta})$ amplifies and ultimately triggers catastrophic overfitting.

Notably, although TRADES does not explicitly regularize mixed curvature, Figure 6a shows that it effectively stabilizes mixed consistency and $\sigma_1(H_{x\theta})$, thereby preventing catastrophic overfitting. We hypothesize that multi-step adversaries implicitly suppress mixed curvature: each PGD step recomputes input gradients along the loss surface, exposing regions where input perturbations strongly affect parameter gradients, and thereby penalizing excessive mixed curvature. A single-step adversary relies on a local first-order linearization and lacks this implicit signal. Our KL divergence-based regularizer explicitly regularize this, achieving comparable stabilization in the single-step regime.



(a) Training Dynamics



(b) Final $\sigma_1(H_{x\theta})$ and PGD Accuracy

Figure 6. **Mixed Hessian control determines robustness.** (a) Training dynamics of robust accuracy under PGD-10 (top), mixed consistency (middle), and $\sigma_1(H_{x\theta})$ (bottom). Catastrophic overfitting coincides with the amplification of mixed consistency and $\sigma_1(H_{x\theta})$ under FGSM, FGSM-RS, ZeroGrad, and even TRADES-S (see red dashed vertical lines). In contrast, methods that successfully prevent catastrophic overfitting consistently suppress the growth of both quantities. (b) Final $\sigma_1(H_{x\theta})$ at the end of training versus final PGD-10 robust accuracy across all methods. Methods with smaller final $\sigma_1(H_{x\theta})$ achieve higher PGD-10 robustness, indicating that stronger robustness is closely associated with better suppression of $\sigma_1(H_{x\theta})$. Together, these results show that controlling $\sigma_1(H_{x\theta})$ is critical for both preventing catastrophic overfitting and achieving strong adversarial robustness.

5.4. Experimental results

We now introduce our experimental results. Detailed experimental setup, implementation details, and additional experiments, which contains different architecture generalization (WideResNet, ViT (Dosovitskiy et al.)) are reported in Appendix B, C and E, respectively.

Table 1. Clean, FGSM, PGD, AutoAttack robust accuracy (%), training time on CIFAR-10, CIFAR-100 with ResNet-18 across various methods. The best PGD and AutoAttack robustness within single-step methods and within multi-step methods are shown in bold. FGSM robustness is not indicative of true robustness, as catastrophic overfitting can yield artificially high FGSM robustness while PGD robustness collapses.

Type	Method	CIFAR-10				CIFAR-100				Time (h)
		Clean	FGSM	PGD	AutoAttack	Clean	FGSM	PGD	AutoAttack	
Single-Step	FGSM	87.50	91.74	0.11	0.00	58.37	58.93	0.11	0.00	1.03
	FGSM-RS	83.15	94.87	0.00	0.00	49.33	66.63	0.00	0.00	1.00
	GradAlign	86.38	54.24	43.75	38.90	57.81	26.67	19.87	17.05	3.46
	ZeroGrad	78.24	92.75	0.00	0.00	60.16	75.00	0.00	0.00	1.14
	MultiGrad	83.59	55.69	47.43	40.37	54.02	25.78	25.54	17.84	2.09
	NFGSM	81.92	55.58	46.32	40.70	52.79	24.44	20.76	18.53	1.00
	ATAS	89.62	53.13	37.72	33.04	62.83	24.88	18.42	14.73	1.29
	AAER	81.92	55.58	46.32	40.72	52.79	24.44	20.76	18.54	1.17
	ELLE	85.16	53.46	42.97	37.75	57.92	25.11	20.76	16.22	2.63
	TRADES-S	91.96	59.82	9.26	0.11	70.76	30.25	5.92	0.49	2.19
	Ours	81.14	57.03	**52.46**	**45.08**	49.11	29.13	**26.56**	**22.41**	2.11
Multi-Step	PGD-10	82.25	56.92	51.79	46.09	54.46	27.79	25.45	20.79	5.99
	PGD-10+Ours	80.80	56.36	**53.68**	**46.57**	49.67	29.91	**27.90**	**23.65**	7.06
	TRADES	82.59	58.04	52.90	48.38	53.24	30.36	28.01	23.33	6.87
	TRADES+Ours	80.13	58.15	**54.02**	**48.51**	53.35	32.14	**30.25**	**24.13**	8.29

