When Uncertainty Is the Target: Adversarial Attacks on Uncertainty-Aware Predictors

A Q M Sazzad Sayyed[1,*] **Michele Caprio**[2] **Nathaniel D. Bastian**[3] **Francesco Restuccia**[1]

[1]Northeastern University [2]University of Warwick [3]Syracuse University

Abstract

As machine learning becomes increasingly deployed in high-stakes scenarios, studying adversarial vulnerability of uncertainty-aware models has become a compelling necessity. In this paper, we focus on adversarial attacks that manipulate *predictive uncertainty* rather than model predictions. We focus on predictors that decompose uncertainty into aleatoric uncertainty (AU) and epistemic uncertainty (EU) and formalize *uncertainty vulnerability* as the maximum increase in AU achievable under a stealth constraint where the EU is kept below a threshold. This framing yields three complementary findings. First, we theoretically show that knowledge of the baseline EU can enlarge the attacker's feasible stealth region and increase the maximal achievable increase in AU. Second, a first-order analysis reveals that vulnerability is governed by the component of the AU gradient orthogonal to EU, enabling stealthy manipulation. Third, contrary to standard margin-based intuition, vulnerability peaks at moderate margins. Experiments across MNIST, CIFAR-10, and CIFAR-100 with standard architectures show that the proposed attack achieves non-trivial aleatoric damage while maintaining zero epistemic-stealth violation (e.g., mean AU increase of 0.0286, 0.0853, and 0.1568, respectively), whereas unconstrained uncertainty attacks violate the stealth constraint on over 90% of samples. Experiments indicate that the largest damage typically occurs in an intermediate uncertainty regime rather than at the extremes. Furthermore, consistent with our theoretical analysis, feasible attackability collapses under increasing epistemic uncertainty (e.g., mean attackability decreases from 0.0363 to 0). Overall, these results intimately connect uncertainty decomposition to adversarial geometry. The code can be found here: Code

1 Introduction

Predictive uncertainty is increasingly used in high-stakes systems to properly inform decision-making. Real-world failures such as biased recognition systems [Guynn, 2015, AlgorithmWatch, 2020, Medina, 2023] and autonomous driving incidents [National Highway Traffic Safety Administration, 2017] show that unreliable confidence or uncertainty estimates can be as dangerous as incorrect predictions themselves. This has motivated a large body of work on uncertainty quantification (UQ) for deep neural networks (DNNs), including Bayesian [Kendall and Gal, 2017, Depeweg et al., 2018], ensemble-based [Lakshminarayanan et al., 2017], imprecise probability [Caprio et al., 2024, 2025, Amini et al., 2020] and distributional [Malinin and Gales, 2018] approaches.

Predictive uncertainty is distinguished into aleatoric uncertainty (AU) and epistemic uncertainty (EU). AU captures ambiguity intrinsic to the input (thus irreducible), such as class overlap or label noise, whereas EU reflects model uncertainty arising from insufficient knowledge or weak support in the training distribution – and it is thus reducible [Kendall and Gal, 2017, Depeweg et al., 2018]. While this decomposition has become central to modern uncertainty quantification, *we argue that*

*Corresponding author: sayyed.a@northeastern.edu

an adversary could only need to manipulate the uncertainty associated with the prediction to inflict damage, instead of attempting to change the predicted labels.

While adversarial machine learning has extensively been studied attacks on model predictions (e.g., evasion attacks), *attacks to model uncertainty are largely underexplored*. To our knowledge, no prior work studies *attacks that increase AU while remaining stealthy with respect to EU*. While existing work on calibration shows that predictive confidence can be made unreliable under adversarial perturbations [Guo et al., 2017, Kull et al., 2019], they largely treat uncertainty as a single scalar quantity (e.g., confidence or total predictive entropy). In contrast, a more fundamental setting – attacking AU while keeping the EU unchanged – remains open.

Figure 1 illustrates the difference between classic evasion attacks and uncertainty-targeted attacks. In the former, the adversary attempts to make the model confidently wrong: for example, a cat image is perturbed so that the predictor assigns high probability to "dog." *Conversely, an attack against an uncertainty-aware predictor can be subtler*. The adversary may avoid forcing a confident wrong answer and make the prediction appear genuinely ambiguous, e.g., by spreading probability mass between "cat" and "dog," while keeping the system's warning signal low. As such, the downstream system may then treat the input as an intrinsically difficult case instead of as a manipulation, since from the defender's perspective, increased AU is indistinguishable from genuine data ambiguity.

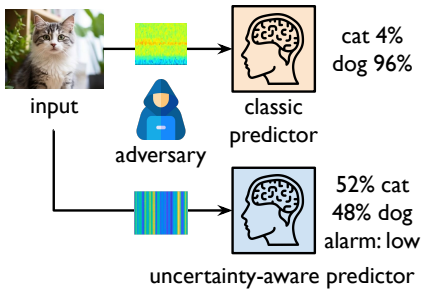


Figure 1: Classical evasion attacks aim to make a predictor confidently wrong; uncertainty-targeted attacks make the model appear legitimately unsure, while the warning signal remains low.

In this work, we formalize and analyze adversarial attacks that maximize AU while keeping EU within a prescribed stealth budget. What makes this problem challenging is that the attacker must increase AU without exceeding a given perturbation budget for the EU. These two quantities are coupled; as such, perturbations that are typically harmful for prediction might not always be successful. This leads to the following intriguing question: *what helps the attacker successfully increase AU without increasing EU beyond a certain level?* One of our key findings is that this form of uncertainty manipulation is not uniformly achievable across samples. Conversely, vulnerability tends to peak at neither very low nor very high baseline uncertainty but at intermediate uncertainty regime. We hypothesize that these samples lie in transition regions of the classifier (i.e., where class ambiguity is appreciable but epistemic disagreement remains sufficiently constrained) and that this region is closely related to traditional notions of adversarial vulnerability such as small margin and high local sensitivity. *This perspective leads to a new view of adversarial robustness: by characterizing which samples are most vulnerable to uncertainty manipulation, we connect uncertainty decomposition with the classifier's geometry and identify the conditions under which uncertainty-aware predictors are most vulnerable.*

Summary of Novel Contributions

• We formulate sample-wise uncertainty vulnerability as a constrained optimization problem – maximize the increase in AU while remaining stealthy in EU. This isolates a class of attacks that alters the reliability signal without necessarily forcing immediate label change.

