

Adversarial Detection for Pre-trained Classifiers via Lipschitz-Margin Analysis with Certification

Rui Zhang*, Xinle Cao[†], Huazhen Yu[‡], Jian Liu[†], Youwen Zhu^{*§}

^{*}Nanjing University of Aeronautics and Astronautics

[†]The State Key Laboratory of Blockchain and Data Security, Zhejiang University

[‡]Nanjing University of Information Science and Technology [§]Corresponding author

Abstract—Deep classifiers deployed in safety-critical services like autonomous-driving perception remain vulnerable to small adversarial perturbations. Certified ℓ_2 defenses today face a costly trade-off: randomized smoothing (RS) requires hundreds of noisy queries per input, while Lipschitz networks (LN) constrain the architecture and sacrifice clean accuracy. We take a detection-oriented approach instead: rather than certifying that predictions are invariant inside an ℓ_2 ball, we certify that adversarial inputs within an ℓ_2 ball are *detected* — a guarantee aligned with security-critical deployments where missing an attack is costlier than flagging a clean input.

To realize this, we prove a Lipschitz-margin reduction from certified ℓ_2 robustness to certified detection, which reduces the problem to handling a single Lipschitz constant L . We introduce two detector families: $\mathbb{D}_{Adv}$, which estimates L from a single defender-side perturbed pair, and $\mathbb{D}_{lc}$, which augments that pair with a local margin centroid to keep the estimate above what the certificate requires. Both deploy on any pretrained classifier without retraining. With a valid global bound, the detector is fully certified at that radius; with a local estimate L' , certification is sample-wise and applies only when our certified condition holds.

Across MNIST, CIFAR-10, and ImageNet-1k on four backbones (LeNet-5, VGG16, ResNet-50, ViT-B/16), our detectors match or outperform five baselines (CTT, SortNet, CRM, ContraNet, A²D). For CIFAR-10 with 1-Lipschitz SortNet, our reduction delivers 100% certified detection at the lowest FPR among compared certified detectors; on the standard CIFAR-10 VGG16, our detectors reject all evaluated in-budget adversarial examples, with 62-82% both rejected and formally certified.

I. INTRODUCTION

Deep classifiers (DNNs) deployed in safety-critical services like autonomous-driving perception [1], content moderation, biometric authentication [2], and access control are vulnerable to *adversarial examples* [3], [4]: small, often imperceptible perturbations δ that flip the prediction $\arg \max f(x + \delta) \neq \arg \max f(x)$.

A large body of work [5]–[9] addresses this by building *certifiably robust* classifiers — whose prediction is provably unchanged inside an ℓ_p -ball around the input. Existing ℓ_2 certifications cluster into two families, each with a structural cost:

- **Randomized smoothing (RS)** [10]–[15] replaces f with a smoothed classifier whose output is a majority vote over Gaussian-noised inputs. Each prediction needs hundreds to thousands of forward passes, ruling out real-time use.
- **Lipschitz networks (LN)** [16]–[18] constrain the architecture so that a small global Lipschitz constant is provable by construction. The constraint sacrifices clean accuracy and forecloses use of standard pretrained classifiers.

An alternative, motivated by Shumailov et al. [19] and Tramer [20], is to defend by *detecting and rejecting* adversarial inputs: the deployed classifier is left unchanged, and a detector decides whether each query is safe to answer. This preserves clean accuracy and inference speed. Empirical detectors are well studied [21]–[27], but they offer no formal guarantee, and *certified detection* has received far less attention. Tramer [20] gave a theoretical reduction from robust classification to robust detection, but the construction is too expensive for deployment. We therefore ask: *Can a certified robust classifier be turned into a certified detector that is efficient enough for real-time use?*

Three challenges. Doing this in practice requires solving three coupled problems.

(C1) *Reduction.* Existing certified-robustness theorems guarantee *prediction invariance* inside a radius, not detectability of adversarial inputs outside it. Reducing detection to robustness requires a fresh argument that ties the margin of any in-budget adversarial example to its perturbation budget.

(C2) *Lipschitz constants are intractable.* Any Lipschitz-based reduction needs a value for the network’s Lipschitz constant L . Architecturally enforced bounds (e.g., 1-Lipschitz networks) cost clean accuracy; numerical bounds on standard pretrained networks are vacuous.

(C3) *Local estimators must be calibrated.* Replacing the unknown global L with a local estimator L' can occasionally yield L' below what the certificate requires, silently breaking the guarantee on those samples.

Our approach. An adversarial example with bounded budget cannot move far from the decision boundary, so the classifier’s margin function $q(\cdot)$ must be small there. Lipschitz continuity makes this quantitative: if f is L -Lipschitz, the change in margin under any perturbation δ is bounded by $cL\|\delta\|$, tying the adversarial margin $q(\hat{x})$ to the clean margin $q(x)$ and to $\|\delta\|$. Each input is certified by a Lipschitz value relevant to it, not a globally enforced one. Two practical

detectors follow: $\mathbb{D}_{Adv}$ replaces the unknown clean counterpart with a defender-side perturbed surrogate, and $\mathbb{D}_{lc}$ additionally anchors the bound to a local centroid of the margin function. The two detectors are tighter in different regimes (cf. Section VI-C): on Lipschitz-constrained networks $\mathbb{D}_{Adv}$ ’s short-pair certificate is naturally tight, while on unconstrained pretrained classifiers $\mathbb{D}_{lc}$ ’s centroid scales the bound with the model’s clean-margin distribution and certifies a substantially larger fraction. Both deploy on any pretrained classifier.

For (C1), we prove a Lipschitz-margin reduction (Lemma IV.1, Theorem IV.2): every successful in-budget adversarial example is provably detectable from L and the margin alone. The certificate is *attack-agnostic*: it applies to any in-budget ℓ_2 perturbation that flips the prediction, regardless of how it was generated. For (C2), on Lipschitz-constrained networks (SortNet [16], CRM [28]) the exact L_{glob} gives a fully certified detector via Lemma IV.1; on unconstrained pretrained classifiers, we replace L_{glob} with a local estimate L' , which yields sample-wise certificates when the local estimate is large enough for certification while allowing substantially lower FPR in practice. For (C3), we quantify this sample-wise guarantee with **cert%**, the fraction of successful in-budget adversarial examples that are both rejected and formally certified. Concretely, for each detected adversarial example, the certificate holds when $L' \geq L^*$, where L^* is the minimum Lipschitz value required by Proposition IV.6 on that sample. To increase this certified subset, we can use a more conservative multi-probe estimate $L'_{strong} = \max_k L'_k$: it deliberately raises L' , making $L' \geq L^*$ easier to satisfy, but the same larger L' also enlarges the rejection threshold, increasing FPR and runtime.

What our guarantee covers. We distinguish three protection levels. *Fully certified detection* applies when a valid Lipschitz bound is known for all inputs, so every successful adversarial example within the declared radius is rejected; the cost is that a conservative global bound can also reject many clean inputs, i.e., high FPR. *Conditionally certified detection* applies on standard pretrained models after replacing the intractable global bound by a local estimate L' : this gives a lower-FPR operating point, but only the subset satisfying the per-sample certified condition is formally certified. The stronger L' variant above deliberately moves along this frontier, trading FPR and runtime for a larger certified subset. *Empirical protection* describes additional successful attacks that are rejected in experiments but do not satisfy this per-sample certificate. We therefore report TPR, FPR, and **cert%** together: the same detector may reject nearly all measured attacks, while only a subset of those rejections are formally certified.

In summary, our contributions are as follows.

- A **Lipschitz-margin reduction** from certified ℓ_2 robustness to certified adversarial detection (Lemma IV.1, Theorem IV.2) with an attack-agnostic guarantee. We empirically validate the reduction on 1-Lipschitz SortNet, where the global Lipschitz L_{glob} is exactly known and Lemma IV.1 delivers fully certified detection (Table I).
- Two **practical approximated detectors** for the case where L_{glob} is intractable: $\mathbb{D}_{Adv}$ (single-perturbation surrogate, tighter on Lipschitz-constrained networks) and $\mathbb{D}_{lc}$ (local-centroid surrogate, tighter on unconstrained pretrained classifiers); both deploy without retraining. We give deterministic bias bounds (Section IV-C) and introduce **cert%**, the fraction of all successful in-budget adversarial examples that are both rejected and formally certified. On CIFAR-10 VGG16, our detectors achieve 100% empirical TPR on in-budget adversarial examples while 62–82% of them are both rejected and formally certified; a stronger multi-probe L' estimate further raises CIFAR-10 **cert%** by up to 34.8 points with explicit FPR and runtime costs. (Tables III and VI).
- A **comprehensive empirical evaluation** on MNIST, CIFAR-10 and ImageNet-1k, four architectures (LeNet-5, VGG16, ResNet-50, ViT-B/16), and five baselines (CTT, SortNet, CRM, ContraNet, A²D). Against SOTA empirical detector A²D we deliver 100% TPR at 3.6% FPR on ViT-B/16 ($\epsilon=0.1$) versus A²D’s 89.5% TPR at 11.2% FPR. Furthermore, our detection runs 9–20 $\times$ faster than randomized smoothing on ImageNet (Section V).
- A **principled bridge from robust training to detection**. A large body of robustness research focuses on enlarging clean-input margins (Section II). Because our detectors deploy on any pretrained classifier without retraining, margin-enlarging advances in that literature can tighten our detection frontier under the same reduction, e.g., ℓ_2 PGD adversarial training on a CIFAR-10 VGG16 cuts $\mathbb{D}_{lc}$ ’s FPR from 93.6% to 32.2% at $\epsilon=0.5$ while maintaining $\geq 99\%$ TPR (Section V-E).

II. RELATED WORKS

A. Empirical Adversarial Detectors

A long line of empirical detectors aims to identify adversarial inputs at inference time without certified guarantees [21], [22], [29], [30]. They roughly fall into three categories. *Auxiliary-network detectors* attach trainable components to the protected classifier and flag inputs whose intermediate or reconstructed representations look anomalous: ContraNet [24] trains a class-conditional generator and uses reconstruction error as the detection signal but does not scale to ImageNet-1k, while U-CAN [31] embeds projection layers and ArcFace-style components into intermediate features and detects adversarial inputs through unsupervised anomaly detection. *Statistic-based detectors* compute test statistics from the model’s predictions or gradients on the input; PASA [32] is a recent representative that combines prediction sensitivity and feature-attribution sensitivity to noise into two threshold-calibrated scores. Finally, *adversarial attacks as detectors* flip the perspective: A²D [26] measures the perturbation budget required to cross the decision boundary, exploiting the fact that already-adversarial inputs need less push than clean ones, and is one of the few empirical detectors evaluated on full ImageNet-1k — making it our strongest empirical baseline (Table IVb).

All three families provide *empirical* guarantees only and remain subject to the standard adaptive-attack critique [33], [34]: an adversary aware of the detector can craft inputs that evade its statistics or auxiliary networks while still flipping the underlying classifier. In contrast, our detectors derive from a Lipschitz-margin reduction (Section IV) and inherit an *attack-agnostic* certificate: any in-budget ℓ_2 perturbation that flips the prediction is provably refused, regardless of how the adversarial input was generated.

B. Certified Robustness

Existing work on certified robustness against adversarial examples primarily focuses on designing robust classifiers that maintain correct predictions even when inputs are adversarially perturbed. These approaches can be broadly categorized into three families: (1) relaxation-based methods [35], (2) Lipschitz network methods [17], and (3) randomized smoothing methods [10]. Among them, randomized smoothing and Lipschitz network methods currently achieve state-of-the-art performance against ℓ_2 -norm perturbations; consequently, we concentrate on these two paradigms.

Randomized smoothing (RS) methods [10], [13], [14], [36], [37] enhance robustness by injecting random noise into the input, thereby transforming an arbitrary base classifier into a *smoothed classifier* that is certifiably robust under the ℓ_2 norm. In particular, Cohen et al. [10] introduced Gaussian smoothing and derived the strongest known ℓ_2 robustness radius. For ℓ_1 robustness, Lecuyer et al. [7] employed Laplacian smoothing. Regarding ℓ_∞ robustness, Yang et al. [12] demonstrated that by leveraging norm inequalities to translate ℓ_2 guarantees derived from Gaussian smoothing into ℓ_p bounds, one can achieve state-of-the-art certified ℓ_∞ robustness.

For Lipschitz network (LN) methods, they exploit the Lipschitz constant of the neural network to formally certify robustness [16]–[18], [38]–[40]. Early efforts in this line typically compute a global Lipschitz constant [17], [40] and introduce novel architectures or training procedures to tighten this bound. More recent works directly enforce tight Lipschitz constraints by designing 1-Lipschitz networks [16], [38]. Fazlyab et al. [28] (CRM) maximises a dynamic margin while regularising the Lipschitz constant; we use both SortNet [16] and CRM as certified-detection baselines (Section V). Notably, Delattre et al. [15] recently proposed a hybrid approach that integrates Lipschitz networks with randomized smoothing.

C. Margin and Adversarial Robustness

A complementary line of work analyses adversarial robustness through the *margin* of the classifier. Tsuzuku et al. [40] formalised the link — a classifier with margin $q(\mathbf{x})$ at point $\mathbf{x}$ is provably robust to ℓ_2 perturbations of size up to $q(\mathbf{x})/L$, motivating margin-maximising training. Xu et al. [41] empirically showed that classifiers’ adversarial robustness is governed by decision-boundary dynamics, with larger clean-input margins translating to harder-to-attack inputs. Lyu et al. [42], [43] prove that gradient descent on homogeneous

neural networks converges to a KKT point of the margin-maximising problem, a result we leverage to compute the local centroid in $\mathbb{D}_{lc}$ (Section IV-C). The same margin lens motivates adversarial training [44]–[47] as a margin-enlarging primitive on the clean-input side.

Our work shares the margin-centric view but reaches it from the opposite direction. Rather than asking *how to push clean-input margins outward*, we ask *where an adversarial example must land given a bounded budget*: Lipschitz continuity forces it to remain close to the decision boundary in margin space, which makes the margin function an attack-agnostic detection signal. The two perspectives converge — both single out the margin as the discriminator between robust and non-robust regimes — but ours yields a quantitative reduction from certified robustness to certified detection (Section IV) rather than a robustness-training prescription. As a consequence, every margin-enlarging primitive in this literature — adversarial training, Lipschitz constraints, Gaussian augmentation — composes with our framework: it lifts the underlying classifier’s clean-margin distribution and directly tightens our certified-detection frontier (Section V-E).

D. Certified Detection

Shumailov et al. [19] (CTT) introduced an early certified detection framework with a one-sided guarantee on the *specificity* side: samples flagged as adversarial are certified to be adversarial, with non-zero false-negative rate on real adversarial inputs. CrossCert [48] (FSE 2024) extends one-sided certified detection to patch attacks via a cross-checking mechanism. Our work parallels this line for ℓ_2 -bounded perturbations on the complementary *recall* side: every in-budget adversarial example is provably refused, with non-zero FPR on clean inputs (cf. Section III-C). The two families thus offer complementary safety properties, and the right choice in deployment depends on whether false-positives or false-negatives are more costly.

III. PROBLEM FORMULATION

In this section we formalize the certified adversarial detection problem. We describe the threat model (§III-A), introduce the notation and preliminaries used throughout the paper (§III-B), and then state the formal detection problem and the certification objective (§III-C).

A. Threat Model

a) *Deployment scenarios.*: We consider deployments where a deep classifier $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ is exposed to untrusted inputs from users or sensors:

- **Autonomous-driving perception.** A traffic-sign or pedestrian classifier ingests camera frames; a misclassification under a small physical-world perturbation (e.g., a sticker on a stop sign) can drive unsafe trajectory decisions and cause physical harm.
- **Content moderation.** A platform-side image classifier flags uploads that violate policy; an adversary perturbs a violating image just enough to evade flagging while remaining recognisably the same content.