Table 1 shows that multi-step AT prevents catastrophic overfitting, whereas single-step AT such as FGSM (Goodfellow et al., 2014), FGSM-RS (Wong et al., 2020), ZeroGrad (Golgooni et al., 2021), and TRADES-S (which we discussed in Section 5.3) exhibit catastrophic overfitting. Consistent with Figure 6, these methods fail to control the growth of mixed consistency and $\sigma_1(H_{x\theta})$.

In contrast, in Table 1, several single-step variants such as GradAlign (Andriushchenko and Flammarion, 2020), ATAS (Huang et al., 2023), AAER (Lin et al., 2023), NFGSM (de Jorge Aranda et al., 2022), and ELLE (Rocamora et al., 2024) successfully prevent catastrophic overfitting. As illustrated in Figure 6, these methods implicitly stabilize mixed consistency and joint curvature during training. This observation suggests that preventing catastrophic overfitting fundamentally requires controlling the joint input-parameter curvature, even when not explicitly formulated.

More importantly, the role of joint curvature extends beyond merely preventing catastrophic overfitting. As reported in Table 1, methods that effectively regulate mixed consistency and $\sigma_1(H_{x\theta})$ achieve stronger robustness under AutoAttack (Croce and Hein, 2020). When considered alongside the training dynamics in Figure 6, these results indicate that controlling mixed curvature is not only essential for preventing catastrophic overfitting, but also closely linked to robustness under strong adversarial attacks.

In particular, our proposed KL divergence-based regularizer consistently suppresses the growth of mixed curvature across training. This stabilization translates into both stable training dynamics and competitive robustness under strong multi-step attacks. Furthermore, as shown in Table 1, our method achieves this robustness improvement with substantially lower computational overhead, compared to multi-step AT.

6. Conclusion

We identify the fundamental mechanism underlying catastrophic overfitting. Our analysis reveals that catastrophic overfitting is driven by the rapid amplification of the spectral norm of the mixed Hessian, which induces severe parameter update direction misalignment and destabilizes training.

Contrary to prevailing explanations based solely on gradient misalignment, we show that the fundamental cause lies in uncontrolled curvature growth in the mixed input-parameter space. As a result, existing single-step adversarial training methods often fail to prevent catastrophic overfitting because they do not explicitly regularize mixed curvature or mixed consistency.

Building on this insight, we propose a KL divergence-based regularizer that directly stabilizes joint curvature dynamics. Extensive experiments demonstrate that our approach consistently prevents catastrophic overfitting across various settings (see Figure 1), while achieving competitive or superior robustness compared to multi-step adversarial training (see Table 1 and Figure 6b). Our findings highlight joint curvature as the key driver of catastrophic overfitting and provide a new perspective on designing robust adversarial training algorithms.

References

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

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

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.

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

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.

Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13409–13418, 2022.

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.

Zhichao Huang, Yanbo Fan, Chen Liu, Weizhong Zhang, Yong Zhang, Mathieu Salzmann, Sabine Süsstrunk, and Jue Wang. Fast adversarial training with adaptive step size. *IEEE Transactions on Image Processing*, 32:6102–6114, 2023.

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.

Elias Abad Rocamora, Fanghui Liu, Grigorios Chrysos, Pablo M Olmos, and Volkan Cevher. Efficient local linearity regularization to overcome catastrophic overfitting. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Sungyoon Lee, Jaewook Lee, and Saerom Park. Lipschitz-certifiable training with a tight outer bound. *Advances in Neural Information Processing Systems*, 33:16891–16902, 2020.