• We show that the attacker's feasible damage budget depends on how baseline EU is monitored or disclosed. In particular, we derive conditions under which knowledge of the baseline epistemic level enlarges the attacker's opportunity set, while imperfect knowledge shrinks it. This is corroborated empirically by the collapse of maximum feasible damage (e.g., mean Γ decreases from 0.0363 to 0 as epistemic uncertainty radius increases). We derive a first-order characterization of local vulnerability, and show that vulnerability is amplified by aleatoric changes that are orthogonal to epistemic changes and is often largest at moderate margins.

• We validate these predictions empirically across MNIST, CIFAR-10, and CIFAR-100 using multiple ensemble architectures. Across settings, the maximal feasible aleatoric damage is typically non-monotone in baseline uncertainty and is concentrated in intermediate uncertainty regimes The constrained attack achieves significant AU damage (e.g., mean ΔAU of 0.0286 on MNIST, 0.0853 on CIFAR-10, and 0.1568 on CIFAR-100), while maintaining zero EU violations, as opposed to unconstrained attacks that violate EU constraints 90% of the time.

2 Related Work

Adversarial vulnerability, margin, and local sensitivity. Literature has shown that samples with smaller effective margin or that lie closer to a decision boundary are generally more susceptible to adversarial perturbations. This view has motivated margin-based robustness methods such as Lipschitz-margin training [Tsuzuku et al., 2018] and adversarial margin maximization [Yan et al., 2019]. More broadly, it is consistent with the intuition that a larger perturbation is required to alter the prediction of a sample that lies deeper inside a class region. Related geometric analyses directly study surrounding decision boundaries and show that boundary distances and adjacent-class structure around a sample help characterize adversarial susceptibility [He et al., 2018]. A complementary line of work relates adversarial vulnerability to local sensitivity, often quantified through input gradients or local Lipschitz behavior. [Goodfellow et al., 2015] argued that adversarial examples arise in part from the locally linear behavior of modern networks, while [Simon-Gabriel et al., 2019] showed that first-order adversarial vulnerability grows with gradients of the training objective. Similarly, [Finlay and Oberman, 2021] derived per-sample robustness bounds based on local gradient information and proposed scalable input-gradient regularization as a robustness mechanism. More recent training methods explicitly reweight samples by geometric vulnerability, emphasizing points closer to the boundary during adversarial training [Zhang et al., 2021]. Our work instead studies where the uncertainty decomposition is easiest to manipulate, which is a fundamentally different problem.

Attacks on confidence, calibration, and uncertainty. Existing work has shown that neural networks are often miscalibrated [Guo et al., 2017, Kull et al., 2019] and that calibration and predictive uncertainty can substantially degrade under adversarial perturbations. These findings imply that uncertainty-based decision systems may fail due to corruption of their reliability. As such, a growing literature now studies direct attacks on uncertainty estimates. Early work showed that confidence scores can be manipulated without necessarily changing the predicted labels [Galil and El-Yaniv, 2021, Tuna et al., 2023]. Other work attacked EU estimates derived from Bayesian approximations or MC-dropout models, showing that uncertainty itself can be exploited as an attack objective [Tuna et al., 2021]. More recently, adversarial perturbations have been shown to severely distort system-level calibration metrics such as ECE and Brier score while leaving accuracy nearly unchanged [Emde et al., 2025]. Complementarily, robustly trained classifiers have been shown to exhibit smaller "uncertainty spans" under perturbation, indicating that adversarial robustness can partially transfer to uncertainty robustness [Ledda et al., 2026]. Instead, our work asks how an attacker can manipulate an AU signal while remaining stealthy with respect to EU monitoring. The closes work related to ours is [Rabanser et al., 2025] which studies a complementary but distinct threat: their Mirage attack is carried out by a malicious model provider who fine-tunes the deployed model at training time (using a KL-divergence penalty) to suppress confidence in targeted input regions while preserving accuracy, thereby targeting abstention mechanisms. Our setting separates itself along three axes: (i) the adversary is an external test-time attacker who perturbs inputs instead of model weights; (ii) the attack targets the aleatoric component of a decomposed AU/EU uncertainty estimate instead of a scalar confidence score; and (iii) stealth is defined by keeping EU within a prescribed band instead of preserving accuracy. Both papers show that uncertainty-aware systems are vulnerable from both the service provider side and the client side (input perturbation).

Summary of Novelty: This work studies a more selective form of uncertainty attack than prior attacks, as we ask whether an adversary can increase AU while keeping the monitored EU nearly unchanged. This leads to a constrained, sample-wise vulnerability measure: the maximum apparent ambiguity that can be induced without triggering an epistemic warning. The resulting analysis connects uncertainty manipulation to local gradient geometry and classifier margin, showing that vulnerability is controlled not only by how easily AU can be increased, but by whether it can be increased in directions that leave EU stable. This explains why attackability often peaks in an intermediate uncertainty regime instead of at the extremes.

3 Problem Setup

Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be our input and output metric spaces, assumed to be *separable*. We use a predictive credal approach [Caprio et al., 2024] so that, given a training set $D = \{(x_i, y_i)\}_{i=1}^{n}$ and a new test input $x_{n+1} \equiv x$, we obtain a predictive set $\mathcal{P}_x$ on the y's, i.e. $\mathcal{P}_x \subseteq \Delta_{\mathcal{Y}}$ convex and weak*-closed, where $\Delta_{\mathcal{Y}}$ is the space of (countably additive) probability measures on $\mathcal{Y}$.