- **Biometric authentication & access control.** A face- or voice-recognition gate decides whether to grant access; an adversary perturbs an unauthorised face to be classified as an authorised identity.

In all three cases f and its parameters are *fixed at deployment*: retraining or modifying f is not available to the defender, and inference must meet a real-time latency budget (tens of milliseconds for AV perception, hundreds for moderation/auth).

b) Attacker goal and capability.: Given a clean input x that f classifies correctly, the attacker seeks a perturbation δ that flips the prediction, $\arg \max f(x + \delta) \neq \arg \max f(x)$, while keeping δ small in the ℓ_2 norm: $\|\delta\|_2 \leq \epsilon$. The budget ϵ is chosen by the defender and announced as a deployment-side protection level: within this declared ℓ_2 radius, the detector can reject adversarial inputs, with high TPR and formal guarantees for the certified subset. We primarily consider ℓ_2 ; extension to other ℓ_p norms is discussed in Section VI-D.

c) Attacker knowledge and adaptivity.: We assume a white-box attacker with full access to f 's architecture and parameters, consistent with the standard threat model in the adversarial ML literature [44], [49]. The defender likewise has white-box access to f , which is natural in model-owner deployments where the defender controls the classifier and inserts the detector, although the service itself may remain black-box to external users. The attacker may use any attack procedure (FGSM, BIM, PGD, C&W, and adaptive variants aware of the detector). Crucially, our certified guarantee is *attack-agnostic*: by construction it holds for every $\hat{x} = x + \delta$ with $\|\delta\|_2 \leq \epsilon$, independent of how δ was generated. An adaptive attacker therefore cannot bypass the certificate by switching optimisation strategy — any in-budget perturbation that flips the prediction is provably detected (Section V-D).

d) Defender architecture: detection-then-refuse.: The defender deploys f together with a detector $\mathbb{D} : \mathbb{R}^d \rightarrow \{0, 1\}$. On input z , the pipeline returns $f(z)$ when $\mathbb{D}(z) = 0$ and *refuses* the input when $\mathbb{D}(z) = 1$. Refusal is a hard reject signal: it does not return any class label and does not attempt to recover the original clean input. Refused inputs are escalated to a deployment-defined fallback — human review for content moderation, an additional verification factor for biometric authentication, or a safe-stop policy for AV perception. This *detection-then-refuse* paradigm contrasts with *correction*-based defenses (which try to recover the clean label from a perturbed input) and with *soft-rejection* defenses (which lower the output confidence without rejecting): we make no recovery attempt and emit a binary refuse signal.

e) Defender objective.: The defender's primary objective is to refuse every successful in-budget adversarial example. This objective is fully certified when a valid global Lipschitz bound is available. For standard pretrained models, where L_{glob} is unavailable, our detectors approximate this objective: they reject successful in-budget adversarial examples empirically, and we report **cert%** for the subset whose rejection is formally certified. Refusing some clean inputs (non-zero FPR) is a willing trade-off, calibrated by deployment context: an AV perception pipeline can absorb conservative refusal by

deferring to the safe-stop fallback, content moderation can route flagged uploads to human reviewers, and biometric gates can request a second factor. A secondary objective is inference efficiency — the per-input overhead of $\mathbb{D}$ must remain within the latency budget of the host pipeline. We make no claim about robustness to attackers operating outside the announced ℓ_2 budget; that is the responsibility of the announced safety envelope.

B. Notation and Preliminaries

A complete summary of symbols used throughout the paper is given in Appendix B (Table IX).

We denote a clean input by x and its adversarial counterpart by $\hat{x} = x + \delta$, where ϵ is the upper bound of perturbation. We further use z for a generic test input at inference time (clean or adversarial — the defender does not know which) and $z' = z + \sigma$ for a deliberate, defender-side perturbation of z used as a numerical surrogate for z 's clean counterpart; the attacker-side $\hat{x}$ and the defender-side z' live on opposite sides of the inference pipeline. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a neural network, with $f_i(x)$ the i^{th} output component (the confidence that x belongs to class i), and let θ denote the parameters of f . The ℓ_p norm of a vector is $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$; unless otherwise stated we default to the ℓ_2 norm.

The theoretical foundation of our method relies on the Lipschitz continuity of f . Intuitively, Lipschitz continuity bounds how much the output changes under a given input perturbation.

a) Lipschitz continuity.: A mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ is L -Lipschitz continuous with respect to norm $\|\cdot\|$ if, for all $x_1, x_2 \in \mathbb{R}^d$,

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|. \quad (1)$$

b) Margin.: For a reference class k , the *class- k margin* of an input u with respect to f is

$$q(u; k) = f_k(u) - \max_{j \neq k} f_j(u). \quad (2)$$

We write $q(u)$ as shorthand for $q(u; \arg \max_k f_k(u))$, i.e., the margin at u 's *own predicted class*; this is non-negative by construction. When the reference class is fixed and clear from context (e.g., the pseudo-label $\hat{y} = \arg \max f(z)$ of an input z throughout a detection step), we keep the shorthand $q(\cdot)$ but note that the value can become *negative* if the second argument differs from the reference class — as happens to $q(z'; \hat{y})$ when the defender-side perturbation z' flips z 's prediction (Algorithm 1). For Lemma IV.1, both $q(x)$ and $q(\hat{x})$ are evaluated at their own predicted classes (so both are non-negative and the bound is symmetric).

c) Certified robustness.: For a neural network f with Lipschitz constant L under ℓ_p , the model is provably robust to perturbations $\|\delta\|_p < \frac{1}{cL}q(x)$, i.e.,

$$\arg \max f(x) = \arg \max f(x + \delta), \quad \forall \|\delta\|_p < \frac{1}{cL}q(x), \quad (3)$$

where $c = 2^{\frac{p-1}{p}}$ depends on the norm. This is the standard robustness certificate [5], [6], [40]; our work reduces *detection* to it.

C. Certified Adversarial Detection

We now formalize the detection problem and the certification objective.

Definition III.1 (Certified Detector). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a classifier and let $\epsilon > 0$ be a perturbation budget. A detector $\mathbb{D} : \mathbb{R}^d \rightarrow \{0, 1\}$ is a *certified detector at budget ϵ* (under ℓ_p) if, for every clean input x and every adversarial example $\hat{x} = x + \delta$ satisfying $\|\delta\|_p \leq \epsilon$ and $\arg \max f(\hat{x}) \neq \arg \max f(x)$, we have $\mathbb{D}(\hat{x}) = 1$.

Remark III.2 (One-sided guarantee and non-triviality). Definition III.1 is a *one-sided* certificate: $\mathbb{D}$ must reject every in-budget successful adversarial example, but may also reject some clean inputs (non-zero FPR). This parallels the complementary guarantee of CTT [19], which certifies that clean samples are *not* flagged and admits non-zero FNR on adversarial inputs, and of CrossCert [48], which defines the same form of one-sided certified detection against patch attacks. It also mirrors certified robustness, which guarantees correct classification inside a radius while admitting non-trivial error outside it. Crucially, joint reporting of TPR and FPR in our evaluation rules out trivial detectors: the constant detector $\mathbb{D} \equiv 1$ satisfies Definition III.1 but has FPR = 1.

a) *Evaluation metrics.*: Given a test set of clean inputs and successful adversarial examples, we report three quantities:

- **TPR**: the fraction of *successful* adversarial examples (those that change f 's prediction) that $\mathbb{D}$ rejects.
- **FPR**: the fraction of clean samples that $\mathbb{D}$ rejects.
- **cert%**: the fraction of all successful in-budget adversarial examples that are both rejected and formally certified, i.e., the local Lipschitz estimate L' used by $\mathbb{D}$ satisfies the per-detector certification condition on that pair (cf. Section IV-E).

b) *Design objective.*: Given a pretrained classifier f and a budget ϵ , we seek a certified detector (Definition III.1) that additionally (i) achieves low FPR at ϵ of practical interest, (ii) is applicable to any pretrained f without retraining or architectural constraints, and (iii) has inference-time overhead within the latency budget of real-time pipelines such as AV perception. Section IV develops the Lipschitz-margin analysis that realizes this objective.

IV. FROM CERTIFIED ROBUSTNESS TO CERTIFIED DETECTION

We now formalize the Lipschitz-margin reduction sketched in Figure 1 and instantiate two approximated detectors that operate on any pretrained classifier without retraining. The load-bearing results are Lemma IV.1 (the core inequality) and Theorem IV.2 (its centroid-distance form); Claims 1–2 instantiate the two approximated detectors; the bias bounds

(Section IV-D) and certification conditions (Section IV-E) tighten the connection between them.

A. Certified Detection

An L -Lipschitz network cannot change its output by more than $L\|\delta\|$ under perturbation $\|\delta\|$. So if a clean input x has a large margin in its true-class direction, an in-budget adversarial example $\hat{x}$ cannot simultaneously achieve a large margin in any wrong-class direction — the budget is too small to “move” that much logit mass. Lemma IV.1 captures this in a one-sided sum bound.

Lemma IV.1. (Certified Detection) *Given a DNN f with Lipschitz constant L under ℓ_p norm $\|\cdot\|_p$, for an adversarial example $\hat{x}$ and its clean counterpart x , the margin satisfies*

$$q(\hat{x}) \leq cL\|\delta\| - q(x), \quad (4)$$

where $q(\hat{x})$ is evaluated w.r.t. the predicted class of $\hat{x}$ (Section III-B) and $c = 2^{(p-1)/p}$.

The proof is deferred to Appendix E. The bound is consistent with certified robustness: $q(\hat{x}) \geq 0$ by definition (Section III-B), so when $\|\delta\|$ is below the certified-robustness threshold $q(x)/(cL)$, the right-hand side becomes negative and Lemma IV.1 cannot be satisfied — equivalently, no in-budget adversarial example exists, recovering the standard robustness certificate.

Based on Lemma IV.1, a certified detector $\mathbb{D}$ can be formally defined as

$$\mathbb{D}(z) = \begin{cases} 1, & \text{if } z \text{ satisfies Lemma IV.1,} \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

where input z may be either a clean sample or a perturbed sample.

However, in practice, the clean counterpart x is unknown for an adversarial example, and it does not even exist for a clean sample. Therefore, this detector cannot be directly applied. To resolve this issue, we propose approximation strategies.

B. Approximation to the Certified Detector

Approximating the bound in Lemma IV.1 requires identifying the preimage x of z , which is generally infeasible because:

- 1) For $z = \hat{x}$, the underlying clean input x is unknown, as the adversarial perturbation is typically generated through an optimization procedure rather than a well-defined invertible mapping.
- 2) For z that is not adversarial, a corresponding x may not exist at all.

We therefore propose two approximation methods: (1) approximating x by perturbing z ; (2) introducing a numerically computable surrogate, namely the local centroid.

1) *Approximation by Perturbation.* We first approximate x by perturbing z : given z , we generate $z' = z + \sigma$ and use it as a surrogate for x in Lemma IV.1 (recall from Section III-B that z' is the defender-side perturbation, not the attacker-side $\hat{x}$).

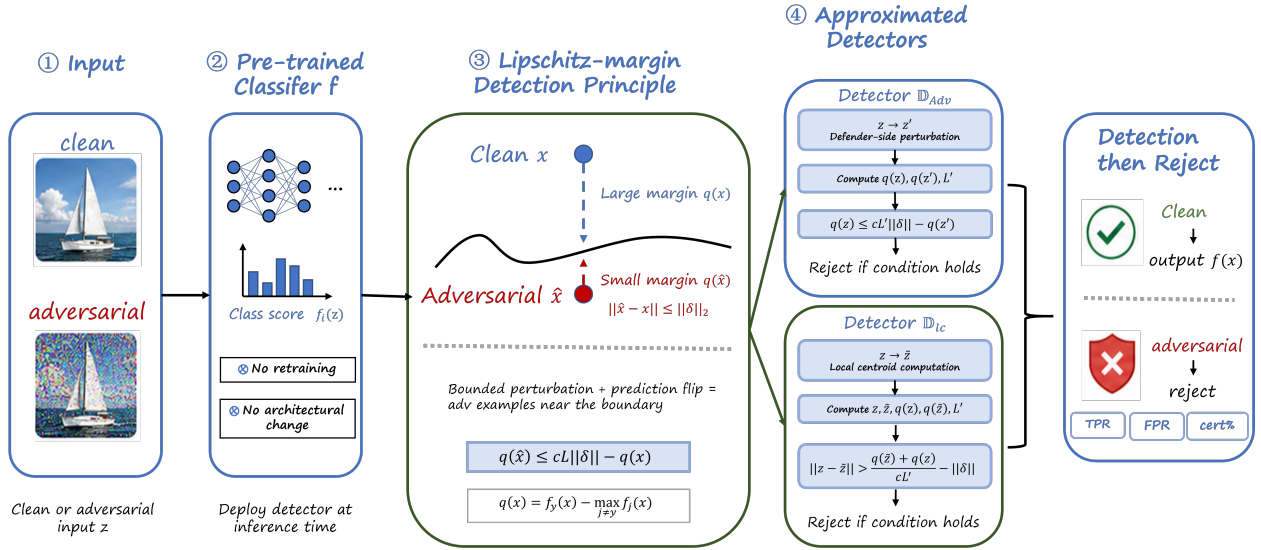


Fig. 1: **Overview of certified adversarial detection.** Taking a pre-trained classifier f as-is (no retraining and no architectural change), we exploit the *Lipschitz-margin* principle: a clean input x exhibits a large margin $q(x) = f_y(x) - \max_{i \neq y} f_i(x)$, whereas a successful adversarial example $\hat{x}$ with bounded perturbation $\|\delta\|$ is forced close to the decision boundary, yielding a small margin upper-bounded by $q(\hat{x}) \leq cL\|\delta\| - q(x)$ (Lemma IV.1). Because the clean counterpart x is unobserved at inference time, we instantiate two approximated detectors — $\mathbb{D}_{Adv}$ (defender-side perturbation surrogate; Claim 1) and $\mathbb{D}_{Lc}$ (local-centroid surrogate; Claim 2) — which flag adversarial inputs while accepting clean ones, and report TPR, FPR, and cert%, where cert% counts successful in-budget adversarial examples that are both rejected and formally certified.

Claim 1. (Detector $\mathbb{D}_{Adv}$) For a DNN f and an input z with pseudo-label $\hat{y} = \arg \max f(z)$, given a defender-side perturbation z' of z , if $z = \hat{x}$, then

$$q(z; \hat{y}) + q(z'; \hat{y}) \leq cL'\|\delta\|, \quad (6)$$

where $L' = (q(z; \hat{y}) - q(z'; \hat{y})) / (c \max(\|z - z'\|, \|\delta\|))$ approximates the Lipschitz constant from the local pair (the numerator is non-negative whenever z' flips z 's prediction, so $L' \geq 0$).

Since $L' \leq L$ and $q(z') \neq q(x)$, this introduces a bias compared to Lemma IV.1; the inequality in Claim 1 is therefore the *approximate* certification test used by $\mathbb{D}_{Adv}$ rather than an unconditional bound. Proposition IV.7 establishes the conditions under which it is provably correct on adversarial inputs.

2) *Approximation by Local Centroid*: To better track the bias on classifiers with large clean margins, we introduce the concept of a *local centroid*, which can be computed numerically. The local centroid of x with respect to f is defined as the local maximizer of the margin function:

$$\tilde{x} = \arg \max_x q(x), \quad q(x) := f_y(x) - \max_{i \neq y} f_i(x).$$

Such local optima have been used in detecting backdoor attacks [50]. Here, we employ them as surrogates for x to strengthen the bound.