Sungyoon Lee, Woojin Lee, Jinseong Park, and Jaewook Lee. Towards better understanding of training certifiably robust models against adversarial examples. *Advances in Neural Information Processing Systems*, 34:953–964, 2021.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *stat*, 1050:22, 2018.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in neural information processing systems*, 33:2983–2994, 2020.

Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4424–4434, 2023.

Binghui Li and Yuanzhi Li. Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data. In *The Thirteenth International Conference on Learning Representations*.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. In *International Conference on Learning Representations*, 2019.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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.

Supplementary Material for A Training-Dynamics View of Catastrophic Overfitting: Understanding and Prevention

A. Proof of Theorem 5.1

Define

$p_0(y) = p_\theta(y \mid x)$, $p_{\eta_x}(y) = p_\theta(y \mid x+\eta_x)$, $q_{\eta_\theta}(y) = p_{\theta+\eta_\theta}(x)$.

Consider the consistency regularizer

$$\mathrm{KL}(p_{\eta_x} \| q_{\eta_\theta}) = \sum_y p_{\eta_x}(y) \log \frac{p_{\eta_x}(y)}{q_{\eta_\theta}(y)}. \tag{17}$$

We analyze its second-order behavior around $(\eta_x, \eta_\theta) = (0,0)$. We denote evaluation at $(\eta_x, \eta_\theta) = (0,0)$ by $\big|_{0,0}$. Assume $p_{\eta_x}(y \mid x)$ is positive and twice continuously differentiable.

First-order terms vanish. At $(\eta_x, \eta_\theta) = (0,0)$ we have KL$= 0$. Differentiating w.r.t. η_θ:

$$\nabla_{\eta_\theta} \mathrm{KL} = -\sum_y p_{\eta_x}(y) \nabla_{\eta_\theta} \log q_{\eta_\theta}(y). \tag{18}$$

At $(0,0)$:

$$\nabla_{\eta_\theta} \mathrm{KL}\big|_{0,0} = -\sum_y p_0(y) \nabla_\theta \log p_\theta(y \mid x). \tag{19}$$

$$\sum_y p_0(y) \nabla_\theta \log p_0(y) = \sum_y p_0(y) \frac{\nabla_\theta p_0(y)}{p_0(y)} \tag{20}$$

$$= \nabla_\theta \sum_y p_0(y) \tag{21}$$

$$= 0. \tag{22}$$

Hence all first-order terms vanish.

Second-order derivatives. Differentiating again:

$$\nabla^2_{\eta_\theta \eta_\theta} \mathrm{KL}\big|_{0,0} = -\sum_y p_0(y) \nabla^2_{\theta\theta} \log p_\theta(y \mid x) \tag{23}$$

and similarly

$$\nabla^2_{\eta_x \eta_x} \mathrm{KL}\big|_{0,0} = -\sum_y p_0(y) \nabla^2_{xx} \log p_\theta(y \mid x), \tag{24}$$

$$\nabla^2_{\eta_x \eta_\theta} \mathrm{KL}\big|_{0,0} = -\sum_y p_0(y) \nabla^2_{x\theta} \log p_\theta(y \mid x). \tag{25}$$

Quadratic expansion. Since $\mathrm{KL}(p_{\eta_x} \| q_{\eta_\theta})\big|_{0,0} = 0$ and all first-order terms vanish, the second-order Taylor expansion around $(\eta_x, \eta_\theta) = (0,0)$ yields

$$\mathrm{KL}(p_{\eta_x} \,\|\, q_{\eta_\theta}) = \frac{1}{2} \eta^\top F(x,\theta) \eta + o(\|\eta\|^2), \quad \eta := [\eta_x; \eta_\theta]. \tag{26}$$