Assumption (Attacker Perfect Knowledge). The attacker knows $\mathcal{P}_x$ and seeks a perturbation $\tilde{x}$ solving

$$\begin{aligned} \inf_{\tilde{x}\in\mathcal{X}} \quad & d_{\mathcal{X}}(\tilde{x}, x) \\ \text{s.t.} \quad & d_H(\mathcal{P}_x, \mathcal{P}_{\tilde{x}}) \leq \delta \\ & AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \geq \gamma, \end{aligned} \tag{1}$$

where d_H is the Hausdorff distance on $\Delta_{\mathcal{Y}}$ and $AU(\cdot)$ denotes AU (e.g., lower entropy or lower prevision). Problem (1) enforces stealth via proximity of the full predictive object. In practice, defenders often monitor only a scalar $EU(\mathcal{P}_x)$, leading to the relaxed formulation

$$\begin{aligned} \inf_{\tilde{x}\in\mathcal{X}} \quad & d_{\mathcal{X}}(\tilde{x}, x) \\ \text{s.t.} \quad & |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta \\ & AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \geq \gamma, \end{aligned} \tag{2}$$

where $EU(\cdot)$ denotes EU (e.g., entropy gap or upper–lower prevision difference). This constraint is strictly weaker than (1), since similar EU values do not imply proximity in d_H.

The defender's notion of stealth depends on the monitored predictive object: **Full monitoring:** $d_H(\mathcal{P}_x, \mathcal{P}_{\tilde{x}}) \leq \delta$. **Scalar EU monitoring:** $|EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta$. **Uncertain baseline EU:** $EU(\mathcal{P}_x)$ is not exactly known to the attacker and lies within an uncertainty set. We focus on scalar EU monitoring (second regime) as this is more practical for deployed systems, with the third treated as a special case capturing partial epistemic disclosure. Throughout this work, we view uncertainty manipulation through three equivalent lenses. The primary object is the predictive credal set $\mathcal{P}_x$. Associated uncertainty functionals AU($\cdot$) and EU($\cdot$) give a scalar view of the predictive credal sets. For local analysis, we further write $a(x) := AU(\mathcal{P}_x)$ and $e(x) := EU(\mathcal{P}_x)$ making the connection to input space explicit. The set-level, scalar, and local view will be used interchangeably.

4 Baseline Epistemic Disclosure and Vulnerability

We focus on Problem (2), as it is more realistic that the defender does not (or cannot) monitor the full predictive object $\mathcal{P}_x$ and instead watches one or a few scalar diagnostics. We define the global and local maximum feasible damage as the following.

Definition 4.1 (Global Maximal Feasible Damage Budget)**.** Fix $\delta > 0$ and let $\mathfrak{P} := \{\mathcal{P}_x : x \in \mathcal{X}\}$ be the family of predictive credal sets. Let $EU, AU : \mathfrak{P} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ be functionals such that $EU(\mathcal{P}) \geq 0$ for all $\mathcal{P} \in \mathfrak{P}$. For $x \in \mathcal{X}$, define

$$\Gamma(x, \delta) := \sup \Big\{ \gamma \in \mathbb{R} : \exists \tilde{x} \in \mathcal{X} \text{ s.t. } |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta \text{ and } AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \geq \gamma \Big\}.$$

Equivalently,

$$\Gamma(x, \delta) = \sup_{\tilde{x}\in\mathcal{X}: |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \Big] \in \overline{\mathbb{R}}.$$

Definition 4.1 formalizes sample-wise vulnerability or attackability under epistemic-stealth monitoring with $\Gamma(x, \delta)$. This denotes the largest increase in AU achievable for a sample while keeping the EU within a δ-band around its baseline value. In practice, it is not feasible to estimate the global maximum feasible damage budget as there would be no bound on the added perturbation. In practice, we estimate a local version of it defined as follows:

Definition 4.2 (Local Maximal Feasible Damage Budget)**.** Fix $\epsilon > 0$ and $\delta > 0$, and let $\mathfrak{P} := \{\mathcal{P}_x : x \in \mathcal{X}\}$ be the family of predictive credal sets. Let $EU, AU : \mathfrak{P} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ be functionals such that $EU(\mathcal{P}) \geq 0$ for all $\mathcal{P} \in \mathfrak{P}$. For $x \in \mathcal{X}$, define

$$\Gamma_\epsilon(x, \delta) := \sup_{\tilde{x}\in\mathcal{X}:\ \|\tilde{x}-x\|_\infty \leq \epsilon,\ |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \Big].$$

Thus $\Gamma_\epsilon(x, \delta)$ measures the largest achievable increase in AU inside an ℓ_∞ perturbation ball of radius ϵ, while remaining stealthy with respect to EU up to tolerance δ. In practice, we estimate $\Gamma_\epsilon(x, \delta)$ via PGD-based constrained optimization.

In the subsequent sections, unless otherwise stated, any Γ or Γ_ϵ refer to $\Gamma_\epsilon(x, \delta)$ introduced in the definition 4.2. The following theorem gives an equivalence characterization for the relation between EU and AU. Specifically, it shows that $EU(\mathcal{P}_x) = 0$ yields a smaller maximum damage budget than $EU(\mathcal{P}_x) > 0$.

Theorem 4.3 (Envelope Characterization of Feasible Damage)**.** *Fix $\delta > 0$ and assume $EU, AU : \mathfrak{P} \to \overline{\mathbb{R}}$ with $EU(\mathcal{P}) \geq 0$ for all $\mathcal{P} \in \mathfrak{P}$. Let $x_0, x_1 \in \mathcal{X}$ satisfy $EU(\mathcal{P}_{x_0}) = 0$, and $EU(\mathcal{P}_{x_1}) =: e > 0$. We define the following two (nonempty) envelope values,*

$$A_0(\delta) := \sup_{\tilde{x} \in \mathcal{X} : EU(\mathcal{P}_{\tilde{x}}) \leq \delta} AU(\mathcal{P}_{\tilde{x}}), \qquad A_1(e, \delta) := \sup_{\tilde{x} \in \mathcal{X} : \max(0, e-\delta) \leq EU(\mathcal{P}_{\tilde{x}}) \leq e+\delta} AU(\mathcal{P}_{\tilde{x}}).$$

Then the following statements are equivalent:

(i) $\Gamma(x_0, \delta) < \Gamma(x_1, \delta)$.