By Lemma A.4 (Appendix G), gradient ascent on $q(\cdot; \hat{y})$ at the input's frozen pseudo-label $\hat{y}$ is monotonically non-decreasing in the gradient-flow limit; starting from $q(z; \hat{y}) > 0$,

the centroid trajectory keeps the margin positive and $\tilde{z}$ stays predicted as $\hat{y}$, so $y(\tilde{x}) = y(\hat{x})$ as required by Theorem IV.2. Following Lyu et al. [42], when f is homogeneous in its parameters (a condition satisfied by ReLU networks without bias terms), this ascent additionally converges to a Karush–Kuhn–Tucker (KKT) point of $\min \frac{1}{2} \|x\|^2$ s.t. $q(x) \geq 1$ — our setting is the input-space counterpart of their parameter-space result.

Intuitively, points near the decision boundary tend to be farther away from their local centroids. This intuition leads to the following guarantee:

Theorem IV.2. (Certified Detection via Local Centroid) Given a neural network f with Lipschitz constant L , for an adversarial example $\hat{x}$ and its clean counterpart x , the distance between $\hat{x}$ and its local centroid $\tilde{x}$ satisfies

$$\|\tilde{x} - \hat{x}\| \geq \frac{q(\tilde{x}) + q(x)}{cL} - \|\delta\|. \quad (7)$$

The proof (Appendix E) applies Lemma IV.1 to the pair $(x, \tilde{x})$ and combines it with the triangle inequality. The theorem strictly enlarges the certified region of Lemma IV.1: any pair $(x, \hat{x})$ satisfying the lemma also satisfies the theorem, because $q(\tilde{x}) \geq q(\hat{x})$ by construction of the centroid. The right-hand side is non-vacuous (positive) whenever $\|\delta\| < (q(\tilde{x}) + q(x)) / (cL)$, which is the regime in which the certified test fires.

Analogous to the first approximation, we further use $q(z')$ as a surrogate for $q(x)$ and replace L with a local Lipschitz constant L' .

Algorithm 1 Perturbed Sample Generation ($z' = z + \sigma$)

Input: Test input z (clean or adversarial), classifier f , radius $\|\delta\|$, step size α , number of steps T
Output: Perturbed version z'
 Compute pseudo-label $y = \arg \max f(z)$
 Initialize $z' \leftarrow z$
for $t = 1$ to T **do**
 Compute loss $\mathcal{L}(f(z'), y)$
 Compute gradient $g \leftarrow \nabla_{z'} \mathcal{L}(f(z'), y)$
 Update $z' \leftarrow z' + \alpha \cdot \text{sign}(g)$ {gradient ascent step}
 Project z' onto the ℓ_2 -ball $\mathcal{B}_2(z, \|\delta\|)$:
 $\Delta \leftarrow z' - z$
 $\Delta \leftarrow \Delta \cdot \min \left\{ 1, \frac{\|\delta\|}{\|\Delta\|_2} \right\}$
 $z' \leftarrow z + \Delta$
 Clip pixel values: $z' \leftarrow \text{clip}(z', 0, 1)$
 if $\arg \max f(z') \neq y$ **then**
 break {break on flip: $y(z') \neq y$, i.e., $q(z'; y) \leq 0$ }
 end if
end for

Claim 2 (Detector $\mathbb{D}_{lc}$). For a DNN f and an input z with pseudo-label $\hat{y} = \arg \max f(z)$, given a defender-side perturbation z' of z and the local centroid $\tilde{z}$ of z , if $z = \hat{x}$, then

$$\|\tilde{z} - z\| \geq \frac{q(\tilde{z}; \hat{y}) + q(z'; \hat{y})}{cL'} - \|\delta\|. \quad (8)$$

As with Claim 1, the inequality is the *approximate* certification test used by $\mathbb{D}_{lc}$. The substitutions $q(x) \rightarrow q(z')$ and $L \rightarrow L'$ make it conditional on $L' \geq L_{(\tilde{x}, x)}^*$ and on the flip property $q(z'; \hat{y}) \leq 0$ (Proposition IV.7, part (ii)).

Remark IV.3 (False Positive Rate in Practice). We note that the false positive rate (FPR) of our detectors can be relatively high in practice. This stems from two factors. First, many clean samples naturally lie close to the decision boundary — an inherent property of the trained model and the dataset — which longer training or stronger regularization can alleviate (Section V-E). Second, larger perturbation bounds $\|\delta\|$ allow adversarial examples to drift farther from their local centroids, but the detection rule must then admit a wider band that inevitably flags more clean samples. The TPR/FPR trade-off is therefore governed by both the model’s clean-margin distribution and the announced budget ϵ .

C. Approximated Detector

We now translate Claims 1–2 into concrete algorithms. The main text describes the detector construction; full pseudocode for perturbed-sample generation, local-centroid optimization, and both detector variants is given below.

1) *Basic Detector ($\mathbb{D}_{Adv}$): Perturbed-sample generation.* Given test input z , we generate $z' \in \mathcal{B}_2(z, \|\delta\|)$ by running PGD [44] on the predicted class (Algorithm 1). z' replaces the unknown clean counterpart x in Claim 1.

Local Lipschitz estimation. Since the global Lipschitz constant is typically intractable, we estimate a *local* Lipschitz

Algorithm 2 Approximate Detector from Claim 1 ($\mathbb{D}_{Adv}$)

Input: Input z , perturbation radius $\|\delta\|$, model f
Output: Detection label (ADVERSARIAL / CLEAN)
 $\hat{y} \leftarrow \arg \max f(z)$ {pseudo-label, fixed reference for all $q(\cdot)$ below}
 Generate perturbed version $z' = z + \sigma$
 Estimate $L' \leftarrow \frac{q(z; \hat{y}) - q(z'; \hat{y})}{c \cdot \max(\|z - z'\|_2, \|\delta\|)}$
 $\tau \leftarrow L' \cdot c \cdot \|\delta\|$
if $q(z; \hat{y}) + q(z'; \hat{y}) \leq \tau$ **then**
 Return: ADVERSARIAL
else
 Return: CLEAN
end if

Algorithm 3 Local Centroid Optimization

1: **Input:** Input z , classifier f , step size η , number of ascent steps T
2: **Output:** Local centroid $\tilde{z}$ and distance $d_z = \|\tilde{z} - z\|_2$
3: Compute pseudo-label $y \leftarrow \arg \max f(z)$ {frozen throughout ascent}
4: Initialize $\tilde{z} \leftarrow z$
5: **for** $t = 1$ to T **do**
6: Compute margin $q(\tilde{z}) = f_y(\tilde{z}) - \max_{j \neq y} f_j(\tilde{z})$
7: Compute gradient $g \leftarrow \nabla_{\tilde{z}} q(\tilde{z})$
8: Update $\tilde{z} \leftarrow \tilde{z} + \eta \cdot g$ {Gradient ascent to enlarge margin}
9: Clip $\tilde{z}$ to valid input range: $\tilde{z} \leftarrow \text{clip}(\tilde{z}, 0, 1)$
10: **end for**
11: Compute distance $d_z \leftarrow \|\tilde{z} - z\|_2$

constant from the pair (z, z') at the input’s pseudo-label $\hat{y} = \arg \max f(z)$:

$$L' \approx \frac{q(z; \hat{y}) - q(z'; \hat{y})}{c \cdot \max(\|z - z'\|_2, \|\delta\|)}, \quad c = \sqrt{2}.$$

This data-dependent surrogate replaces L in Claim 1; Algorithm 2 implements the resulting detector $\mathbb{D}_{Adv}$.

2) *Refined Detector ($\mathbb{D}_{lc}$):* The Refined Detector is the practical-side counterpart of the Lemma IV.1 $\rightarrow$ Theorem IV.2 refinement. In Claim 1, z' alone plays a double role: it is both the surrogate for x in $q(z')$ and the second point used to estimate L' . This double-use ties the bias on both quantities to the same geometric distance between z' and the unobserved x , making the substitution non-negligible whenever z' is not approximately on the segment $(x, \hat{x})$. Introducing the local centroid $\tilde{x}$ (Section IV-B) decouples the two roles: $\tilde{x}$ supplies a precisely-computable, high-margin anchor for the centroid-distance bound, while z' continues to drive the local Lipschitz estimator. We compute $\tilde{x}$ by iterative gradient ascent on $q(\cdot)$ starting from z (Algorithm 3); Algorithm 4 then implements $\mathbb{D}_{lc}$ following Claim 2.

Algorithm 4 Approximate Detector from Claim 2 ($\mathbb{D}_{lc}$)

```

1: Input: Input  $z$ , perturbation budget  $\|\delta\|$ , model  $f$ 
2: Output: Detection label (ADVERSARIAL / CLEAN)
3:  $\hat{y} \leftarrow \arg \max f(z)$  {pseudo-label, fixed reference for all  $q(\cdot)$  below}
4: Generate perturbed version  $z' = z + \sigma$ 
5: Compute local centroid  $\tilde{z}$  and distance  $d_z = \|\tilde{z} - z\|_2$ 
6: Estimate  $L' \leftarrow \frac{q(\tilde{z}; \hat{y}) - q(z'; \hat{y})}{c \cdot \max(\|\tilde{z} - z'\|_2, \|\delta\|)}$ 
7:  $\text{RHS} \leftarrow \frac{q(\tilde{z}; \hat{y}) + q(z'; \hat{y})}{c \cdot L'} - \|\delta\|$ 
8: if  $d_z \geq \text{RHS}$  then
9:   Return: ADVERSARIAL
10: else
11:   Return: CLEAN
12: end if

```

D. Bias Analysis of Approximated Detector

The substitution $q(z') \rightarrow q(x)$ contributes a bias bounded by $|q(z') - q(x)|$; we focus here on the harder term, the deviation between the local estimator L' and its theoretical counterpart L .

1) *Bias Analysis of the Basic Detector:* Since Claim 1 replaces x with its perturbed counterpart z' , the estimated local Lipschitz constant may deviate from the ideal one computed on $(x, \hat{x})$. We characterize this deviation as follows.

Lemma IV.4 (Bounded Bias of the Approximated Lipschitz Constant). *Let $a = \hat{x}$, $b = x$, and $z' = z + \sigma$ denote the adversarial, clean, and perturbed samples, respectively. Define $\alpha = \|a - b\|$, $\beta = \|a - z'\|$, and $\gamma = \|b - z'\|$, and let*

$$L^* = \frac{q(a) + q(b)}{c\alpha}, \quad L' = \frac{|q(a) - q(z')|}{c\beta}.$$

Then the deviation between L' and L^* satisfies

$$|L' - (\alpha/\beta)L^*| \leq \frac{\gamma}{\beta} L_{\text{glob}} + \frac{\alpha}{\beta} (L^* - D),$$

where $D = \frac{|q(a) - q(b)|}{c\alpha} \leq L^*$ and L_{glob} is a global Lipschitz constant of $q(\cdot)$.

The proof is deferred to Appendix H. Intuitively, the first term $\frac{\gamma}{\beta} L_{\text{glob}}$ captures the geometric bias introduced by using z' instead of the clean sample x , while the second term $\frac{\alpha}{\beta} (L^* - D)$ reflects the margin bias between sum- and difference-based margin quotients. When z' lies approximately along the segment $(x, \hat{x})$, both terms diminish, and L' provides a tight local approximation of the ideal Lipschitz constant.

2) *Bias Analysis of the Refined Detector:* Recall that the ideal constant in Lemma IV.1 can be written as

$$L^* = \frac{q(\hat{x}) + q(x)}{c\|\hat{x} - x\|},$$

where x is the clean input and $\hat{x}$ its adversarial counterpart. In contrast, our detector employs

$$L' = \frac{|q(\tilde{z}) - q(z')|}{c\|\tilde{z} - z'\|},$$

where z' is the perturbed version of z and $\tilde{z}$ is the local centroid. The difference between these two quantities can be bounded as

$$L' - L^* \leq (\kappa - 1)L_{\text{glob}} + (L_{\text{glob}} - L^*),$$

$$\kappa = \frac{\|\tilde{z} - z\| + \|z' - z\|}{\|\tilde{z} - z'\|} \geq 1,$$

where L_{glob} is the global Lipschitz constant. This inequality shows that the bias arises from two sources: (i) the geometric factor κ , which depends on the relative positions of $\tilde{z}$, z , and z' , and (ii) the inherent looseness of the global bound L_{glob} compared to L^* . In particular, when z' lies close to $\tilde{z}$, κ may become large, leading to a systematic overestimation of L^* . Conversely, when z' is well separated, L' better approximates L^* .

The above analysis indicates that the bias of the refined detector stems from both geometric and Lipschitz factors. To make this relationship more explicit and to enable practical estimation, we next derive a deterministic and fully computable upper bound on the deviation between the approximation and the ideal certified term.

Theorem IV.5 (Deterministic bound for the approximation). *Let $z = \hat{x}$ be an adversarial input with clean counterpart x ($\|z - x\| \leq \|\delta\|$), and $z' = z + \sigma$ with $\|\sigma\| \leq \|\delta\|$. Assume that*

- (A1) (**Lipschitz of margin**) q is L_q -Lipschitz under $\|\cdot\|$ with $L_q \leq c\bar{L}$: $|q(u) - q(v)| \leq L_q\|u - v\|$ for all u, v in a neighborhood of the data manifold.
- (A2) (**Global Lipschitz bound**) There exists a computable global Lipschitz upper bound $L_{\text{glob}} \leq \bar{L}$, and $L' \leq L_{\text{glob}}$ (i.e., the local estimator does not exceed the global upper bound).

Then $0 < L' \leq L_q/c \leq \bar{L}$ and, for

$$\Delta_T := \left| \frac{q(\tilde{z}) + q(z'; \hat{y})}{cL'} - \frac{q(\tilde{z}) + q(x)}{cL_{\text{glob}}} \right|.$$

We write the relative gap $\eta := 1 - L'/\bar{L} \in [0, 1)$ and have the fully computable bound

$$\Delta_T \leq \frac{1}{c} \left[\frac{2L_q}{L'} \|\delta\| + \frac{\eta}{L'} (q(\tilde{z}) + |q(z'; \hat{y})| + L_q \|\delta\|) \right]. \quad (9)$$

Δ_T is the gap between the approximated certified threshold (computed at inference from L' and z') and the ideal one in Theorem IV.2 (computed from the global L_{glob} and the unobserved clean margin $q(x)$). The two summands isolate two error sources. The first, $\frac{2L_q}{cL'} \|\delta\|$, is the *geometric* stretch from substituting z' for the unobserved x (the factor 2 absorbs the triangle $x \rightarrow z \rightarrow z'$), scaled by how tight L' is relative to the margin's Lipschitz constant L_q . The second, scaled by the relative gap $\eta = 1 - L'/\bar{L} \in [0, 1)$, is the *looseness* from replacing the global upper bound $\bar{L}$ with the local estimator L' : when L' approaches $\bar{L}$, $\eta \rightarrow 0$ and the second term vanishes. Both terms are observable at inference from $\|\delta\|$, L' , $\bar{L}$, and the margins.

E. When the Practical Detectors Provably Detect Adversarial Examples?

We now connect the approximated detectors $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ back to the formal certificate of Lemma IV.1 and Theorem IV.2. Two gaps separate the approximated detectors from the formal certificate: the local estimator L' may fall short of the pairwise Lipschitz L^* (the deviation $|L' - L^*|$ is bounded for $\mathbb{D}_{Adv}$ by Lemma IV.4, and the resulting threshold gap for $\mathbb{D}_{lc}$ is bounded by Theorem IV.5), and the surrogate $q(z')$ replaces the unobserved $q(x)$ (closed by the break-on-flip rule in Algorithm 1, which guarantees $q(z'; \hat{y}) \leq 0$ whenever the inner attack succeeds within T steps). To state the certification condition cleanly, we first define the pairwise Lipschitz constants involved.