Using the second derivatives derived above, the Hessian with respect to (η_x, η_θ) at $(0,0)$ can be written in block form as

$$\nabla^2_{(\eta_x,\eta_\theta)} \mathrm{KL}\big|_{0,0} = \begin{bmatrix} F_{xx} & F_{x\theta} \\ F_{\theta x} & F_{\theta\theta} \end{bmatrix}, \tag{27}$$

where

$$F_{xx} = -\mathbb{E}_{y \sim p_0} \left[\nabla^2_{xx} \log p_\theta(y \mid x) \right], \tag{28}$$

$$F_{\theta\theta} = -\mathbb{E}_{y \sim p_0} \left[\nabla^2_{\theta\theta} \log p_\theta(y \mid x) \right], \tag{29}$$

$$F_{x\theta} = -\mathbb{E}_{y \sim p_0} \left[\nabla^2_{x\theta} \log p_\theta(y \mid x) \right]. \tag{30}$$

Relation to Fisher information and loss Hessians. The block matrix in (27) is precisely the (joint) Fisher information of the predictive distribution with respect to (x, θ), since under standard regularity conditions,

$$-\mathbb{E}_{y \sim p_0} \left[\nabla^2 \log p_\theta(y|x) \right] = \mathbb{E}_{y \sim p_0} [\nabla \log p_\theta(y|x) \, \nabla \log p_\theta(y|x)^\top]. \tag{31}$$

Moreover, for cross-entropy loss $\mathcal{L}(x, y; \theta) = -\log p_\theta(y \mid x)$, we have the exact identity

$$\nabla^2_{x\theta} \mathcal{L}(x, y; \theta) = -\nabla^2_{x\theta} \log p_\theta(y \mid x), \tag{32}$$

$$\nabla^2_{xx} \mathcal{L}(x, y; \theta) = -\nabla^2_{xx} \log p_\theta(y \mid x), \tag{33}$$

(and similarly for $\nabla^2_{\theta\theta}$), so each Fisher block coincides with the *expected* loss-Hessian block:

$$F_{x\theta} = \mathbb{E}[\nabla^2_{x\theta} \mathcal{L}], \quad F_{xx} = \mathbb{E}[\nabla^2_{xx} \mathcal{L}], \quad F_{\theta\theta} = \mathbb{E}[\nabla^2_{\theta\theta} \mathcal{L}]. \tag{34}$$

Therefore minimizing

$$\mathrm{KL}(p_{\eta_x}(y) \,\|\, q_{\eta_\theta}(y)) \tag{35}$$

locally minimizes the Fisher / expected Hessian blocks $H_{xx}, H_{x\theta}, H_{\theta\theta}$ simultaneously, which suppresses joint curvature and stabilizes gradient dynamics, helping prevent catastrophic overfitting.

B. Experimental Settings

We conduct a set of experiments on CIFAR-10 and CIFAR-100 with ResNet-18 at Section 5.4 (additional experiments are reported in Appendix E). During training, we set the maximum perturbation magnitude ϵ as 8/255 (see Figure 1 for experiments with varying ϵ). We evaluate train-time PGD-10/FGSM accuracy every 200 iterations. At the end of training, we evaluate robustness on the test set using FGSM, PGD-10, and AutoAttack, proposed by Croce and Hein (2020), to assess robustness under stronger attacks. We compare our method with several representative single-step and multi-step adversarial training methods. We use a SGD with cyclic learning rate scheduling, momentum of 0.9 and weight decay of $5e-4$. To examine whether catastrophic overfitting emerges at later stages of training, we train for 200 epochs. We train all methods on single A100 GPU.

FGSM AT (Goodfellow et al., 2014) Fast Gradient Sign Method adversarial training uses a single-step perturbation

$$\delta = \epsilon \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(x, y; \theta)), \tag{36}$$

and optimizes $\mathcal{L}(x+\delta, y; \theta)$.

FGSM-RS (Wong et al., 2020) Fast adversarial training introduces a random start $\eta \sim \mathcal{U}(-\epsilon, \epsilon)^d$ before the FGSM update. We follow the recommended setting with step size $\alpha = 1.25\epsilon$ and project perturbation into ϵ ball.