(ii) The strict envelope inequality holds:

$$A_1(e, \delta) - A_0(\delta) > AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0}). \tag{3}$$

(iii) There exists $\hat{x} \in \mathcal{X}$ such that $|EU(\mathcal{P}_{\hat{x}}) - e| \leq \delta$ and

$$AU(\mathcal{P}_{\hat{x}}) - AU(\mathcal{P}_{x_1}) > \sup_{\tilde{x} \in \mathcal{X} : EU(\mathcal{P}_{\tilde{x}}) \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_{x_0}) \Big]. \tag{4}$$

In particular, Equation (3) *gives a necessary and sufficient condition for the maximum feasible damage budget (i.e., the largest γ for which the constraints in Equation* (2) *are feasible) to be strictly smaller when $EU(\mathcal{P}_x) = 0$ than when $EU(\mathcal{P}_x) = e > 0$.*

Key Finding: *Theorem 4.3 compares two potentially different samples, x_0 and x_1. It characterizes, rather than universally guarantees, when the positive-EU sample has the larger feasible damage budget. The deciding quantity is the opportunity gap $A_1(e, \delta) - A_0(\delta)$: the strict ordering $\Gamma(x_0, \delta) < \Gamma(x_1, \delta)$ holds if and only if this gap exceeds the difference in their baseline AU. This cross-sample characterization is distinct from the fixed-sample disclosure result in Appendix F.*

5 First-order condition for local attackability

While our analysis of the Problem 2 provides an equivalence characterization, it does not indicate when the conditions may hold. In this section, we perform first order analysis to get conditions when a sample can be successfully attacked and subsequently, Theorem 4.3 can hold.

Proposition 5.1 (First-Order vulnerability Condition)**.** *Let $a(x) := AU(\mathcal{P}_x)$ and $e(x) := EU(\mathcal{P}_x)$ be continuously differentiable in a neighborhood of x. Define $\Gamma_{\text{lin}}(x, \varepsilon) := \sup_{\|r\|_2 \leq \varepsilon,\ \nabla e(x)^\top r = 0} \nabla a(x)^\top r$. If $\nabla e(x) \neq 0$, then $\Gamma_{\text{lin}}(x, \varepsilon) = \varepsilon \left\| \Pi_{\nabla e(x)^\perp} \nabla a(x) \right\|_2$, where Π_m denotes orthogonal projection onto the subspace m. If $\nabla e(x) = 0$, then $\Gamma_{\text{lin}}(x, \varepsilon) = \varepsilon \|\nabla a(x)\|_2$.*

Proposition 5.1 identifies the local mechanism behind stealthy uncertainty attacks. The key quantity is not the full aleatoric gradient, but the component that remains after removing the direction aligned with epistemic change. When $\nabla AU(x)$ points mostly along $\nabla EU(x)$, increasing AU inevitably changes EU as well, making stealth difficult. When a large orthogonal component exists, the attacker can increase AU while remaining locally invisible to EU-based monitoring.

Proposition 5.2 (Moderate-Margin Sensitivity of a Binary Top-Two AU Surrogate)**.** *Let $m(x) = z_{(1)}(x) - z_{(2)}(x) > 0$ denote the top-two logit margin of a single classifier, and define the binary top-two uncertainty surrogate $a_{\text{bin}}(x) := h(\sigma(m(x)))$, where $\sigma(t) = \frac{1}{1+e^{-t}}$ and $h(p) = -p \log p - (1-p) \log(1-p)$. Suppose that m is differentiable in a neighborhood of x. Then $\nabla a_{\text{bin}}(x) = -\phi(m(x)) \nabla m(x)$, where $\phi(t) := t\, \sigma(t) \big(1 - \sigma(t)\big)$. Moreover, $\phi(0) = 0$ and $\lim_{t \to \infty} \phi(t) = 0$, and ϕ attains its unique maximum at the positive solution $t^\star \approx 1.5434$ of $t \tanh(t/2) = 1$. Consequently, $\|\nabla a_{\text{bin}}(x)\|_2 = \phi(m(x)) \|\nabla m(x)\|_2$. For a fixed margin-gradient norm, the local sensitivity of the binary surrogate is therefore largest at a moderate margin.*

If $e(x) := EU(\mathcal{P}_x)$ is also differentiable, then under exact first-order epistemic stealth, $\Gamma_{\text{lin}}^{\text{bin}}(x, \epsilon) = \epsilon\, \phi(m(x)) \left\| \Pi_{\nabla e(x)^\perp} \nabla m(x) \right\|_2$.

Relation to attacked credal AU. Proposition 5.2 is exact for the binary top-two surrogate a_{bin}; it does not assert that this surrogate is identical to the credal AU used by our attack. In the experiments, the attacked functional is $AU(\mathcal{P}_x) = \min_k H(p_k(\cdot \mid x))$. We can make the connection locally under additional assumptions. Suppose that the entropy-minimizing ensemble member $k^\star$ is unique and remains the minimizing member in a neighborhood of x. Then, within that neighborhood, $AU(\mathcal{P}_x) = H(p_{k^\star}(\cdot \mid x))$. Let $p_{(1)}$ and $p_{(2)}$ be the two largest probabilities of this active member and define $s := p_{(1)} + p_{(2)}$, $q := \frac{p_{(1)}}{p_{(1)}+p_{(2)}}$, and $m_{k^\star} := z^{k^\star}_{(1)} - z^{k^\star}_{(2)}$. For a softmax predictor, $q = \sigma(m_{k^\star})$. Let r denote the normalized distribution over the remaining classes. The entropy of the active member admits the exact decomposition $H(p_{k^\star}) = h(s) + s\,h(q) + (1-s)H(r)$. Thus, the binary top-two entropy $h(\sigma(m_{k^\star}))$ is one component of the full multiclass entropy, rather than the full experimental AU. Differentiating the decomposition gives $\nabla AU(\mathcal{P}_x) = -s\,\phi(m_{k^\star})\nabla m_{k^\star} + R(x)$, where $R(x) = [h'(s) + h(q) - H(r)]\,\nabla s + (1-s)\nabla H(r)$. After projection onto directions locally orthogonal to the EU gradient, $\Pi_{\nabla EU(x)^\perp}\nabla AU(\mathcal{P}_x) = -s\,\phi(m_{k^\star})\Pi_{\nabla EU(x)^\perp}\nabla m_{k^\star} + \Pi_{\nabla EU(x)^\perp}R(x)$. Consequently, when the active member is locally stable, the top-two mass is concentrated, and the projected residual term is small, $\Pi_{\nabla EU(x)^\perp}\nabla AU(\mathcal{P}_x) \approx -s\,\phi(m_{k^\star})\Pi_{\nabla EU(x)^\perp}\nabla m_{k^\star}$. Under these conditions, Proposition 5.2 identifies one mechanism contributing to the local geometry of the attacked credal AU. The approximation may weaken when substantial probability mass lies outside the top two classes, when the tail distribution varies rapidly, or when the entropy-minimizing ensemble member changes locally.