Proposition IV.6 (Pairwise Lipschitz). *For an adversarial example $\hat{x}$, its clean counterpart x , and the local centroid $\tilde{x}$ of $\hat{x}$, define the pairwise Lipschitz constants*

$$L_{(x, \hat{x})}^* := \frac{q(\hat{x}) + q(x)}{c \|\hat{x} - x\|}, \quad L_{(\tilde{x}, x)}^* := \frac{q(\tilde{x}) + q(x)}{c \|\tilde{x} - x\|}.$$

Both are saturations of Lemma IV.1: $L_{(x, \hat{x})}^$ is the minimum L for which the lemma holds on the $(x, \hat{x})$ pair; and $L_{(\tilde{x}, x)}^*$ is the minimum L for which the lemma holds on the $(\tilde{x}, x)$ pair (both $x, \tilde{x}$ have differing predicted classes since $y(\tilde{x}) = y(\hat{x}) \neq y(x)$). Theorem IV.2 can be derived from a single application of Lemma IV.1 on the $(x, \tilde{x})$ pair plus the triangle inequality (see Appendix E); it therefore holds for any $L \geq L_{(\tilde{x}, x)}^*$.*

Proposition IV.7 (When approximated detectors are certified). *For every successful in-budget adversarial example $\hat{x}$:*

- (i) $L' \geq L_{(x, \hat{x})}^*$ implies the bound of Claim 1 ($\mathbb{D}_{Adv}$);
- (ii) $L' \geq L_{(\tilde{x}, x)}^*$ together with the flip property $q(z'; \hat{y}) \leq 0$ — guaranteed by Algorithm 1’s break-on-flip whenever the inner attack succeeds within T steps — implies the bound of Claim 2 ($\mathbb{D}_{lc}$).

Proof sketch. For input $z = \hat{x}$: (i) The max-clamp denominator in L' combined with $L' \geq L_{(x, \hat{x})}^*$ and Lemma IV.1 forces $q(z'; \hat{y}) \leq -q(x) \leq 0$, so the flip property emerges as a *consequence*; substituting $q(z'; \hat{y}) \leq -q(x)$ into the detector’s threshold then recovers the Claim 1 bound. (ii) The flip property gives $q(z'; \hat{y}) \leq 0 \leq q(x)$, so substituting $q(x) \rightarrow q(z'; \hat{y})$ in the detector’s right-hand side only loosens it, and $L' \geq L_{(\tilde{x}, x)}^*$ closes the chain back to Theorem IV.2. Full proof in Appendix F. $\square$

Remark IV.8 (What cert% certifies). The framework operates in two regimes.

(i) **Global Lipschitz** L_{glob} . On networks with a known global Lipschitz constant (e.g., 1-Lipschitz SortNet, Table I), $L_{glob} \geq L_{(x, \hat{x})}^*$ and $L_{glob} \geq L_{(\tilde{x}, x)}^*$ for every pair, so plugging $L = L_{glob}$ certifies both Lemma IV.1 and Theorem IV.2 for every in-budget adversarial example.

(ii) **Local estimator** L' . On standard pretrained classifiers L_{glob} is intractable, so we estimate L' at inference. The

cert% metric (Section V) measures the empirical fraction of all successful in-budget adversarial examples that are both rejected and satisfy the per-detector certification hypothesis of Proposition IV.7: $L' \geq L_{(x, \hat{x})}^*$ for $\mathbb{D}_{Adv}$, and $L' \geq L_{(\tilde{x}, x)}^*$ together with the flip property $q(z'; \hat{y}) \leq 0$ for $\mathbb{D}_{lc}$. **All cert% numbers reported in the abstract, experiment tables, and discussion use these per-detector definitions**; we keep the shorthand “ $L' \geq L^*$ ” in subsequent table captions to avoid clutter. Empirically, on every rejected adversarial that satisfies $L' \geq L^*$, $q(z'; \hat{y}) \leq 0$ always holds in all our experiments, so the conjoint hypothesis collapses to $L' \geq L^*$ in practice. By Proposition IV.7, cert% is exactly the fraction on which the approximated detectors provide a formal certificate.

V. EXPERIMENTS

A. Experimental Setup

We evaluate on **MNIST** [51] (LeNet-5), **CIFAR-10** [52] (VGG16), and **ImageNet-1k** [53] (PyTorch-pretrained ResNet-50; ViT-B/16 in the A²D comparison). Adversarial examples are generated under ℓ_2 with FGSM [54], BIM [55], PGD (40 steps, 10 restarts) [44], and CW (40 steps) [49].

Metrics. We report **TPR**, the fraction of *successful* adversarial examples that are detected (success means the prediction changes), **FPR**, the fraction of clean samples flagged as adversarial, and **cert%**. cert% is reported only on detectors whose guarantee depends on L' (i.e., $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$).

Implementation. Both detectors are implemented in PyTorch on top of the pretrained classifiers without any retraining or fine-tuning. The perturbed surrogate z' is generated by $T = 10$ BIM steps with step size $\alpha = \epsilon/T$ inside the ℓ_2 -ball of radius $\|\delta\|$ (Algorithm 1); the local centroid $\tilde{z}$ uses $T_{cent} = 10$ gradient-ascent steps on $q(\cdot)$ at step size 0.1 with $[0, 1]^d$ clipping (Algorithm 3). Unless otherwise noted, all results are evaluated on a single NVIDIA V100 GPU, averaged over 5 runs with different random seeds.

Detection Baselines. We compare against five external baselines spanning certified and empirical detection. Certified detection remains a relatively nascent setting, so the baselines differ in what they certify. CTT [19] is the closest prior certified-detection method, but it provides a complementary specificity-side certificate rather than our recall-side guarantee in Definition III.1. For a broader defense-level comparison, we therefore instantiate certified-robustness methods (RS, SortNet, CRM) as detectors via their certified radii, and include state-of-the-art empirical detectors (ContraNet and A²D).

(i) *SortNet* [16] is a 1-Lipschitz network whose certified robustness radius gives a natural detection rule (flag inputs whose certified ℓ_2 radius $r < \epsilon$). SortNet doubles as the testbed for our fully certified Lemma IV.1 (Table I), since L_{glob} is exactly known and used directly.

(ii) *CRM* [28] is a Lipschitz-regularised classifier with dynamic margin maximisation; its certified-radius detection rule mirrors SortNet’s.

(iii) *CTT* [19] provides the complementary specificity-side certified-detection baseline; we evaluate both CTT-loose and CTT-strict thresholding variants.

(iv) *ContraNet* [24] trains a class-conditional generator and detects adversarial inputs via reconstruction error; it does not scale to ImageNet-1k.

(v) A^2D [26] is a recent state-of-the-art empirical detector that uses the perturbation budget required to flip the prediction as the detection signal; it is one of the few empirical detectors evaluated on full ImageNet-1k and serves as our headline empirical comparison (Table IVb).

In addition to these five external baselines, we adapt the randomized smoothing of Cohen *et al.* [10] into a certified detection rule for completeness: compute the certified radius r for each input and output ADVERSARIAL iff $r < \epsilon$. We use $n_0 = 64$, $n = 512$, and $\sigma = 0.25$ throughout the main body; the conservative variant (smoothed-mismatch + certified radius) is reported in the Conservative RS detector paragraph below.

B. Main Results

We evaluate our two approximated detectors $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ on MNIST, CIFAR-10, and ImageNet-1k. Across datasets, our methods match or outperform existing detection baselines on the TPR/FPR frontier. Unlike RS, CTT, or ContraNet, our detectors require neither Monte Carlo sampling, retraining, nor additional models to distinguish adversarial examples.

a) Fully Certified Detection on Lipschitz Networks.:

On 1-Lipschitz SortNet [16], where L_{glob} is exactly known, applying Lemma IV.1 with $L = L_{glob}$ yields *fully* certified detection: every successful ℓ_2 adversarial example within budget ϵ is provably rejected (TPR = 100%). Table I compares this fully certified path against the baseline detectors built on top of SortNet’s and CRM’s [28] own certified-radius computations (flag if the certified radius $r < \epsilon$), and against our approximated detectors $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ that use the local Lipschitz estimator L' in place of L_{glob} . All methods reach 100% TPR (Except $\mathbb{D}_{Adv}$ with TPR $\geq 99\%$); the differentiator is FPR. Our fully certified path delivers substantially lower FPR than either baseline (e.g., 31.5% vs. 59.0% / 68.0% at $\epsilon=0.3$). The practical $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ use the same detection-rule families with local L' estimates, which can further tighten FPR but are certified only on samples where the per-detector L' condition holds.

MNIST. Both $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ achieve perfect detection on MNIST, maintaining 100% TPR across all attacks (FGSM, BIM, PGD, CW) at every ϵ tested (Table II). FPR stays at most 1% for $\epsilon \leq 0.1$ and grows gradually thereafter (17% at $\epsilon=1.0$). The Cohen detector by contrast attains $\leq 45\%$ TPR at $\epsilon=0.1$ and reaches 98% FPR by $\epsilon=1.0$ at the same noise level. Full per-attack and per- σ breakdowns are in Appendix D.

CIFAR-10. Table III reports representative results on CIFAR-10 (VGG16). Both $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ reach 100% TPR at every ϵ , with FPR rising from $\sim 15\%$ at $\epsilon=0.05$ to $\sim 93\%$ at $\epsilon=0.5$. On a standard-trained VGG16, $\epsilon \leq 0.1$ gives a practical operating point with FPR $\leq 30\%$. We return to the FPR–budget trade-off and ways to shift this frontier in Section VI-B. In contrast, RS (Cohen) maintains moderate TPR but incurs very high FPR (35–100%) even at small perturbations. Our

cert% at $\epsilon=0.1$ on VGG16 reaches 75%, while empirical TPR remains 100%. Compared with ContraNet, our detectors achieve 100% empirical TPR and additionally certify the cert% subset, whereas ContraNet provides only empirical separation. Additional CIFAR-10 ResNet-18 and MNIST VGG-style CNN results in Appendix D, Table XIV, further validate that the trend is not tied to the primary classifier. Full per-attack tables are in Appendix D.

ImageNet-1k. Table IV reports representative ϵ values on PyTorch pretrained ResNet-50; the full $\epsilon \in [0.1, 1.0]$ at step 0.1 is in Appendix D, Table XVII. Both $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ achieve $\geq 98.8\%$ TPR while keeping FPR within 9–55% even at $\epsilon=1.0$, where the Cohen detector saturates to 100% FPR. On the certificate, $\mathbb{D}_{lc}$ formally certifies 64–80% of all successful in-budget adversarial examples while empirically rejecting almost all of them. The magnitude is comparable to reported RS certified-accuracy rates on the same backbone, but the metrics measure different events and are not directly interchangeable.

b) ViT-B/16 vs. a SOTA empirical detector (A^2D).:

We further evaluate on a PyTorch pretrained ViT-B/16 and compare against A^2D [26], a recent state-of-the-art empirical adversarial detector. Table IVb (right half of Table IV) shows that our $\mathbb{D}_{lc}$ reaches 100% TPR at 3.6% FPR for $\epsilon=0.1$ on ViT-B/16, while A^2D plateaus at 89.5% TPR with 11.2% FPR. $\mathbb{D}_{lc}$ maintains a higher TPR than A^2D at every ϵ and a lower FPR at the small budgets ($\epsilon \leq 0.3$) most relevant to security deployments; unlike A^2D , which provides no formal guarantee, $\mathbb{D}_{lc}$ additionally certifies at least 63% of all successful in-budget adversarial examples.

c) Conservative RS detector.: The “RS (Conservative)” columns of Table III report a stronger RS variant that combines two certified-robustness checks: (i) compare the base classifier’s prediction $f(z)$ to the smoothed prediction $f_s(z)$ and flag mismatches; (ii) for those flagged inputs compute the certified radius r [10] and output ADVERSARIAL iff $r < \epsilon$. The conservative variant raises TPR to 93–100%, but FPR exceeds 93% across every ϵ , effectively rejecting nearly every clean input. Even a strengthened RS detector therefore retains a substantially worse safety–usability trade-off than our certified detection.

d) Comparison with CTT.: CTT [19] is the closest prior certified-detection framework, but it certifies a complementary specificity-side property. Under ℓ_2 FGSM and PGD on MNIST and CIFAR-10, our detectors match or beat CTT in TPR at comparable FPR without retraining; the full comparison is in Appendix D, Table XI.

Trends across perturbation scales. Empirical TPR climbs to 100% within a small budget on all datasets; FPR rises smoothly with ϵ , as the detection band must widen to admit the larger adversarial radius. Full perturbation-scale curves are in Appendix D, Figure 2. The per-sample detection-statistic distribution for $\mathbb{D}_{lc}$ is also reported in Appendix D, Figure 3; the clean and adversarial distributions are clearly separated across FGSM, PGD, BIM, and CW. Appendix D, Table XVI, further reports the realised L'/L^* distribution on CIFAR-

TABLE I: **Fully certified detection on 1-Lipschitz SortNet (CIFAR-10).** With L_{glob} exactly known, Lemma IV.1 (using $L = L_{\text{glob}}$) achieves 100% *certified* detection at FPR substantially below the baseline detectors built on SortNet’s and CRM’s certified radii. Our approximated detectors $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ use the local estimator L' and reduce FPR further; we report cert% as the fraction of all successful in-budget adversarial examples that are both rejected and formally certified.

ϵ	SortNet [16]		CRM [28]		Lem IV.1, L_{glob} (ours)		$\mathbb{D}_{Adv}$ (ours)			$\mathbb{D}_{lc}$ (ours)		
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	cert%	TPR	FPR	cert%
0.1	100.0	22.0	100.0	25.2	100.0	10.5	99.0	9.3	58.1	100.0	10.1	36.5
0.3	100.0	59.0	100.0	68.0	100.0	31.5	100.0	21.3	59.0	100.0	25.8	41.0
0.5	100.0	86.0	100.0	96.0	100.0	58.5	99.5	35.3	65.3	100.0	40.7	50.0
1.0	100.0	100.0	100.0	100.0	100.0	95.0	100.0	59.0	71.5	100.0	63.5	57.5

TABLE II: **MNIST (LeNet-5): detection at representative ℓ_2 budgets.** TPR on BIM; FPR on clean inputs; cert% is the fraction of all successful in-budget adversarial examples that are both rejected and formally certified ($L' \geq L^*$). RS (Cohen) reported at $\sigma=0.25$. Full per-attack and per- σ results in Appendix D, Tables XII–XIII.

ϵ	$\mathbb{D}_{Adv}$ (ours)			$\mathbb{D}_{lc}$ (ours)			RS (Cohen)	
	TPR	FPR	cert%	TPR	FPR	cert%	TPR	FPR
0.1	100.0	0.5	92.0	100.0	1.0	90.0	44.5	4.0
0.5	100.0	8.0	94.0	100.0	8.5	86.0	100.0	22.0
1.0	100.0	17.0	97.0	100.0	17.5	80.0	100.0	98.0
2.0	100.0	47.0	98.0	100.0	47.5	75.0	100.0	98.0

10. Most samples concentrate near or above the sufficient threshold 1, and the below-threshold cases remain close to 1 on average, explaining why the practical detector keeps high empirical TPR while cert% captures only the formally certified subset.

C. Stronger L' Estimation via Multi-Probe Search

To increase certified coverage without changing the detector families, we replace the single-probe local estimate with $L'_{\text{strong}} = \max_k L'_k$. A larger L' makes $L' \geq L^*$ easier to satisfy but also enlarges the rejection threshold. Table VI shows this trade-off: cert% reaches 99.5–100% on MNIST, gains 14.5–34.8 points on CIFAR-10, and gains up to 66.3 points on ImageNet, with at most 7.3 FPR points added and negligible TPR change. Table V gives cross-dataset latency; Appendix D, Table XV, shows diminishing returns beyond $K=1, T=6$.