GradAlign (Andriushchenko and Flammarion, 2020) GradAlign introduces a regularizer that maximizes the cosine similarity between input gradients at x and $x+\eta$, where $\eta \sim \mathcal{U}(-\epsilon, \epsilon)^d$:

$$\mathrm{GradAlign}(x, y, \theta) = \cos(\nabla_x \mathcal{L}(x, y; \theta), \nabla_x \mathcal{L}(x+\eta, y; \theta)). \tag{37}$$

The objective is $\mathcal{L} + \lambda(1 - \mathrm{GradAlign})$. We use $\lambda = 0.2$.

ZeroGrad and MultiGrad (Golgooni et al., 2021) ZeroGrad suppresses small components of the input gradient when generating perturbations. We follow the recommendation setting with $q = 0.35$ for CIFAR-10 and $q = 0.45$ for CIFAR-100, which is threshold for zero out gradient, and step size $\alpha = 2\epsilon$. MultiGrad aggregates gradients from N randomly initialized perturbations and retains directions with consistent signs. We use $N = 3$ and $\alpha = 2\epsilon$.

N-FGSM (de Jorge Aranda et al., 2022) N-FGSM increases the random initialization range and removes the final projection step. We follow the recommended setting with initialization $\eta \sim \mathcal{U}(-k, k)^d$ where $k = 2\epsilon$ and step size $\alpha = \epsilon$.

ATAS (Huang et al., 2023) ATAS adapts the step size based on a moving average of squared gradient norms

$$v_i^j = \beta v_i^{j-1} + (1-\beta)\|\nabla_x \mathcal{L}\|_2^2. \tag{38}$$

The step size is set to $\alpha_i^j = \frac{\gamma}{c + \sqrt{v_i^j}}$. We follow the recommended hyperparameters $\beta = 0.5$, $c = 0.01$, and $\gamma = 2c\epsilon$.

AAER (Lin et al., 2023) AAER identifies abnormal adversarial examples (AAEs) as a primary cause of catastrophic overfitting. AAEs are defined as adversarial examples whose loss is lower than that of their corresponding clean samples:

$$\begin{aligned} \text{AAE}: \ & \mathcal{L}(x+\eta, y) > \mathcal{L}(x+\eta+\delta, y), \\ \text{NAE}: \ & \mathcal{L}(x+\eta, y) \le \mathcal{L}(x+\eta+\delta, y). \end{aligned} \tag{39}$$

To mitigate the influence of AAEs, AAER introduces a regularization term that penalizes abnormal loss variations and logit discrepancies between AAEs and normal adversarial examples. The AAER regularizer is defined as

$$\begin{aligned} &\text{AAER} = \\ &\lambda_1 \frac{n}{m} \left(\lambda_2 \text{AAE-CE} + \lambda_3 \max(\text{AAE-L2} - \text{NAE-L2}, 0)\right), \end{aligned} \tag{40}$$

where m is the batch size and n is the number of AAEs. We adopt the N-AAER variant and follow the authors' recommendations with $\lambda_1 = 1$, $\lambda_2 = 1.5$, and $\lambda_3 = 0.15$.

ELLE (Rocamora et al., 2024) ELLE (Efficient Local Linearity Enforcement) encourages local linearity of the loss function by penalizing deviations from linear interpolation between adversarial samples. Two perturbed samples x_a and x_b are drawn as

$$x_a, x_b \sim x + \mathcal{U}(-\epsilon, \epsilon)^d, \quad \alpha \sim \mathcal{U}(0, 1), \tag{41}$$

and an interpolated point is formed as

$$x_c = (1-\alpha)x_a + \alpha x_b. \tag{42}$$

The ELLE regularizer penalizes deviations from linear interpolation of losses:

$$E_{\text{lin}} = \left(\mathcal{L}(x_c, y) - (1-\alpha)\mathcal{L}(x_a, y) - \alpha\mathcal{L}(x_b, y)\right)^2. \tag{43}$$

The overall objective augments the adversarial training loss with the ELLE penalty:

$$\min_\theta \ \mathcal{L}_{\text{adv}}(x, y; \theta) + \lambda E_{\text{lin}}(x, y; \theta), \tag{44}$$

so the parameter update includes $\nabla_\theta \mathcal{L}_{\text{adv}} + \lambda \nabla_\theta E_{\text{lin}}$. We follow the recommended setting $\lambda = 1000$.