Key Finding: *Proposition 5.2 identifies a mechanism for a binary top-two uncertainty surrogate which clarifies why vulnerability is not expected to peak at either vanishing or very large margins. Its relevance to the attacked credal AU is local and conditional and the same factor contributes to the credal-AU gradient when the entropy-minimizing member is locally stable and the projected residual term is small. Moreover, realized first-order attackability depends on both the modulation factor $\phi(m)$ and the projected margin-gradient norm. Therefore, empirical attackability need not attain its maximum exactly at $m^\star \approx 1.5434$, the maximizer of ϕ alone.*

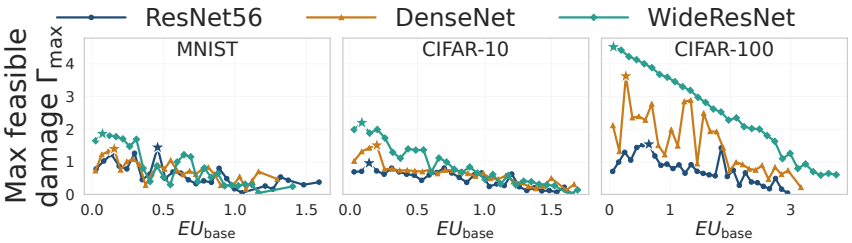


Figure 2: Upper envelope of feasible aleatoric damage across baseline uncertainty budgets. Columns correspond to architectures. The bins represent baseline EU EU_{base}. Across settings, the envelope bends and eventually declines rather than increasing monotonically, consistent with the theoretical prediction that the largest feasible damage occurs in an intermediate regime.

6 Experimental Setup and Results

In this section we empirically validate the theoretical predictions of Theorem 4.3, and Propositions 5.1-5.2. We provide more details of our experimental setup in Appendix C and I.

Datasets: MNIST [Deng, 2012], CIFAR-10, and CIFAR-100 [Krizhevsky and Hinton, 2009]. For these benchmarks, we train ensemble models using the CredRO [Wang et al., 2026] procedure.

Architectures: ResNet56 [He et al., 2016], WideResNet [Zagoruyko and Komodakis, 2016], and DenseNet [Huang et al., 2017] trained on each dataset mentioned above. Training details and hyperparameters are given in Appendix C.

Non-monotonic Maximum Feasible Damage Budget. We first study the maximum local feasible damage to identify which samples are more vulnerable. The relevant empirical object is therefore the upper envelope of attainable vulnerability Γ_ϵ. Figure 2 reports this envelope as a function of baseline EU (EU_{base}) and total baseline uncertainty ($U_0 = AU_{\text{base}} + EU_{\text{base}}$) across datasets and architectures.

The non-monotone pattern is clearest when conditioning on EU_{base}. On MNIST, the envelope peaks in low-to-intermediate uncertainty bins for all three architectures and then declines. The same qualitative behavior persists on CIFAR-10 – peak vulnerability is attained away from the highest-EU regime, while the largest-EU bins exhibit substantially smaller $\Gamma_\epsilon(x, \delta)$. On CIFAR-100, the precise peak location becomes more architecture-dependent, but the envelope again bends and declines rather than rising monotonically. Thus, across all nine dataset–architecture pairs, the most attackable samples are generally not the most epistemically uncertain ones.

Table 1: Comparison of Spearman's rank correlation (ρ) between raw and projected gradient and maximum feasible aleatoric damage.

Dataset	Arch	$\rho(\text{raw})$	$\rho(\text{projected})$	Gain
MNIST	ResNet56	0.7836	0.7901	**0.0065**
	WideResNet	0.6058	0.6601	**0.0543**
	DenseNet	0.6685	0.7042	**0.0357**
CIFAR-10	ResNet56	0.1247	0.1250	**0.0003**
	WideResNet	0.3011	0.3011	**0.0001**
	DenseNet	0.2130	0.2134	**0.0004**
CIFAR-100	ResNet56	-0.3298	-0.3256	**0.0042**
	WideResNet	0.0750	0.0830	**0.0080**
	DenseNet	-0.2030	-0.1964	**0.0066**

First-Order Attack Criterion. We next test Proposition 5.1, which predicts that local vulnerability is governed not by the full aleatoric gradient $\|\nabla AU\|_2$, but by the component orthogonal to epistemic change,$\|\Pi_{\nabla EU^\perp} \nabla AU\|_2$. Intuitively, only directions that increase AU while leaving EU unchanged remain feasible under stealth constraints.

Table 1 summarizes the correlations between realized vulnerability Γ and these first-order quantities. On MNIST, the projected gradient consistently exhibits stronger monotone association with Γ across all architectures than raw gradient. For example, on ResNet56 the Spearman correlation increases from 0.784 for $\|\nabla AU\|_2$ to 0.790 after projection, with larger gains on WideResNet (0.606 to 0.660) and DenseNet (0.669 $to 0.704$). These results closely match the first-order geometric mechanism of Proposition 5.1. On CIFAR-10, the same trend is weaker but still partially visible: the projected statistic is marginally stronger on DenseNet and nearly unchanged on the remaining models. On CIFAR-100, correlations become mixed and occasionally negative, indicating that first-order local geometry alone is insufficient in harder recognition settings. This is consistent with a more complex attack landscape in which curvature, optimization effects, and higher-order interactions become increasingly important. Although, the projected gradient is the theoretically appropriate first-order descriptor under exact EU stealth, its empirical advantage over the raw AU-gradient (as observed from table 2) norm is modest and is clearly distinguished only in some MNIST settings. We investigate this behavior through an AU/EU gradient-alignment diagnostic in Appendix I.1. The gradients generally exhibit weak-to-moderate directional alignment, so projection retains most of the AU-gradient norm for the majority of samples. Consequently, the raw and projected norms often induce similar sample rankings. The small difference between their correlations in Table 1 is therefore consistent with Proposition 5.1, but does not demonstrate a large or uniform empirical advantage from projection.