D. Adaptive Attack Evaluation

a) *Adaptive strategy.*: The detectors fire when the observed margin is *too small* for the declared budget (Lemma IV.1). A detector-aware attacker therefore seeks a misclassification with the largest possible wrong-class margin. Since $q(\mathbf{x})$ is fixed, maximizing the left-hand side of Lemma IV.1, $q(\mathbf{x} + \delta) + q(\mathbf{x})$, is exactly maximizing $q(\mathbf{x} + \delta)$:

$$\max_{\|\delta\|_2 \leq \epsilon} q(\mathbf{x} + \delta) + q(\mathbf{x}),$$

using 100 Adam steps with ℓ_2 projection and $[0, 1]^d$ clipping. We write $\hat{\mathbf{x}}_{\text{adp}} = \mathbf{x} + \delta$ for the attacker-side input, distinct from the defender-side surrogate $\mathbf{z}'$.

We report the classification attack success rate before detection as ASR_{cls} , and the end-to-end evasion rate after detection as $\text{ASR}_{\text{joint}} = \text{ASR}_{\text{cls}} \cdot (1 - \text{TPR})$.

Table VII shows that the adaptive objective raises classification ASR over BIM, especially on ImageNet, yet end-to-end evasion remains low: $\text{ASR}_{\text{joint}} = 0$ on CIFAR-10 and at most 2.8%/0.9% for $\mathbb{D}_{Adv}/\mathbb{D}_{lc}$ on ImageNet. The stronger multi-probe estimate restores much of the certified subset.

E. Robustness-Oriented Training Improves Detection

A central implication of our framework is that robust-training advances that enlarge the underlying classifier’s clean-margin distribution can improve the detector’s TPR/FPR frontier. We confirm this empirically on two complementary robust-training primitives: ℓ_2 PGD adversarial training and Gaussian-augmented training.

a) *Adversarial training.*: We adversarially train a CIFAR-10 VGG16 classifier with ℓ_2 PGD at radius 0.5 and apply $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ on top of the trained model. Compared to the standard-trained baseline (Table III), FPR drops sharply at the same TPR — e.g., $\mathbb{D}_{lc}$ ’s FPR at $\epsilon=0.5$ falls from 93.6% to 32.2% while $\mathbb{D}_{lc}$ ’s TPR remains $\geq 99\%$ across the reported ϵ on ℓ_2 PGD (Table VIII). Adversarial training therefore acts as a margin-enlarging primitive that translates directly into a better TPR/FPR Pareto frontier under our certificate.

b) *Gaussian augmentation.*: A second standard primitive is training with Gaussian-augmented data ($\sigma=0.25$). Table VIII compares our detectors against RS (Cohen) and the conservative-RS variant on CIFAR-10 (ℓ_2 PGD): noise augmentation lifts RS’s TPR but its FPR stays substantially higher than ours at every reported ϵ .

Full robust-training and Gaussian-augmentation results are in Table VIII.

Taken together, both settings make a single point: robust training is a compounding lever on our detection. Stronger base-model margins improve the TPR/FPR frontier without changing the detector, so robustness gains translate directly into detection gains.

Efficiency. Table V reports per-sample detection time (mean $\pm$ std over 500 inputs). $\mathbb{D}_{Adv}$ is the fastest base detector (single PGD pass), $\mathbb{D}_{lc}$ adds centroid optimisation on top, and

TABLE III: **CIFAR-10: detection at representative ℓ_2 budgets on VGG16.** TPR on BIM attacks (averaged over ~ 500 successful adversarials per ϵ); FPR on the full CIFAR-10 test set; cert% is the fraction of all successful in-budget adversarial examples that are both rejected and formally certified ($L' \geq L^*$). RS (Conservative) is the strengthened RS variant that combines smoothed-classifier mismatch with the certified radius (Section V). Full per-attack results in Appendix D.

ϵ	$\mathbb{D}_{Adv}$ (ours)			$\mathbb{D}_{Ic}$ (ours)			RS (Cohen)		RS (Conservative)		ContraNet	
	TPR	FPR	cert%	TPR	FPR	cert%	TPR	FPR	TPR	FPR	TPR	FPR
0.05	100.0	15.3	63.2	100.0	15.6	62.4	38.5	46.0	93.5	93.2	87.0	9.0
0.1	100.0	29.5	76.4	100.0	29.9	75.0	53.0	60.2	95.0	93.8	87.8	9.0
0.3	100.0	74.8	83.4	100.0	75.8	80.4	95.8	96.0	99.6	99.2	74.6	9.0
0.5	100.0	92.7	82.6	100.0	93.6	78.8	100.0	100.0	100.0	100.0	85.0	9.0

TABLE IV: **Detection on ImageNet-1k (500 images) with PyTorch pretrained models.** TPR on ℓ_2 BIM attacks; FPR on clean samples; cert% is the fraction of all successful in-budget adversarial examples that are both rejected and formally certified ($L' \geq L^*$). The full $\epsilon \in [0.1, 1.0]$ breakdown for ResNet-50 is in Appendix D, Table XVII.

(a) ResNet-50 pretrained model.										(b) ViT-B/16 pretrained model (vs. A ² D [26]).									
ϵ	$\mathbb{D}_{Adv}$ (ours)			$\mathbb{D}_{Ic}$ (ours)			RS (Cohen)			ϵ	$\mathbb{D}_{Adv}$ (ours)			$\mathbb{D}_{Ic}$ (ours)			A ² D		
	TPR	FPR	cert%	TPR	FPR	cert%	TPR	FPR			TPR	FPR	cert%	TPR	FPR	cert%	TPR	FPR	
0.1	99.5	9.0	20.4	99.7	9.4	63.8	34.0	17.0		0.1	97.4	2.8	35.1	100.0	3.6	63.0	89.5	11.2	
0.3	99.8	24.6	42.9	99.6	23.7	72.7	62.5	36.0		0.3	97.5	6.5	38.0	99.0	8.2	67.3	90.3	11.2	
0.5	99.6	34.4	38.1	99.3	32.8	80.4	85.5	51.0		0.5	97.0	14.0	39.8	99.0	15.4	72.3	90.3	11.2	
1.0	99.8	54.8	41.9	99.8	53.5	76.3	100.0	100.0		1.0	98.0	25.0	58.8	99.3	26.6	77.5	93.1	11.2	

TABLE V: Average per-sample detection time (seconds, mean $\pm$ std over $n=500$ inputs). Experiments are conducted on a single NVIDIA V100 GPU.

Method	MNIST	CIFAR-10	ImageNet-1k
$\mathbb{D}_{Adv}$ (base)	0.020 $\pm$ 0.002	0.071 $\pm$ 0.023	0.190 $\pm$ 0.073
$\mathbb{D}_{Adv}$ (strong L')	0.048 $\pm$ 0.002	0.183 $\pm$ 0.017	0.451 $\pm$ 0.084
$\mathbb{D}_{Ic}$ (base)	0.063 $\pm$ 0.006	0.151 $\pm$ 0.026	0.421 $\pm$ 0.074
$\mathbb{D}_{Ic}$ (strong L')	0.155 $\pm$ 0.008	0.341 $\pm$ 0.036	0.765 $\pm$ 0.087
RS (Cohen)	0.335 $\pm$ 0.003	0.437 $\pm$ 0.132	3.757 $\pm$ 0.502

both are markedly faster than RS (Cohen) at every scale. The stronger L' estimator increases latency by roughly 1.8–2.6 $\times$, but remains below RS on all three datasets; on ImageNet-1k, strong $\mathbb{D}_{Adv}$ and strong $\mathbb{D}_{Ic}$ are still about 8.3 $\times$ and 4.9 $\times$ faster than RS, respectively.

VI. DISCUSSION

A. Three Levels of Certification

A natural concern with our framework is the gap between the formal certificate (Lemma IV.1, Theorem IV.2) and the approximated detectors $\mathbb{D}_{Adv}$, $\mathbb{D}_{Ic}$ that use the local estimator L' . We make this gap explicit by distinguishing three levels of certification, each with its own empirical signature in our results.

Level 1 — Fully certified. On a network whose global Lipschitz constant L_{glob} is exactly known (e.g., 1-Lipschitz SortNet [16]), Lemma IV.1 with $L = L_{\text{glob}}$ delivers a formal certificate at the declared radius: every successful in-budget adversarial example is provably refused. Since $q(\mathbf{x})$ is unobserved at inference, we lower-bound it by $q(\mathbf{x}) \geq 0$ (from clean correct classification), giving the conservative test $q(\hat{\mathbf{x}}) < c L_{\text{glob}} \|\delta\|$, which remains a sound certificate. Our

SortNet evaluation (Table I) realises this level, attaining 100% TPR at FPR substantially below the baseline detectors built on SortNet’s and CRM’s own certified radii.

Level 2 — Conditionally certified ($L' \geq L^*$). On standard pretrained classifiers L_{glob} is intractable, so we replace it with the local estimator L' . The certificate continues to hold for any specific successful in-budget adversarial example whenever the per-detector L' condition holds on the relevant pair (Proposition IV.7). We quantify this subset by cert%, defined over all successful in-budget adversarial examples rather than only over rejected ones; on PyTorch pretrained ImageNet models, cert% reaches 63–80% for $\mathbb{D}_{Ic}$. This number is comparable in magnitude to RS certified accuracy on the same backbone (cf. Cohen et al. [10]), but the two metrics measure different events and are not directly comparable.

Level 3 — Empirical protection. The difference between TPR and cert% is the empirically rejected but not formally certified subset. These examples are still protected by the deployed reject decision in our experiments (TPR $\geq 99\%$ on ImageNet), but we do not call them certified detections unless the per-detector L' condition is verified. Our empirical TPR matches or exceeds dedicated empirical detectors (A²D [26], ContraNet [24]) on the same evaluation, so the empirical layer is competitive with SOTA empirical detection while remaining separate from the formal guarantee.

B. FPR Growth with the Declared Radius

FPR grows with ϵ because larger declared radii require wider rejection bands, mirroring the certified-robustness frontier. Thus the defender selects the largest protection radius whose clean-FPR budget is acceptable, while cert% reports which successful in-budget attacks are also formally certified.

TABLE VI: **Stronger L' estimation improves certified coverage.** Only the local estimate is replaced by a multi-probe L'_{strong} ($K=1, T=6$ unless stated), trading cert% gains for FPR/runtime.

Dataset	ϵ	$\mathbb{D}_{Adv}$			$\mathbb{D}_{Lc}$		
		ΔTPR	ΔFPR	cert% base $\rightarrow$ strong	ΔTPR	ΔFPR	cert% base $\rightarrow$ strong
MNIST	0.50	+0.0	+0.9	92.0 $\rightarrow$ 100.0	+0.0	+1.5	94.0 $\rightarrow$ 100.0
	1.00	+0.0	+2.3	97.0 $\rightarrow$ 99.5	+0.0	+2.5	98.0 $\rightarrow$ 100.0
CIFAR-10	0.05	+0.0	+0.4	63.2 $\rightarrow$ 98.0	+0.0	+1.0	62.4 $\rightarrow$ 87.0
	0.10	+0.0	+2.0	76.4 $\rightarrow$ 93.5	+0.0	+2.0	75.0 $\rightarrow$ 89.5
	0.30	+0.0	+5.8	83.4 $\rightarrow$ 98.0	+0.0	+7.3	80.4 $\rightarrow$ 97.0
	0.50	+0.0	+5.7	82.6 $\rightarrow$ 100.0	+0.0	+5.0	78.8 $\rightarrow$ 100.0
ImageNet	0.10	+0.0	+1.0	20.2 $\rightarrow$ 86.5	+0.3	+1.9	64.0 $\rightarrow$ 82.0
	0.50	+0.2	+4.7	38.2 $\rightarrow$ 95.0	+0.5	+5.8	80.8 $\rightarrow$ 89.6

TABLE VII: **Detector-aware adaptive attack evaluation.** $\text{ASR}_{\text{joint}} = \text{ASR}_{\text{cls}} \cdot (1 - \text{TPR})$ is end-to-end evasion after detection; cert% is shown before/after stronger L' .

Dataset	ϵ	BIM ASR	Adp. ASR	$\mathbb{D}_{Adv}$			$\mathbb{D}_{Lc}$		
				TPR	$\text{ASR}_{\text{joint}}$	cert% base $\rightarrow$ strong	TPR	$\text{ASR}_{\text{joint}}$	cert% base $\rightarrow$ strong
CIFAR-10	0.05	8.9	11.4	100.0	0.0	3.0 $\rightarrow$ 96.5	100.0	0.0	5.0 $\rightarrow$ 67.5
	0.10	20.3	26.1	100.0	0.0	9.0 $\rightarrow$ 98.0	100.0	0.0	8.5 $\rightarrow$ 77.5
	0.30	61.0	80.7	100.0	0.0	16.5 $\rightarrow$ 97.5	100.0	0.0	17.0 $\rightarrow$ 97.0
	0.50	78.5	98.5	100.0	0.0	24.0 $\rightarrow$ 99.5	100.0	0.0	30.5 $\rightarrow$ 99.0
ImageNet	0.10	4.7	14.1	87.4	1.8	2.8 $\rightarrow$ 89.1	99.0	0.1	11.6 $\rightarrow$ 92.6
	0.30	14.3	33.6	91.6	2.8	7.4 $\rightarrow$ 92.8	97.2	0.9	20.0 $\rightarrow$ 95.6
	0.50	23.8	46.6	94.0	2.8	12.6 $\rightarrow$ 92.5	98.0	0.9	25.4 $\rightarrow$ 95.7
	1.00	45.8	70.3	98.8	0.8	20.5 $\rightarrow$ 94.1	99.6	0.3	33.2 $\rightarrow$ 96.4

TABLE VIII: **Robust training boosts our detection on CIFAR-10 (VGG16).** TPR (%) on ℓ_2 PGD attacks; FPR (%) on clean inputs. Two robust-training primitives applied to the same VGG16 base model: ℓ_2 PGD adversarial training (radius 0.5, AT columns) and Gaussian augmentation ($\sigma=0.25$). RS baselines do not apply to PGD-adversarial-trained models.

ϵ	Adv-trained (ℓ_2 PGD, $r=0.5$)				Gaussian aug ($\sigma=0.25$)							
	$\mathbb{D}_{Adv}$ (AT)		$\mathbb{D}_{Lc}$ (AT)		$\mathbb{D}_{Adv}$		$\mathbb{D}_{Lc}$		RS (Cohen)		RS (Conservative)	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
0.05	92.0	4.0	99.0	4.0	99.1	5.8	99.5	7.2	38.4	17.0	91.5	27.2
0.1	94.9	7.0	99.2	7.6	99.6	11.2	100.0	13.8	47.6	21.4	93.8	30.8
0.3	97.0	18.0	99.0	20.0	99.8	29.5	100.0	35.4	80.8	43.8	98.6	46.4
0.5	98.4	29.8	99.6	32.2	100.0	43.9	100.0	51.5	95.8	65.0	100.0	65.8

On SortNet at $\epsilon=0.5$, our fully certified detector keeps FPR at 35–41%, versus 86%/96% for SortNet/CRM certified-radius detectors; adversarial training cuts CIFAR-10 $\mathbb{D}_{Lc}$ FPR from 93.6% to 32.2%.

C. Which detector certifies more, and when

The two approximated detectors do not dominate one another on cert%; which one is tighter depends on the underlying classifier’s clean-margin scale. On Lipschitz-constrained 1-Lipschitz SortNet (Table I), $\mathbb{D}_{Adv}$ certifies more: 58–72% versus 37–58% for $\mathbb{D}_{Lc}$. On unconstrained pretrained ImageNet classifiers (Table IV), the order reverses: $\mathbb{D}_{Lc}$ certifies 63–80% versus 20–60% for $\mathbb{D}_{Adv}$. Standard CIFAR-10 VGG16 falls in between, with the two within 1–4 percentage points of each other (Table III).

cert% tracks local Lipschitz stability rather than dataset complexity alone. If data complexity were the dominant driver, one would expect a sharp drop from CIFAR-10 to ImageNet-1k; instead the $\mathbb{D}_{Lc}$ certified fraction is comparable across the two settings, while larger variation appears across architectures

on the same dataset and across constrained versus unconstrained backbones. Table XIV adds ResNet-18 and VGG-style CNN results that reproduce the same profile on an architecture variation for each small-scale dataset.