Table 2. Effect of input perturbation choice in the KL regularizer on CIFAR-10 with ResNet18.

Method	Clean	FGSM	PGD-10
Ours (Random η_x)	79.69	53.57	45.42
Ours (FGSM δ/4)	81.14	57.03	52.46

PGD (Madry et al., 2018) Projected Gradient Descent (PGD) adversarial training optimizes the worst-case loss within an ℓ_∞ perturbation set:

$$\min_\theta \left[\max_\delta f(x+\delta, y; \theta)\right]. \tag{45}$$

The inner maximization is approximated by K steps of PGD:

$$\delta_t^{(k+1)} = \Pi_{\|\delta\|\le\epsilon}\Big(\delta_t^{(k)} + \alpha \cdot \mathrm{sgn}\big(\nabla_x \mathcal{L}(x+\delta_t^{(k)}, y; \theta_t)\big)\Big), \tag{46}$$

In our experiments we use $K = 10$ attack steps with step size $\alpha = 2\epsilon/10$.

TRADES (Zhang et al., 2019) TRADES improves the robustness-accuracy trade-off by decomposing the objective into a natural accuracy term and a robustness regularization term:

$$\min_\theta \mathbb{E}[\phi(p_0(y)) + \max_{\|\delta\|_\infty\le\epsilon} \phi(p_0(y), p_\delta(y))] \tag{47}$$

The adversarial perturbation is generated using PGD on the KL divergence term. We follow the standard setting with 10 attack steps and step size $\alpha = \epsilon/4$.

And for TRADES-S, we use one attack step and step size $\alpha = \epsilon$, which is consistent with single-step adversarial training.

C. KL-based Regularizer Experimental Setting

Our method introduces a KL-based regularization term that enforces joint input-parameter consistency:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda\, \mathrm{KL}\big(p_\delta(y) \,\|\, q_\eta(y)\big) \tag{48}$$

The KL regularization weight is fixed to $\lambda = 50$ in all experiments.

C.1. Input Perturbation η_x

Two variants of η_x are considered:

Random η_x. A random perturbation is sampled uniformly within the ℓ_∞ ball:

$$\eta_x \sim \mathcal{U}(-\epsilon', \epsilon')^d, \quad \|\eta_x\|_\infty \le \epsilon', \tag{49}$$

where $\epsilon' = 2/255$.

Table 3. Ablation study about parameter perturbation η_θ

η_θ	Clean	FGSM	PGD-10
10^{-10}	81.14	57.03	52.46
10^{-8}	81.03	56.58	51.56
10^{-6}	81.14	58.48	53.79
10^{-4}	81.03	57.81	53.01
10^{-2}	81.81	56.58	51.34

FGSM-based η_x. In the main experiments we instead use a scaled FGSM perturbation direction:

$$\eta_x = \frac{1}{4}\epsilon \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(x, y; \theta)). \tag{50}$$

This choice provides a more informative local perturbation direction while maintaining a small perturbation magnitude, and empirically yields stronger robustness compared to random perturbations (see Table 2).

C.2. Parameter perturbation η_θ.

Parameter perturbations are sampled from an isotropic Gaussian distribution:

$$\eta_\theta \sim \mathcal{N}(0, \sigma^2 I_d), \tag{51}$$

where d is the parameter dimension. In all experiments we set $\sigma = 10^{-10}$.

We additionally provide an ablation study over different perturbation scales in Table 3, showing that the proposed method remains effective across a broad range of σ values.