Table 2: Relation between classifier margin and maximum feasible attack damage.

Dataset	Arch	$\rho(\Gamma, \text{margin})$	Interpretation
MNIST	ResNet56	-0.6218	Strong negative
	WideResNet	-0.4784	Strong negative
	DenseNet	-0.5914	Strong negative
CIFAR-10	ResNet56	-0.1743	Moderate negative
	WideResNet	-0.3199	Moderate negative
	DenseNet	-0.2281	Moderate negative
CIFAR-100	ResNet56	0.3139	Mixed
	WideResNet	-0.1031	Mixed
	DenseNet	0.1745	Mixed

Relation Between Classifier Margin and vulnerability. We next examine whether uncertainty vulnerability is related to classical margin geometry. We use a classifier-level top-two margin surrogate. For each ensemble member with classifier weights $\{w_i\}$ and predicted label $\hat{y}$, we compute $m_{\text{cls}}(x) = \frac{z_{\hat{y}}(x) - z_{(2)}(x)}{\|w_{\hat{y}} - w_{(2)}\|_2}$, and average across ensemble members. This quantity is not the exact input-space boundary distance, but it is the natural representation-space analogue. Table 2 shows a clear pattern on MNIST and CIFAR-10. On MNIST, vulnerability Γ is strongly negatively associated with classifier margin across all architectures, with Spearman correlations of -0.622 (ResNet56), -0.478 (WideResNet), and -0.591 (DenseNet). The same trend persists on CIFAR-10, though more weakly: -0.174, -0.320, and -0.228, respectively. Thus, on simpler and moderately difficult datasets, samples farther from the decision boundary are generally harder to manipulate under uncertainty-stealth constraints. We also test the more specific prediction of Proposition 5.2 by replacing raw margin with the modulation factor $\phi(m) = m\,\sigma(m)(1 - \sigma(m))$. Figure 3 visualizes this effect for the representative

ResNet56 run on CIFAR-100. We provide the figures for MNIST and CIFAR-10 with ResNet56 architecture in Appendix H. We notice that this quantity has a noticeably more stable positive relation with vulnerability. We can clearly observe the trend that the attackability or vulnerability peaks near the center(at a value of 5) of the top-two logit margin distribution. While the margin modulation does not follow this trend exactly due to the nonlinear effect of the sigmoid, it still peaks at an intermediate value and starts sloping downwards after that. This corroborates our analysis in Proposition 5.2 that the vulnerability is not necessarily at the extreme margin values but rather at an intermediate value.

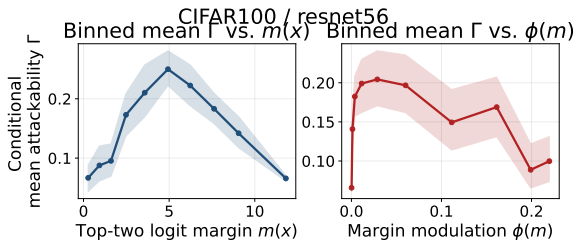


Figure 3: Binned conditional mean attackability $\Gamma_{\epsilon,\delta}(x)$ versus top-two logit margin $m(x)$ (left) and versus the modulation factor $\phi(m(x))$ (right) for CIFAR-100 / ResNet56. Shaded bands show 95% bootstrap confidence intervals. In this harder setting, the raw-margin relation is non-monotone with a peak at intermediate margins, and the $\phi(m)$ panel similarly concentrates larger vulnerability away from both extremes.

Disclosure Ablation. We test the claim that uncertainty about the baseline epistemic level can shrink the feasible stealth region by running a disclosure-ablation experiment on MNIST ResNet56 over the first 1000 test samples. We compare three attackers: a perfect-knowledge attacker that enforces stealth relative to the exact baseline e, a robust attacker that only knows an interval $[e - r, e + r]$, and a midpoint attacker whose perturbation is optimized using the midpoint but evaluated against the robust interval. Figure 4 summarizes how performance degrades as the uncertainty radius r grows. At $r = 0$, all three attackers coincide. As r increases, the robust attacker's feasible damage decreases monotonically: its mean Γ drops from 0.0363 at $r = 0$ to 0.0341 at $r = 0.01$, 0.0288 at $r = 0.02$, 0.0142 at $r = 0.05$, and finally 0 at $r = 0.10$. The midpoint attacker fails even faster under robust evaluation: its feasible rate falls from 1.000 at $r = 0$ to 0.948, 0.937, 0.483, and then 0 across the same radii. The mean stealth-band width contracts from 0.0550 at $r = 0$ to 0.0514, 0.0504, 0.0450, and finally 0. Once the robust band collapses at the largest radius, robust attackability vanishes entirely. This provides direct empirical support for the claim that non-disclosure of baseline EU can reduce attackability by shrinking or eliminating the stealth-feasible action set.

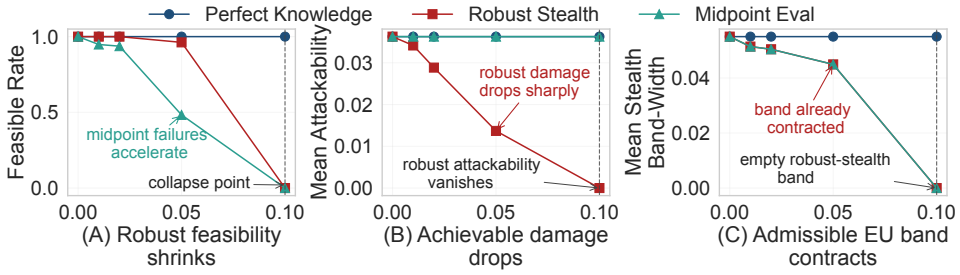


Figure 4: Disclosure-ablation summary on MNIST ResNet56. The x-axis is the attacker's uncertainty radius r around the baseline epistemic level. Panel (A) shows that midpoint-calibrated attacks begin to fail the robust-stealth test well before complete band collapse. Panel (B) shows that the robust attacker achieves progressively less aleatoric damage than the perfect-knowledge attacker as uncertainty about e grows. Panel (C) visualizes the mechanism directly: the admissible epistemic stealth band contracts with r and eventually becomes empty, at which point robust attackability collapses to zero.