The structural reason is that $L^*_{(\mathbf{x}, \hat{\mathbf{x}})}$ for $\mathbb{D}_{Adv}$ scales with $q(\mathbf{x}) + q(\hat{\mathbf{x}})$, while $L^*_{(\hat{\mathbf{x}}, \mathbf{x})}$ for $\mathbb{D}_{Lc}$ scales with $q(\hat{\mathbf{x}}) + q(\mathbf{x})$. On a 1-Lipschitz network, $q(\hat{\mathbf{x}})$ is bounded by the Lipschitz constant times the centroid travel and cannot escape the model’s clean-margin scale; the simpler $\mathbb{D}_{Adv}$ certificate (whose two terms live on the same short-distance pair) is in fact tighter. On deep unconstrained classifiers, $q(\mathbf{x})$ on a confident clean input is very large, $q(\hat{\mathbf{x}})$ scales with it (the centroid moves through the network’s high-margin region), and the centroid construction allows L' to track L^* rather than being dominated by the clean-margin term. In effect, the centroid’s value as a high-margin anchor compounds with the model’s own margin headroom; on Lipschitz-constrained networks there is no headroom to compound. We deploy the two detectors as a complementary pair: $\mathbb{D}_{Adv}$ when a

tight global Lipschitz constant is available, $\mathbb{D}_{lc}$ on standard pretrained classifiers.

D. Beyond ℓ_2 : ℓ_∞ Extension

Lemma IV.1 extends to any ℓ_p norm with $c = 2^{(p-1)/p}$. We empirically tested an ℓ_∞ instantiation ($c=2$, ℓ_∞ PGD as the surrogate, $\|\cdot\|_\infty$ for distances) on standard CIFAR-10 VGG16: $\mathbb{D}_{lc}$ reaches TPR 99.5% at $\epsilon=1.5/255$ (FPR 80%) and 100% at $\epsilon=2/255$ (FPR 90.5%). TPR confirms that our reduction transfers, but FPR is much higher than in the ℓ_2 regime. The reason is purely norm-geometric: since $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty$, the ℓ_∞ Lipschitz constant of a network is substantially larger than its ℓ_2 counterpart, making L' approximation looser and the detection threshold more permissive. Tightening the ℓ_∞ certificate — for instance via norm-specific Lipschitz estimators or per-dimension bounds — is a natural direction we leave for future work.

E. Practical Considerations and Limitations

We close by discussing three considerations that arise when deploying our framework, each with its own empirical or design-side mitigation.

Coupling to the classifier’s clean margin. The detector’s FPR is governed by the classifier’s clean-margin distribution. This follows from the Lipschitz-margin reduction: the margin distribution determines how tight the rejection band can be. Margin theory links larger normalized margins to robustness [56], [57]; empirically, ViT-B/16 improves the ImageNet FPR profile over ResNet-50, and adversarial training cuts FPR sharply without changing the detector.

Single-pair L' estimation. Because L' is computed from one perturbed pair (z, z') , Theorem IV.5 bounds the induced threshold gap. Under the revised denominator, pretrained-model cert% remains 63–80% for $\mathbb{D}_{lc}$. Tables XVI and X (Appendix D) show that L'/L^* concentrates near or above 1 with mild violations; Table VI shows the multi-probe cost/cert% trade-off. Hybrid pairwise or layer-wise estimators are compatible with our framework.

Norm scope. Our primary focus is ℓ_2 , which already covers many security-relevant adversarial threats — physical-world stickers on traffic signs, frequency-domain perturbations, low-energy patches — via principled ℓ_2 formulations. The ℓ_∞ extension is sketched in Section VI-D.

Attention-based architectures. We do not rely on a global Lipschitz bound, which can be loose or unbounded for softmax attention outside a bounded input region [58]. Our certificate needs only the local pairwise constant L^* on the bounded ball $B(x, \epsilon)$, where both L^* and L' are finite; the guarantee on attention models is therefore conditional on $L' \geq L^*$ in the same way as other unconstrained pretrained classifiers. The ViT-B/16 results in Table IVb support this scope, with cert% at least 63% alongside near-perfect empirical TPR.

Beyond image classification. The reduction requires a scalar score margin and a norm-bounded perturbation model, not an image-specific architecture. Extending the same certificate to detection, segmentation, or multimodal pipelines is

therefore conceptually feasible whenever an appropriate task margin can be defined, but certified robustness for those tasks is less mature; we leave such extensions to future work.

VII. CONCLUSION

We propose a detection-oriented perspective on defending DNNs against adversarial examples: rather than certifying that predictions remain unchanged, we certify that adversarial inputs are detected when the stated Lipschitz conditions hold. We prove via Lipschitz-margin analysis that certified ℓ_2 robustness implies certified adversarial detection (Lemma IV.1, Theorem IV.2), and instantiate two practical detector families, $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$, that operate on any pretrained classifier without retraining or architectural change. Across MNIST, CIFAR-10, and ImageNet-1k on four backbones, the framework matches or outperforms five baselines (CTT, SortNet, CRM, ContraNet, A²D), empirically rejects nearly all measured attacks, and formally certifies 63–80% of successful in-budget ImageNet adversarial examples under the revised cert% metric. Furthermore, robust training can improve the operating point of the same detector families by enlarging margins.

REFERENCES

- [1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.
- [2] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 1528–1540.
- [3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
- [4] P. Jing, Q. Tang, Y. Du, L. Xue, X. Luo, T. Wang, S. Nie, and S. Wu, “Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 3237–3254.
- [5] L. Li, T. Xie, and B. Li, “Sok: Certified robustness for deep neural networks,” in *2023 IEEE symposium on security and privacy (SP)*. IEEE, 2023, pp. 1289–1310.
- [6] T.-W. Weng, P. Zhao, S. Liu, P.-Y. Chen, X. Lin, and L. Daniel, “Towards certified model robustness against weight perturbations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6356–6363.
- [7] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, “Certified robustness to adversarial examples with differential privacy,” in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 656–672.
- [8] M. Pautov, N. Tursynbek, M. Munkhoeva, N. Muravev, A. Petiushko, and I. Oseledets, “Cc-cert: A probabilistic approach to certify general robustness of neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7975–7983.
- [9] C. Li, S. Ji, H. Weng, B. Li, J. Shi, R. Beyah, S. Guo, Z. Wang, and T. Wang, “Towards certifying the asymmetric robustness for neural networks: quantification and applications,” *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 3987–4001, 2021.
- [10] J. Cohen, E. Rosenfeld, and Z. Kolter, “Certified adversarial robustness via randomized smoothing,” in *international conference on machine learning*. PMLR, 2019, pp. 1310–1320.
- [11] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, “Provably robust deep learning via adversarially trained smoothed classifiers,” *Advances in neural information processing systems*, vol. 32, 2019.

- [12] G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 693–10 705.
- [13] B. Li, C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive noise," *Advances in neural information processing systems*, vol. 32, 2019.
- [14] N. Carlini, F. Tramèr, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter, "(certified!!) adversarial robustness for free!" in *The Eleventh International Conference on Learning Representations*. OpenReview, 2023.
- [15] B. Delattre, A. Araujo, Q. Barthélemy, and A. Allauzen, "The lipschitz-variance-margin tradeoff for enhanced randomized smoothing," in *ICLR International Conference on Learning Representations*, 2024.
- [16] B. Zhang, D. Jiang, D. He, and L. Wang, "Rethinking lipschitz neural networks and certified robustness: A boolean function perspective," *Advances in neural information processing systems*, vol. 35, pp. 19 398–19 413, 2022.
- [17] K. Leino, Z. Wang, and M. Fredrikson, "Globally-robust neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 6212–6222.
- [18] Q. Li, S. Haque, C. Anil, J. Lucas, R. Grosse, and J.-H. Jacobsen, "Preventing Gradient Attenuation in Lipschitz Constrained Convolutional Networks," Nov. 2019, arXiv:1911.00937 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1911.00937>
- [19] I. Shumailov, Y. Zhao, R. Mullins, and R. Anderson, "Towards certifiable adversarial sample detection," in *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security*, 2020, pp. 13–24.
- [20] F. Tramer, "Detecting adversarial examples is (nearly) as hard as classifying them," in *International Conference on Machine Learning*. PMLR, 2022, pp. 21 692–21 702.
- [21] Y. Gao, H. Peng, H. Ma, Z. Dai, S. Wang, H. Hu, A. Fu, and M. Xue, "From pixels to trajectory: Universal adversarial example detection via temporal imprints," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–17, 2026.
- [22] S. Ma and Y. Liu, "Nic: Detecting adversarial samples with neural network invariant checking," in *Proceedings of the 26th network and distributed system security symposium (NDSS 2019)*, 2019.
- [23] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: a review and experimental comparison," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022. [Online]. Available: <https://doi.org/10.1007/s10462-021-10125-w>
- [24] Y. Yang, R. Gao, Y. Li, Q. Lai, and Q. Xu, "What you see is not what the network infers: Detecting adversarial examples based on semantic contradiction," in *Proceedings of the 29th network and distributed system security symposium (NDSS 2022)*, 2022.
- [25] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, and F. Roli, "Deep neural rejection against adversarial examples," *EURASIP Journal on Information Security*, vol. 2020, pp. 1–10, 2020.
- [26] Z. Zhao, G. Chen, T. Liu, T. Li, F. Song, J. Wang, and J. Sun, "Attack as detection: Using adversarial attack methods to detect abnormal examples," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 3, pp. 1–45, 2024.
- [27] Y. Lai, Y. Zhu, Y. Sun, Y. Wu, B. Xiao, G. Li, J. Li, and K. Zhou, "Auditvotes: A framework towards more deployable certified robustness for graph neural networks," *arXiv preprint arXiv:2503.22998*, 2025.
- [28] M. Fazlyab, T. Entesari, A. Roy, and R. Chellappa, "Certified robustness via dynamic margin maximization and improved lipschitz regularization," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 451–34 464, 2023.
- [29] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for dnn models: A review and experimental comparison," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4403–4462, 2022.
- [30] C. Zhao, H. Li, D. Wang, and R. Liu, "Adversarial example detection for deep neural networks: A review," in *2023 8th International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2023, pp. 468–475.
- [31] E. Mizrahi, R. Lapid, and M. Sipper, "Pulling back the curtain: Unsupervised adversarial detection via contrastive auxiliary networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2294–2305.
- [32] D. Bhusal, M. T. Alam, M. K. Veerabhadran, M. Clifford, S. Rampazzi, and N. Rastogi, "Pasa: Attack agnostic unsupervised adversarial detection using prediction & attribution sensitivity analysis," in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2024, pp. 21–40.
- [33] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in neural information processing systems*, vol. 33, pp. 1633–1645, 2020.
- [34] J. Wan, J. Fu, L. Wang, and Z. Yang, "Bounceattack: A query-efficient decision-based adversarial attack by bouncing into the wild," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023, pp. 68–68.
- [35] S. Goyal, K. D. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli, "Scalable verified training for provably robust image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4842–4851.
- [36] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [37] M. N. Mueller, F. Eckert, M. Fischer, and M. Vechev, "Certified training: Small boxes are all you need," in *The Eleventh International Conference on Learning Representations*, 2023.
- [38] X. Xu, L. Li, and B. Li, "Lot: Layer-wise orthogonal training on improving l2 certified robustness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 904–18 915, 2022.
- [39] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," *Advances in neural information processing systems*, vol. 30, 2017.
- [40] Y. Tsuzuku, I. Sato, and M. Sugiyama, "Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.
- [41] Y. Xu, Y. Sun, M. Goldblum, T. Goldstein, and F. Huang, "Exploring and exploiting decision boundary dynamics for adversarial robustness," *arXiv preprint arXiv:2302.03015*, 2023.
- [42] K. Lyu and J. Li, "Gradient descent maximizes the margin of homogeneous neural networks," in *International Conference on Learning Representations*, 2019.
- [43] K. Lyu, Z. Li, R. Wang, and S. Arora, "Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 12 978–12 991. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/6c351da15b5e8a743a21ee96a86e25df-Abstract.html
- [44] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.
- [45] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.
- [46] Z. Yan, Y. Guo, and C. Zhang, "Adversarial Margin Maximization Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1129–1139, Apr. 2021. [Online]. Available: <https://ieeexplore.ieee.org/document/8877866/>
- [47] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [48] Q. Zhou, Z. Wei, H. Wang, B. Jiang, and W.-K. Chan, "Crosscert: A cross-checking detection approach to patch robustness certification for deep learning models," *Proceedings of the ACM on Software Engineering*, vol. 1, no. FSE, pp. 2725–2746, 2024.
- [49] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. Ieee, 2017, pp. 39–57.
- [50] H. Wang, Z. Xiang, D. J. Miller, and G. Kesidis, "Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 1994–2012.
- [51] Y. LeCun, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998.
- [52] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.
- [53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

- [54] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [55] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.
- [56] P. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," Dec. 2017, arXiv:1706.08498 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1706.08498>
- [57] J. Sokolic, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, "Robust Large Margin Deep Neural Networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 16, pp. 4265–4280, Aug. 2017, arXiv:1605.08254 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/1605.08254>
- [58] H. Kim, G. Papamakarios, and A. Mnih, "The lipschitz constant of self-attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5562–5571.
- [59] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

TABLE IX: Notation summary.

Symbol	Meaning
<i>Inputs and perturbations</i>	
$\mathbf{x}$	Clean input.
y	True class of $\mathbf{x}$.
δ	Adversarial perturbation.
ϵ	Perturbation budget.
$\hat{\mathbf{x}}$	Adversarial example, $\mathbf{x} + \delta$.
z	Inference-time test input.
z'	Defender's perturbation of z .
<i>Network and margin</i>	
f	DNN classifier.
$f_i(\mathbf{x})$	i -th class score.
$q(u; k)$	Class- k margin: $f_k(u) - \max_{j \neq k} f_j(u)$.
$q(u)$	Shorthand for $q(u; \arg \max_k f_k(u))$ (own-label, ≥ 0).
$\hat{y}$	Input pseudo-label $\arg \max f(z)$, fixed reference.
$\tilde{\mathbf{x}}$	Local centroid of $\mathbf{x}$.
$\hat{\tilde{\mathbf{x}}}$	Local centroid of $\hat{\mathbf{x}}$.
<i>Lipschitz quantities</i>	
$\ \cdot\ _p$	ℓ_p norm; default $p=2$.
c	Norm constant $2^{(p-1)/p}$.
L	Lipschitz constant of f .
L_{glob}	Global Lipschitz constant.
L'	Local Lipschitz estimator.
L^*	Optimal pairwise Lipschitz.
$L^*_{(\mathbf{x}, \hat{\mathbf{x}})}$	Pairwise on $(\mathbf{x}, \hat{\mathbf{x}})$.
$L^*_{(\tilde{\mathbf{x}}, \mathbf{x})}$	Pairwise on (centroid, clean).
<i>Detectors and metrics</i>	
$\mathbb{D}$	Detector, $\{0, 1\}$ output.
$\mathbb{D}_{Adv}$	Via Claim 1.
$\mathbb{D}_{lc}$	Via Claim 2.
TPR	Adversarial-detection rate.
FPR	Clean-rejection rate.
cert%	Adversarials both rejected and formally certified.

APPENDIX

A. Ethical Considerations

This work develops an adversarial example detection method that can be deployed in situations like Autonomous-driving perception, content moderation, and access control. While our detectors are shown to have almost 100% recall, similar to certified robustness literature, we acknowledge that there exists false negatives when the adversary have large perturbation budget in reality. Our framework is designed to detect adversarial examples with limit ℓ_2 perturbations, clean samples are also possible to be detected. However, the service provider in our situation already have a meaningful build-in fallback for refused inputs.