C.3. Effect of KL Regularization Weight λ

The KL regularization weight λ in Equation (15) controls the strength of joint curvature suppression. We fix $\lambda = 50$ in all main experiments. To validate this choice, we vary λ over {10, 50, 500} and report robustness on CIFAR-10 with ResNet-18.

Table 4. Effect of KL regularization weight λ on CIFAR-10 with ResNet-18.

λ	Clean	FGSM	PGD-10
10	81.70	56.70	49.00
50 (default)	81.14	57.03	52.46
500	68.97	51.90	51.00

As expected from the joint curvature interpretation (Theorem 5.1), excessively small λ fails to suppress mixed Hessian growth and leads to degraded PGD robustness, while overly large λ over-regularizes and reduces clean accuracy. $\lambda = 50$ provides a balanced trade-off across all evaluation metrics, and we adopt this value throughout the main experiments.

D. Robustness under Extended Training

A natural concern for any catastrophic-overfitting prevention method is whether the stabilization persists over substantially longer training. While the main experiments in Section 5.4 adopt the standard 200-epoch protocol, we additionally train our method for 1000 epochs on CIFAR-10 with ResNet-18, keeping all other hyperparameters identical.

Table 5. Final robust accuracy of our method under extended training on CIFAR-10 with ResNet-18.

Epochs	Clean	FGSM	PGD-10
200 (default)	81.14	57.03	52.46
1000	82.70	56.03	49.22

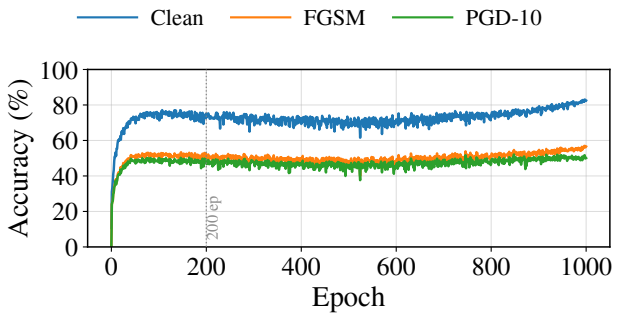


Figure 7. **Robust accuracy of our method throughout 1000-epoch training on CIFAR-10 with ResNet-18.** Clean, FGSM, and PGD-10 accuracy are evaluated periodically during training. The vertical dotted line marks the standard 200-epoch boundary (Section 5.4). Our method exhibits no robustness collapse across the entire extended training horizon, in contrast to single-step baselines such as FGSM, FGSM-RS, ZeroGrad, and TRADES-S, which already exhibit catastrophic overfitting within the first 200 epochs (Figure 6).

Table 5 and Figure 7 together show that our method maintains robust accuracy throughout extended training, with no signs of catastrophic overfitting. Figure 7 visualizes the full training trajectory: PGD-10 accuracy stays stable well past the 200-epoch boundary and through 1000 epochs, while Table 5 reports the final accuracies at both horizons. This stability is consistent with our analysis in Section 4: by directly suppressing mixed Hessian spectral growth via joint KL consistency, our regularizer prevents the positive feedback loop that drives catastrophic overfitting, regardless of training duration. We note that single-step methods which fail to control mixed consistency (e.g., FGSM, FGSM-RS, ZeroGrad, TRADES-S) already exhibit catastrophic overfitting well within 200 epochs (Figure 6), so the relevant question at 1000 epochs is whether our method remains stable—which it does.

E. Additional Experiments

E.1. Vision Transformer Architecture

To assess whether our method generalizes beyond convolutional architectures, we additionally evaluate it on a Vision Transformer. We use ViT-Small with patch size 4 on CIFAR-10, trained from scratch. All training settings are kept identical to our ResNet-18 main experiments, isolating the effect of architecture. We compare with FGSM (the canonical single-step baseline) and PGD-10 (the multi-step reference).

Table 6. Clean, FGSM, PGD-10 robust accuracy (%) on CIFAR-10 with ViT-Small. All methods share the same training protocol as our ResNet-18 experiments.