Baseline Comparison to Other Attack Objective. In order to understand how attacking the decomposed uncertainty is different than attacking alternate uncertainty objectives, we compare with four baselines: AU-only maximization, entropy maximization, PGD for evasion, and random ℓ_∞. Figure 5 summarizes the central tradeoff: mean AU increase versus the rate at which the epistemic-stealth constraint is violated.

From Figure 5, it is readily observable that attacking the decomposed uncertainty is the only approach which can satisfy the stealth condition while successfully increasing the AU. The proposed attack on

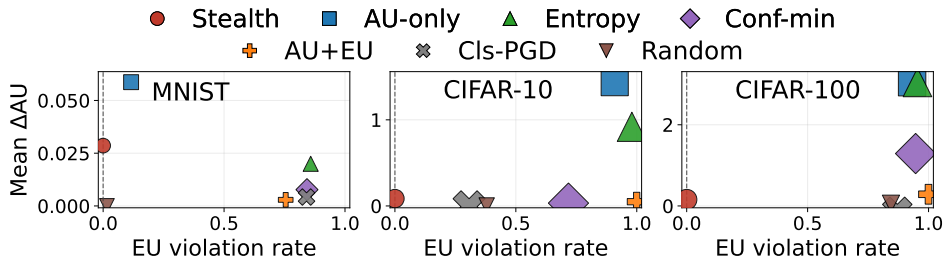


Figure 5: Baseline comparison on ResNet56 across MNIST, CIFAR-10, and CIFAR-100. Each point shows one attack objective, with the x-axis giving the rate of epistemic-stealth violation and the y-axis giving the mean increase in AU. Marker size is proportional to classification accuracy drop. The proposed attack is the only method that stays at zero violation across all three datasets while still producing non-trivial aleatoric damage.

decomposed uncertainty achieves mean AU increases of 0.0286 on MNIST, 0.0853 on CIFAR-10, and 0.1568 on CIFAR-100. On MNIST, the unconstrained AU-only baseline attains a larger raw mean increase (0.0587), but it already violates the stealth constraint on 11.6% of samples; in contrast, the proposed attack preserves feasibility on all samples and still outperforms the feasible subset of that baseline in guaranteed mean damage (0.0286 versus 0.0181 feasible-only mean AU increase). On CIFAR-10 and CIFAR-100, the gap becomes sharper: the largest raw AU gains come from AU-only and entropy-maximizing attacks, but these violate the epistemic constraint on roughly 91–98% of samples and induce large accuracy collapse, making them poor baselines for the stealth-constrained threat model.

Table 3: Compact baseline summary for ResNet56. Each entry reports mean ΔAU / EU-violation rate / accuracy drop. Higher ΔAU is better for the attacker, but lower EU violation is required by the threat model.

Attack objective	MNIST	CIFAR-10	CIFAR-100
Proposed AU+EU stealth	0.0286 / 0.000 / 0.001	0.0853 / 0.000 / 0.147	0.1568 / 0.000 / 0.234
AU-only maximization	0.0587 / 0.116 / 0.010	1.4375 / 0.909 / 0.610	3.0497 / 0.932 / 0.629
Entropy maximization	0.0200 / 0.858 / 0.015	0.9191 / 0.980 / 0.747	3.0427 / 0.954 / 0.704
Classification PGD	0.0042 / 0.841 / 0.073	-0.0093 / 0.307 / 0.950	-0.1438 / 0.871 / 0.768
Random ℓ_∞	0.0000 / 0.016 / 0.000	0.0053 / 0.379 / 0.026	0.0544 / 0.845 / 0.095

Table 3 makes two additional points explicit. First, standard classification attacks are not a good surrogate for uncertainty manipulation as PGD strongly damages accuracy, but it produces negligible or even negative mean AU change on the harder datasets. Second, maximizing scalar uncertainty summaries such as entropy or total uncertainty is not sufficient either, because those objectives typically violate the stealth condition. Overall, these baselines support our claim that uncertainty manipulation is a distinct attack problem – once epistemic stealth is imposed, what matters is how much the DNN is confused without raising any flag to the defenders.

Large-Scale Evaluation on ImageNet To examine whether uncertainty manipulation persists beyond MNIST and CIFAR, we evaluate the proposed attack on 10,000 images from the ImageNet validation set. We provide the complete experimental details in Appendix I. Table 4 reports the achievable AU damage sweeping across ϵ and δ values. At the representative setting $\epsilon = 4/255$ and $\delta = 0.05$, the attack produces a mean AU increase of 0.155. Positive AU damage is obtained for 81.1% of the evaluated images, with no violation of the EU-stealth constraint. Across the six sweeps, increasing δ consistently increases the attainable damage because it enlarges the EU-stealth region. In contrast, increasing ϵ from $4/255$ to $8/255$ produces only a small change at a fixed δ, suggesting that the EU constraint, rather than the input-space budget, is the active restriction here.

Table 4: ImageNet budget sweep using four pretrained ViT predictors and 10,000 validation images. Statistics include all evaluated images; unsuccessful attacks are included as $\Gamma = 0$.