B. Notation Reference

Table IX summarizes the symbols used throughout the paper. Definitions in the main text are authoritative; this table is a quick lookup.

TABLE X: **Empirical tightness of Δ_T on SortNet.** Median actual Δ_T and its share of the per-sample theoretical envelope from (9), computed on detected adversarials with $L' \leq L_{\text{glob}}$. The remaining samples satisfy $L' \geq L_{\text{glob}} \geq L^*$ and are trivially certified.

ϵ	$\mathbb{D}_{Adv}$	$\mathbb{D}_{lc}$
0.3	0.31 (74%)	0.20 (16%)
0.5	0.51 (72%)	0.29 (19%)
1.0	1.02 (63%)	0.85 (44%)

C. Attacks Settings

We use BIM with 10 steps, PGD with 40 steps and 10 random restarts, and CW- ℓ_2 with 40 steps. All attacks are the ℓ_2 version, or projected to the ℓ_2 ball of radius $\|\delta\|$.

D. Detailed Results across Datasets

This subsection reports the per- ϵ , per-attack and per-noise-level breakdowns that did not fit in the body. It also includes the Δ_T bias-validation table (Table X). The trends parallel the main-text summary: $\mathbb{D}_{lc}$ achieves near-perfect TPR across all settings, $\mathbb{D}_{Adv}$ trades a small TPR drop for substantially lower FPR, and the Cohen detector incurs FPR that escalates rapidly with ϵ .

Tables XII–XIII present per-attack TPR and clean FPR on MNIST and CIFAR-10 under noise levels $\sigma \in \{0.125, 0.25\}$. Table XVII extends the body's ImageNet-1k results (Table IV) to the full $\epsilon \in [0.1, 1.0]$ range at step 0.1.

Table X reports the SortNet bias-bound check used to validate the deterministic Δ_T envelope.

Table XIV adds the architecture variations on small-scale datasets; Table XI reports the complementary CTT comparison omitted from the body for space.

Figures 2 and 3 provide the full perturbation-scale curves and per-sample detection-statistic distributions summarized in Section V.

Table XVI and Table XV give the supporting diagnostics for the revised certified-coverage analysis: the realised L'/L^* distribution and the cost-tightness sweep for stronger L' estimation.

On **MNIST**, both $\mathbb{D}_{Adv}$ and $\mathbb{D}_{lc}$ attain 100% TPR across the full range; FPR stays below 3% for $\epsilon \leq 0.2$ and grows gradually thereafter (about 17% at $\epsilon=1.0$, 80% at $\epsilon=3.0$).

On **CIFAR-10**, the difference is more pronounced. $\mathbb{D}_{Adv}$ achieves 69–96% TPR with much lower FPR (7.5–70%); $\mathbb{D}_{lc}$ reaches near-perfect TPR ($\approx 100\%$) at the cost of higher FPR (16–100%) for large ϵ .

On **ImageNet-1k**, both our detectors stay $\geq 98.8\%$ TPR throughout, while keeping FPR under 55% even at $\epsilon=1.0$. Cohen exhibits FPR exceeding 50% once $\epsilon \geq 0.4$ and saturates at 100% for $\epsilon \geq 0.6$.

E. Proof of Lemma IV.1 and Theorem IV.2

Lemma A.1. (Proposition 3.1 in [16]) For a neural network f with Lipschitz constant L under ℓ_p -norm $\|\cdot\|_p$, f is provably

TABLE XI: Comparison of our approximated detectors with CTT [19] on MNIST and CIFAR-10. We report true positive rate (TPR, %) and false positive rate (FPR, %) under different attacks and perturbation levels. CTT-loose and CTT-strict correspond to two thresholding variants in [19].

Dataset	Attack	ϵ	$\mathbb{D}_{Adv}$ (Ours)		$\mathbb{D}_{Ic}$ (Ours)		CTT-loose		CTT-strict	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	FGSM	0.1	100.0	0.5	100.0	1.0	100.0	1.6	100.0	1.1
	FGSM	0.2	100.0	1.5	100.0	3.0	100.0	1.6	100.0	1.1
	PGD	0.1	100.0	0.5	100.0	1.0	100.0	1.6	100.0	1.1
	PGD	0.2	100.0	1.5	100.0	3.0	100.0	1.6	100.0	1.1
CIFAR-10	FGSM	0.02	95.0	2.5	100.0	6.0	95.7	6.4	96.5	3.0
	FGSM	0.04	93.0	5.0	100.0	13.0	93.6	6.4	93.1	3.0
	PGD	0.01	89.0	1.0	100.0	3.0	24.3	6.4	16.6	3.0
	PGD	0.02	96.0	2.5	100.0	6.0	70.3	6.4	49.9	3.0

TABLE XII: TPR (%) of different attacks under varying ℓ_2 budgets ϵ . For the Cohen detector [10], we report certified TPR against **BIM** under two noise levels.

Dataset	ϵ	$\mathbb{D}_{Adv}$				$\mathbb{D}_{Ic}$				Cohen	
		FGSM	PGD	BIM	CW	FGSM	PGD	BIM	CW	$\sigma = 0.125$	$\sigma = 0.25$
MNIST	0.05	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.0	33.0
	0.10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	44.5
	0.30	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	63.9
	0.50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1.50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	2.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	2.50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	3.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
CIFAR-10	0.05	99.8	100.0	100.0	98.8	99.8	100.0	100.0	99.8	67.0	38.5
	0.10	100.0	100.0	100.0	99.6	100.0	100.0	100.0	99.8	74.5	53.0
	0.20	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.8	89.4	70.5
	0.30	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.8
	0.40	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	0.50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	0.75	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLE XIII: Clean FPR (%) of our detectors vs. RS (Cohen) across perturbation budgets. Lower is better.

Dataset	ϵ	Ours		Cohen	
		$\mathbb{D}_{Adv}$	$\mathbb{D}_{Ic}$	$\sigma=0.125$	$\sigma=0.25$
MNIST	0.05	0.0	0.0	0.5	2.5
	0.10	0.5	1.0	1.5	4.0
	0.30	5.0	5.5	100.0	8.5
	0.50	8.0	8.5	100.0	22.0
	1.00	17.0	17.5	100.0	98.0
	1.50	25.5	28.0	100.0	98.5
	2.00	47.0	47.5	100.0	98.0
	2.50	68.0	68.5	100.0	98.5
	3.00	81.0	80.5	100.0	98.5
CIFAR-10	0.05	15.3	15.6	35.0	46.0
	0.10	29.4	29.9	40.5	60.2
	0.20	55.4	56.1	68.1	68.8
	0.30	74.7	75.7	100.0	96.0
	0.40	86.0	87.1	100.0	99.8
	0.50	92.7	93.5	100.0	99.8
	0.75	98.3	98.7	100.0	100.0
	1.00	99.6	99.7	100.0	100.0
	1.00	99.6	99.7	100.0	100.0

TABLE XIV: **Architecture variation on CIFAR-10 and MNIST.** BIM attacks with $n_{adv}=500$ and $n_{clean}=1000$. We evaluate CIFAR-10 on ResNet-18 and MNIST on a VGG-style CNN to check that the detection profile is not tied to the primary classifiers.

Dataset	Backbone	ϵ	TPR	FPR		cert%	
				$\mathbb{D}_{Adv}$	$\mathbb{D}_{Ic}$	$\mathbb{D}_{Adv}$	$\mathbb{D}_{Ic}$
CIFAR-10	ResNet-18	0.05	100.0	12.2	12.2	70.0	71.3
		0.10	100.0	29.7	29.7	86.0	85.3
		0.30	100.0	74.9	75.1	78.7	77.0
		0.50	100.0	90.5	91.0	71.7	67.7
MNIST	VGG-style CNN	0.50	100.0	2.4	3.6	95.0	100.0
		1.00	100.0	7.5	16.3	84.7	96.3
		1.50	100.0	26.1	54.2	73.0	90.7

where $c = 2^{\frac{p-1}{p}}$ is the constant depending only on the norm $\|\cdot\|_p$, which is 2 for ℓ_∞ (and $\sqrt{2}$ for ℓ_2).

The proof for Lemma IV.1 is as follows.

robust against perturbations $\|\delta\|_p < \frac{1}{cL} \text{margin}(f(\mathbf{x}))$, i.e.

$$\argmax f(\mathbf{x}) = \argmax f(\mathbf{x} + \delta), \forall \|\delta\|_p < \frac{1}{cL} \text{margin}(f(\mathbf{x})), \quad (10)$$

Proof. Let y, j be the class of $\mathbf{x}$ and $\hat{\mathbf{x}}$ predicted by f , Then

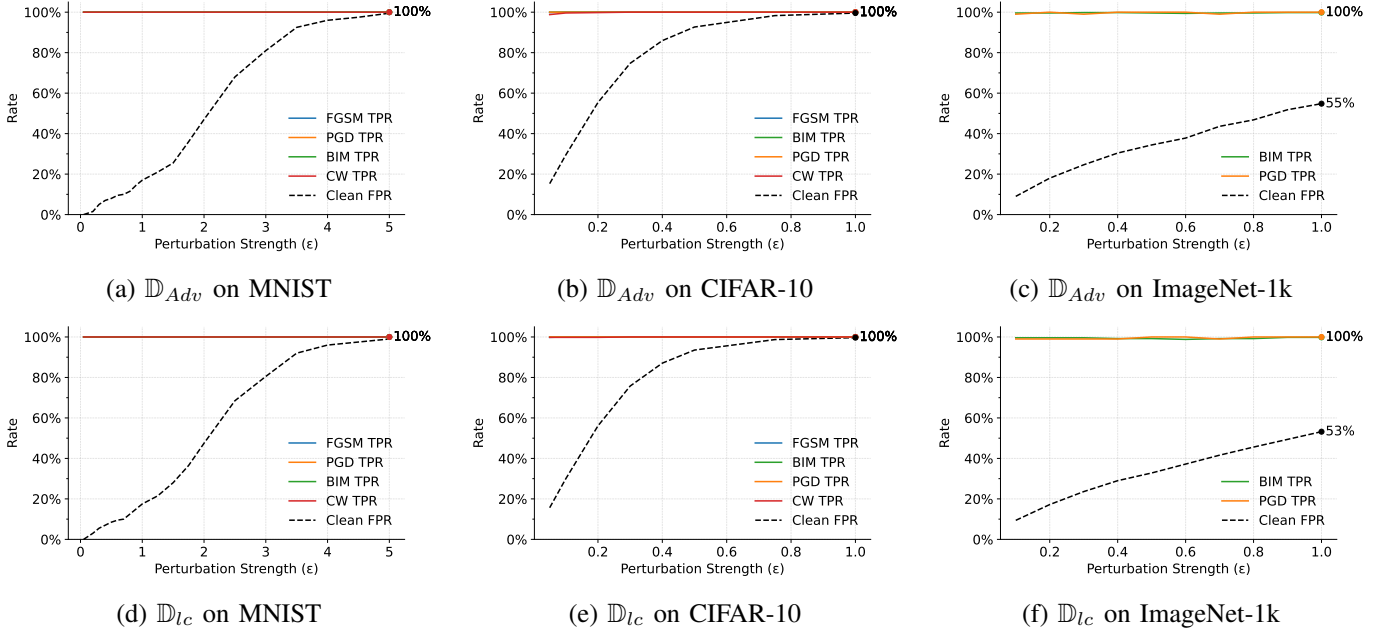


Fig. 2: Summary curves: TPR on adversarial inputs and FPR on clean inputs across ℓ_2 perturbation magnitudes ϵ , for $\mathbb{D}_{Adv}$ (top row) and $\mathbb{D}_{lc}$ (bottom row). Columns sweep MNIST (LeNet-5), CIFAR-10 (VGG16), and ImageNet-1k (ResNet-50).

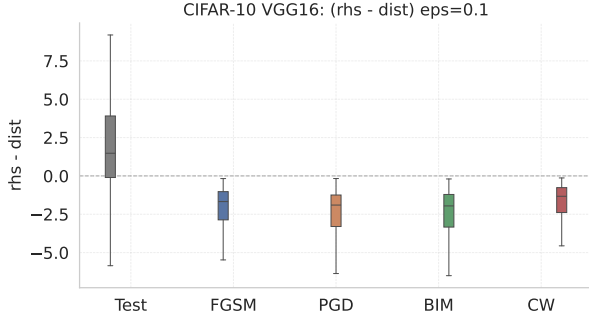


Fig. 3: Distribution of $(RHS - d_z)$ values under different attacks. Negative values ($RHS < d_z$) indicate samples flagged as adversarial.

TABLE XV: Cost-tightness sweep for the stronger estimator. CIFAR-10 VGG16, $\mathbb{D}_{Adv}$, $\epsilon=0.1$.

Estimator	cert%	ms/sample
Baseline	76.4	22
$K=1, T=3$	91.6	108
$K=1, T=6$	93.5	201
$K=1, T=9$	93.7	293
$K=1, T=12$	93.7	383
$K=2, T=12$	95.4	563

we have

$$\|f(\hat{x}) - f(x)\|_p = (|f_y(\hat{x}) - f_y(x)|^p + \sum_{i \neq y} |f_i(\hat{x}) - f_i(x)|^p)^{\frac{1}{p}} \text{ and } f_j(x) \leq f_y(x). \text{ Then,}$$

$$\geq (|f_y(\hat{x}) - f_y(x)|^p + |f_j(\hat{x}) - f_j(x)|^p)^{\frac{1}{p}}. \quad (11)$$

TABLE XVI: Empirical distribution of L'/L^* . CIFAR-10 VGG16, BIM, $n=500$. Violations are samples with $L'/L^* < 1$; the last column reports their ratio distribution. $\mathbb{D}_{lc}$ has a right-skewed tail at $\epsilon=0.10, 0.50$ from a few samples with tiny L^* ; its median remains near 1.

$\mathbb{D}_{Adv}$						
ϵ	cert%	L'/L^* mean $\pm$ std	Median	Viol. %	Viol. mean $\pm$ std	
0.05	62.6	1.037 ± 0.20	1.007	37.4	0.949 ± 0.082	
0.10	76.6	1.067 ± 0.24	1.034	23.4	0.908 ± 0.107	
0.30	83.0	1.153 ± 0.29	1.118	17.0	0.855 ± 0.133	
0.50	82.4	1.194 ± 0.30	1.162	17.6	0.861 ± 0.130	

$\mathbb{D}_{lc}$						
ϵ	cert%	L'/L^* mean $\pm$ std	Median	Viol. %	Viol. mean $\pm$ std	
0.05	62.0	0.985 ± 0.13	1.001	38.0	0.947 ± 0.200	
0.10	75.2	1.123 ± 2.63	1.006	24.8	0.952 ± 0.132	
0.30	80.4	1.064 ± 0.14	1.039	19.6	0.939 ± 0.065	
0.50	78.8	1.167 ± 1.40	1.060	21.2	0.924 ± 0.081	

Let $dz_y = f_y(\hat{x}) - f_y(x)$, $dz_j = f_j(\hat{x}) - f_j(x)$. Considering a polynomial $g(r) = r^p + (dz_j - dz_y - r)^p$, then

$$\begin{aligned} g(-dz_y) &= (-dz_y)^p + dz_j^p \\ &= (-f_y(\hat{x}) + f_y(x))^p + (f_j(\hat{x}) - f_j(x))^p \\ &\leq \|f(\hat{x}) - f(x)\|_p^p, \end{aligned}$$

where the last line inequality is caused by (11). By analyzing the derivatives of $g(z)$, we can find that $g(z)$ has a global minimum at $r = \frac{dz_j - dz_y}{2}$, using the fact that $f_j(\hat{x}) > f_y(\hat{x})$

$$\frac{(dz_j - dz_y)^p}{2^{p-1}} = g\left(\frac{dz_j - dz_y}{2}\right) \leq g(-dz_y) \leq \|f(\hat{x}) - f(x)\|_p^p, \quad (12)$$

TABLE XVII: Full ImageNet-1k results (PyTorch pretrained ResNet-50) at step $\epsilon = 0.1$. TPR on ℓ_2 BIM attacks; FPR on clean samples.