Method	Clean	FGSM	PGD-10
FGSM	52.79	35.38	34.38
Ours	46.88	32.59	32.59
PGD-10	53.12	35.71	34.71

Table 6 shows that our method continues to prevent catastrophic overfitting and achieve strong PGD robustness under the ViT architecture. This indicates that the joint curvature mechanism we identify in Section 4 is not specific to convolutional networks, and that controlling $H_{x\theta}$ via our KL regularizer generalizes across architectural families.

E.2. Extended Experiments

To further evaluate the generality of our method, we additionally conduct experiments on Tiny-ImageNet with ResNet-18 (see Table 7) and with a larger architecture, WideResNet-28 on CIFAR-10 and CIFAR-100 (see Table 8). Tables 7 and 8 confirm that our approach consistently stabilizes training dynamics and maintains strong robustness across different datasets and model scales.

Table 7. Clean, FGSM, PGD, AutoAttack accuracy (%) on Tiny ImageNet with ResNet-18 across various methods.

Method	TinyImageNet			
	Clean	FGSM	PGD	AA
FGSM	47.32	19.20	16.07	13.37
FGSM-RS	47.66	17.86	14.96	13.00
GradAlign	48.66	80.25	0.0	0.0
ZeroGrad	47.21	18.08	14.29	13.66
MultiGrad	47.77	19.31	15.18	14.16
NFGSM	46.76	20.65	16.96	14.56
ATAS	56.81	16.85	10.94	9.32
AAER	46.76	20.65	16.96	14.60
ELLE	50.45	19.42	15.85	12.51
TRADES-S	58.15	13.62	0.45	0.00
(Ours)	41.63	21.21	**19.08**	**16.49**
PGD-10	45.98	19.87	17.52	16.02
TRADES	46.43	22.54	**19.53**	**16.03**

Table 8. Clean, FGSM, PGD, AutoAttack accuracy (%) on CIFAR-10, CIFAR-100, Tiny ImageNet with WideResNet-28 across various methods

Method	CIFAR-10				CIFAR-100				TinyImageNet			
	Clean	FGSM	PGD	AA	Clean	FGSM	PGD	AA	Clean	FGSM	PGD	AA
FGSM	88.73	84.71	0.00	0.00	64.17	61.27	0.11	0.00	42.63	15.96	12.17	9.30
FGSM-RS	84.93	95.31	0.00	0.00	60.83	75.89	0.00	0.00	16.74	10.16	0.00	0.00
GradAlign	89.06	92.52	0.11	0.00	68.08	71.09	0.45	0.00	43.19	46.32	0.11	0.00
ZeroGrad	83.93	96.21	0.00	0.00	63.17	81.47	0.22	0.00	54.02	31.36	0.00	0.00
MultiGrad	86.50	55.25	43.75	39.87	58.82	26.67	21.76	19.33	51.67	22.21	16.41	13.86
NFGSM	84.82	56.70	46.54	41.90	55.36	27.34	23.10	20.10	50.56	20.87	16.85	15.22
ATAS	91.52	54.46	41.74	37.14	66.74	28.35	21.32	17.44	58.82	18.08	10.83	10.15
AAER	84.82	56.70	46.54	41.88	55.36	27.34	23.10	20.10	50.56	20.87	16.85	**15.23**
ELLE	71.21	42.30	37.83	31.78	50.56	23.21	21.76	17.58	38.84	14.96	13.50	11.36
TRADES-S	92.97	58.71	5.80	0.06	73.88	29.91	3.13	0.25	60.04	17.97	0.67	0.00
(Ours)	79.24	55.25	**51.34**	**44.69**	50.45	31.14	**28.57**	**23.71**	36.38	17.97	**17.19**	14.27
PGD-10	85.94	58.15	48.88	46.24	56.58	29.91	25.45	21.84	50.78	22.43	**17.52**	**15.78**
TRADES	84.49	58.71	**53.68**	**50.88**	56.03	33.26	**30.13**	**25.81**	48.10	24.55	16.74	11.21