ϵ	$\mathbb{D}_{Adv}$ (%)		$\mathbb{D}_{lc}$ (%)		RS (Cohen) (%)	
	TPR $\uparrow$	FPR $\downarrow$	TPR $\uparrow$	FPR $\downarrow$	TPR $\uparrow$	FPR $\downarrow$
0.1	99.5	9.0	99.7	9.4	34.0	17.0
0.2	99.6	18.0	99.6	17.2	53.0	30.0
0.3	99.8	24.6	99.6	23.7	62.5	36.0
0.4	99.9	30.3	99.2	29.0	82.0	68.0
0.5	99.6	34.4	99.3	32.8	85.5	51.0
0.6	99.4	37.8	98.8	37.2	100.0	100.0
0.7	99.6	43.5	99.5	41.6	100.0	100.0
0.8	99.3	46.7	99.2	45.3	100.0	100.0
0.9	99.7	51.9	99.8	49.4	100.0	100.0
1.0	99.8	54.8	99.8	53.5	100.0	100.0

which implies

$$f_j(\hat{\mathbf{x}}) - f_y(\hat{\mathbf{x}}) + f_y(\mathbf{x}) - f_j(\mathbf{x}) \leq \frac{2}{2^{\frac{1}{p}}} \|f(\hat{\mathbf{x}}) - f(\mathbf{x})\| \quad (13)$$

$$\leq cL \|\hat{\mathbf{x}} - \mathbf{x}\|,$$

where the last inequality follows from the L-Lipschitz continuity of f . Moreover, since $f_j(\hat{\mathbf{x}}) - f_y(\hat{\mathbf{x}}) + f_y(\mathbf{x}) - f_j(\mathbf{x}) \geq q(\hat{\mathbf{x}}) + q(\mathbf{x})$, we have

$$q(\hat{\mathbf{x}}) \leq cL \|\hat{\mathbf{x}} - \mathbf{x}\| - q(\mathbf{x}).$$

□

The proof for Theorem IV.2 follows from a *single* application of Lemma IV.1 on the pair $(\mathbf{x}, \hat{\mathbf{x}})$ together with the triangle inequality, requiring only $L \geq L_{(\hat{\mathbf{x}}, \mathbf{x})}^*$.

Proof. By construction the local centroid $\tilde{\mathbf{x}}$ retains the predicted class of $\hat{\mathbf{x}}$, i.e., $y(\tilde{\mathbf{x}}) = y(\hat{\mathbf{x}})$. Combined with the assumption that $\hat{\mathbf{x}}$ is a successful adversarial example ($y(\hat{\mathbf{x}}) \neq y(\mathbf{x})$), the pair $(\mathbf{x}, \hat{\mathbf{x}})$ has *distinct* predicted classes. Lemma IV.1 applied to this pair therefore yields

$$q(\tilde{\mathbf{x}}) + q(\mathbf{x}) \leq cL \|\tilde{\mathbf{x}} - \mathbf{x}\|. \quad (14)$$

By the triangle inequality and $\|\hat{\mathbf{x}} - \mathbf{x}\| = \|\delta\|$,

$$\|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\| + \|\hat{\mathbf{x}} - \mathbf{x}\| = \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\| + \|\delta\|. \quad (15)$$

Combining (14) and (15) and dividing both sides by $cL > 0$,

$$\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\| \geq \frac{q(\tilde{\mathbf{x}}) + q(\mathbf{x})}{cL} - \|\delta\|,$$

which is the bound of Theorem IV.2. The inequality $L \geq L_{(\hat{\mathbf{x}}, \mathbf{x})}^*$ is exactly the saturation condition for (14) to hold, and is the only Lipschitz requirement invoked. □

F. Proof of Proposition IV.7

Proof. For an adversarial input $z = \hat{\mathbf{x}}$, the pseudo-label is $\hat{y} = y(\hat{\mathbf{x}}) \neq y(\mathbf{x})$, and $q(\mathbf{x}) > 0$ at $\mathbf{x}$'s own label.

(i) $\mathbb{D}_{Adv}$. Lemma IV.1 on the $(\mathbf{x}, \hat{\mathbf{x}})$ pair at $L = L_{(\mathbf{x}, \hat{\mathbf{x}})}^*$ gives $q(\hat{\mathbf{x}}) + q(\mathbf{x}) \leq cL_{(\mathbf{x}, \hat{\mathbf{x}})}^* \|\delta\|$. Combining with $L_{(\mathbf{x}, \hat{\mathbf{x}})}^* \leq L'$

and the local-Lipschitz formula $cL' \cdot \max(\|\hat{\mathbf{x}} - z'\|, \|\delta\|) = q(\hat{\mathbf{x}}) - q(z'; \hat{y})$,

$$q(\hat{\mathbf{x}}) + q(\mathbf{x}) \leq cL' \|\delta\| \leq cL' \cdot \max(\|\hat{\mathbf{x}} - z'\|, \|\delta\|) = q(\hat{\mathbf{x}}) - q(z'; \hat{y}),$$

so $q(z'; \hat{y}) \leq -q(\mathbf{x}) \leq 0$ — the flip property is a *consequence* of $L' \geq L_{(\mathbf{x}, \hat{\mathbf{x}})}^*$, not a hypothesis. Substituting back,

$$q(\hat{\mathbf{x}}) + q(z'; \hat{y}) \leq q(\hat{\mathbf{x}}) - q(\mathbf{x}) \leq q(\hat{\mathbf{x}}) + q(\mathbf{x}) \leq cL' \|\delta\|,$$

which is the Claim 1 bound.

(ii) $\mathbb{D}_{lc}$. By Proposition IV.6, $L = L_{(\hat{\mathbf{x}}, \mathbf{x})}^*$ certifies Theorem IV.2: $\|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\| \geq \frac{q(\tilde{\mathbf{x}}) + q(\mathbf{x})}{cL_{(\hat{\mathbf{x}}, \mathbf{x})}^*} - \|\delta\|$. The detector's right-hand side is $\frac{q(\tilde{\mathbf{x}}) + q(z'; \hat{y})}{cL'} - \|\delta\|$. Combining the flip property $q(z'; \hat{y}) \leq 0 \leq q(\mathbf{x})$ with $L_{(\hat{\mathbf{x}}, \mathbf{x})}^* \leq L'$,

$$\frac{q(\tilde{\mathbf{x}}) + q(z'; \hat{y})}{cL'} \leq \frac{q(\tilde{\mathbf{x}})}{cL'} \leq \frac{q(\tilde{\mathbf{x}})}{cL_{(\hat{\mathbf{x}}, \mathbf{x})}^*} \leq \frac{q(\tilde{\mathbf{x}}) + q(\mathbf{x})}{cL_{(\hat{\mathbf{x}}, \mathbf{x})}^*},$$

so the detector's right-hand side is at most the certified one, and the detector flags $\hat{\mathbf{x}}$ whenever Theorem IV.2 does. □

G. Margin Monotonicity Under Gradient Ascent

We give a formal justification for the class-preservation property used by Theorem IV.2: under gradient ascent on the margin function at a fixed reference label, the margin is monotonically non-decreasing along the flow, hence the predicted class is preserved.

Definition A.2 (Chain rule). A locally Lipschitz function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ admits a chain rule if for any absolutely continuous $u : \mathbb{R}_+ \rightarrow \mathbb{R}^d$,

$$(h \circ u)'(t) = \langle \partial^\circ h(u(t)), u'(t) \rangle \quad \text{for a.e. } t \geq 0.$$

Lemma A.3 (Davis et al [59], Lemma 5.2). Let $h : \mathbb{R}^d \rightarrow \mathbb{R}$ be locally Lipschitz and admit a chain rule, and let $u : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ be absolutely continuous with $du/dt \in \partial^\circ h(u(t))$. Then $\|du/dt\|^2 = dh(u(t))/dt$ for a.e. $t \geq 0$.

Lemma A.4 (Margin monotonicity). Let f be a locally Lipschitz network and $q(\cdot; \hat{y})$ the margin function at fixed reference label $\hat{y} = \arg \max f(z)$ for input z , and assume $q(\cdot; \hat{y})$ admits a chain rule (Definition A.2); this holds whenever f is locally Lipschitz and tame, including piecewise-linear ReLU/max-pool networks [59]. Under the gradient ascent flow $du/dt \in \partial^\circ q(u(t); \hat{y})$ initialized at $u(0) = z$,

$$\frac{dq(u(t); \hat{y})}{dt} = \left\| \frac{du}{dt} \right\|^2 \geq 0 \quad \text{for a.e. } t \geq 0,$$

with equality iff $du/dt = 0$. In particular, $q(u(t); \hat{y})$ is non-decreasing along the flow; if $q(z; \hat{y}) > 0$, then $q(u(t); \hat{y}) > 0$ for all $t \geq 0$, so $\arg \max f(u(t)) = \hat{y}$ throughout.

Proof. The equality is Lemma A.3 applied to $h = q(\cdot; \hat{y})$. Non-negativity gives monotonicity. The class-preservation conclusion is $q(u; \hat{y}) > 0 \iff \hat{y} = \arg \max f(u)$, combined with the trajectory's non-decreasing margin keeping the strict positivity from initialization onward. □

H. Proof of Lemma IV.4

Proof. Let $a = \hat{\mathbf{x}} = \mathbf{z}$ and $b = \mathbf{x}$ denote the adversarial and its (unknown) clean counterparts, respectively, and let $\mathbf{z}' = \mathbf{z} + \sigma$ be the perturbed sample used in our approximation. Define the pairwise distances

$$\alpha = \|a - b\|, \quad \beta = \|a - \mathbf{z}'\|, \quad \gamma = \|b - \mathbf{z}'\|.$$

We write the *ideal* constant from Lemma IV.1 as

$$L^* \triangleq \frac{q(a) + q(b)}{c\alpha},$$

and the *computable* local quotient as

$$L' \triangleq \frac{|q(a) - q(\mathbf{z}')|}{c\beta}.$$

Let L_{glob} be any global Lipschitz constant of $q(\cdot)$, i.e.,

$$|q(u) - q(v)| \leq c L_{\text{glob}} \|u - v\| \quad \forall u, v.$$

Using the triangle inequality, we decompose:

$$|q(a) - q(\mathbf{z}')| \leq |q(a) - q(b)| + |q(b) - q(\mathbf{z}')|.$$

Since $|q(a) - q(b)| \leq q(a) + q(b) = c\alpha L^*$ and $|q(b) - q(\mathbf{z}')| \leq c L_{\text{glob}} \gamma$, we have

$$L' \leq \frac{c\alpha L^* + c L_{\text{glob}} \gamma}{c\beta} = \frac{\alpha}{\beta} L^* + \frac{\gamma}{\beta} L_{\text{glob}}.$$

Similarly, by reversing the inequality,

$$|q(a) - q(\mathbf{z}')| \geq |q(a) - q(b)| - |q(b) - q(\mathbf{z}')| \geq c\alpha D - c L_{\text{glob}} \gamma,$$

where $D \triangleq \frac{|q(a) - q(b)|}{c\alpha} \leq L^*$. Hence,

$$L' \geq \frac{\alpha}{\beta} D - \frac{\gamma}{\beta} L_{\text{glob}}.$$

Combining both sides yields the two-sided bound:

$$\frac{\alpha}{\beta} D - \frac{\gamma}{\beta} L_{\text{glob}} \leq L' \leq \frac{\alpha}{\beta} L^* + \frac{\gamma}{\beta} L_{\text{glob}}.$$

Consequently, the deviation of L' from the scaled ideal constant $(\alpha/\beta)L^*$ is bounded by

$$|L' - (\alpha/\beta)L^*| \leq \frac{\gamma}{\beta} L_{\text{glob}} + \frac{\alpha}{\beta} (L^* - D).$$

This reveals two distinct sources of bias:

- A *geometric bias* $\frac{\gamma}{\beta} L_{\text{glob}}$, governed by the relative placement of $\mathbf{z}'$ with respect to a and b ;
- A *margin bias* $\frac{\alpha}{\beta} (L^* - D)$, reflecting the gap between the sum- and difference-based margin quotients.

In particular, when $\mathbf{z}'$ lies approximately along the segment (a, b) , we have $\beta \approx \alpha$ and $\gamma \approx |\alpha - \beta|$, so both terms become small and L' closely tracks L^* . Conversely, when $\mathbf{z}'$ is very close to a (small β) or far from b (large γ), the approximation tends to overestimate the ideal constant. $\square$

I. Proof of Theorem IV.5

The proof of Theorem IV.5 is as follows.

Proof. We prove the theorem by bounding the difference term by term. We use the assumptions of Theorem IV.5: $L_q \leq c\bar{L}$ (A1) and $L' \leq L_{\text{glob}} \leq \bar{L}$ (A2); recall $\eta := 1 - L'/\bar{L} \in [0, 1)$.

Let

$$D = \left| \frac{q(\tilde{z}) + q(\mathbf{x})}{c L_{\text{glob}}} - \frac{q(\tilde{z}) + q(\mathbf{z}'; \hat{y})}{c L'} \right|.$$

By adding and subtracting $\frac{q(\tilde{z}) + q(\mathbf{z}'; \hat{y})}{c L_{\text{glob}}}$,

$$D \leq \frac{1}{c} \left(\frac{|q(\mathbf{x}) - q(\mathbf{z}'; \hat{y})|}{L_{\text{glob}}} + |q(\tilde{z}) + q(\mathbf{z}'; \hat{y})| \cdot \left| \frac{1}{L_{\text{glob}}} - \frac{1}{L'} \right| \right).$$

By Lipschitz continuity of q (A1) and the fact that $\|\mathbf{x} - \mathbf{z}'\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{z}'\| \leq 2\|\delta\|$ (using $\|\mathbf{z} - \mathbf{x}\| \leq \|\delta\|$ from the threat model and $\|\mathbf{z} - \mathbf{z}'\| \leq \|\delta\|$ from Algorithm 1), we have $|q(\mathbf{x}) - q(\mathbf{z}'; \hat{y})| \leq L_q \|\mathbf{x} - \mathbf{z}'\| \leq 2L_q \|\delta\|$. Combined with $L' \leq L_{\text{glob}}$,

$$\frac{|q(\mathbf{x}) - q(\mathbf{z}'; \hat{y})|}{L_{\text{glob}}} \leq \frac{2L_q \|\delta\|}{L_{\text{glob}}} \leq \frac{2L_q \|\delta\|}{L'}.$$

For the second term, $L' \leq L_{\text{glob}} \leq \bar{L}$ gives $1 - L'/L_{\text{glob}} \geq 0$ and $1 - L'/L_{\text{glob}} \leq 1 - L'/\bar{L} = \eta$. Therefore

$$\left| \frac{1}{L_{\text{glob}}} - \frac{1}{L'} \right| = \frac{1}{L'} \cdot \frac{L_{\text{glob}} - L'}{L_{\text{glob}}} \leq \frac{\eta}{L'}.$$

Putting the two bounds together,

$$\Delta_T \leq \frac{1}{c} \left[\frac{2L_q}{L'} \|\delta\| + \frac{\eta}{L'} (q(\tilde{z}) + |q(\mathbf{z}'; \hat{y})| + L_q \|\delta\|) \right],$$

which is the bound (9) (with the geometric stretch factor doubled to absorb the $\mathbf{x} \rightarrow \mathbf{z} \rightarrow \mathbf{z}'$ triangle). $\square$