ϵ	δ	Mean Γ	Median Γ	Max Γ	$\Pr(\Gamma > 0)$
$4/255$	0.02	0.080	0.025	2.101	68.3%
$4/255$	0.05	0.155	0.085	1.943	81.1%
$4/255$	0.10	0.342	0.256	2.710	92.0%
$8/255$	0.02	0.081	0.026	2.101	68.5%
$8/255$	0.05	0.156	0.085	1.945	81.0%
$8/255$	0.10	0.345	0.259	2.734	91.9%

For the representative $\epsilon = 4/255, \delta = 0.05$ configuration, the per-image attack outputs have a mean absolute EU change of 0.0244. The prescribed EU tolerance does not exceed for any of the evaluated images. The corresponding AU-damage standard deviation is 0.209, and the 95th percentile of Γ is 0.592. Therefore, the reported AU increases are obtained without violations of the monitored EU constraint. This experiment establishes that AU damage under an EU-stealth constraint is observable on high-resolution natural images and is therefore not confined to smaller datasets like MNIST or CIFAR. For this experiment, we did not impose label preservation as a constraint nor did we measure it after the attack. We therefore do not claim that the ImageNet attacks preserve the predicted class. The result should instead be interpreted as large-scale evidence for the feasibility of manipulating AU while controlling EU.

Operational Impact on Selective Prediction To illustrate a concrete downstream consequence of increasing aleatoric uncertainty, we consider a selective-prediction policy that defers inputs with high AU to abstention or manual review. For each dataset–architecture pair, we set a fixed threshold $\tau_{0.9} = Q_{0.9}\Big(\{AU(\mathcal{P}_{x_i})\}_{i=1}^{500}\Big)$ where $Q_{0.9}$ is the empirical 90th percentile of the clean AU values. The resulting clean deferral rate is 10%. We then apply the label-preserving attack, recompute AU on each perturbed input, and evaluate the attacked inputs using the same threshold without recalibration.

Table 5: Effect of the label-preserving attack on an illustrative AU-threshold deferral policy. The threshold $\tau_{0.9}$ is fixed at the 90th percentile of the 500 clean AU values for each setting, producing a 10% clean deferral rate. "Induced" denotes inputs that are not deferred when clean but are deferred after attack. All settings have 0% label flips and 0% EU-constraint violations.

Dataset	Architecture	$\tau_{0.9}$	Attacked deferral	Induced deferral	Workload multiplier	Mean ΔAU
MNIST	ResNet56	0.0025	93.6%	83.6%	9.36×	0.0264
	DenseNet	0.0074	99.8%	89.8%	9.98×	0.0412
	WideResNet	0.0064	99.8%	89.8%	9.98×	0.1051
CIFAR-10	ResNet56	0.0060	97.2%	87.2%	9.72×	0.0682
	DenseNet	0.0119	97.6%	87.6%	9.76×	0.1065
	WideResNet	0.0114	98.8%	88.8%	9.88×	0.1947
CIFAR-100	ResNet56	0.4346	18.4%	8.4%	1.84×	0.1260
	DenseNet	0.4877	19.0%	9.0%	1.90×	0.1439
	WideResNet	0.8507	17.8%	7.8%	1.78×	0.3387

Table 5 reports the resulting deferral rates. The attack increases the deferral rate to 93.6–99.8% on MNIST, 97.2–98.8% on CIFAR-10, and 17.8–19.0% on CIFAR-100. These correspond to 9.36–9.98×, 9.72–9.88×, and 1.78–1.90× increases, respectively, relative to the clean workload. Across all nine settings, the attack causes no label flips and no violations of the EU-stealth constraint. Thus, the attack directly causes many inputs that would normally receive an automated decision to be abstained on or sent for review, substantially increasing workload and potentially delaying or selectively denying automated service without either an incorrect class prediction or an EU-based alarm. This policy is representative of settings in which AU governs deferral while high EU is treated as evidence of OOD or adversarial input Kotelevskii et al. [2022], Shoeb et al. [2025].

7 Summary of Limitations and Next Steps

In this work, we have studied manipulation of predictive uncertainty through a decomposition-aware lens. We show, both empirically and theoretically, that the maximum aleatoric uncertainty increase is dependent on the precision of information of the epistemic uncertainty disclosed to the attacker. Even then, the samples at intermediate margin region are the ones most vulnerable. Going beyond the credal framework, we show that our findings hold for alternate uncertainty decompositions (e.g., Bayesian, Evidential) as well. Moreover, while our work primarily focus on the vision applications, it also holds for regression tasks. Despite the empirical and theoretical findings, our theoretical results are primarily first-order and local, capturing gradient-level geometry under small perturbations. As observed empirically, these approximations are most predictive in simpler regimes and become less sufficient on more complex datasets and architectures. In future, we plan to extend the analysis to incorporate higher order approximations, diverse settings, and large-scale models (e.g., vision-language models). Beyond analysis, developing defenses, monitoring strategies, and robustness guarantees against such manipulations would be the next immediate step.

Acknowledgments

This work was supported by the U.S. National Science Foundation Award CNS-2312875 and OAC-2530896, by the Office of Naval Research under grant N00014-23-1-2221 and by the Air Force Office

of Scientific Research under grant FA9550-23-1-0261 and by DARPA under cooperative agreement D25AC00374-00.

References

AlgorithmWatch. Google apologizes after its vision ai produced racist results, April 2020. URL `https://algorithmwatch.org/en/google-vision-racism/`. Accessed: 2026-03-25.

A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

M. Caprio, S. Dutta, K. J. Jang, V. Lin, R. Ivanov, O. Sokolsky, and I. Lee. Credal bayesian deep learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=4NHF9AC5ui`.

M. Caprio, S. K. Manchingal, and F. Cuzzolin. Credal and interval deep evidential classifications. *arXiv preprint arXiv:2512.05526*, 2025.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

C. Emde, F. Pinto, T. Lukasiewicz, P. H. S. Torr, and A. Bibi. Towards certification of uncertainty calibration under adversarial attacks. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, Singapore, 2025. Also available as arXiv:2405.13922.

C. Finlay and A. M. Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021.

I. Galil and R. El-Yaniv. Disrupting deep uncertainty estimation without harming accuracy. In *Advances in Neural Information Processing Systems*, 2021.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017.

J. Guynn. Google photos labeled black people 'gorillas'. *USA Today*, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

W. He, J. W. Li, B. So, D. Song, et al. Decision boundary analysis of adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR) Workshop*, 2018.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

N. Kotelevskii, A. Artemenkov, K. Fedyanin, F. Noskov, A. Fishkov, A. Shelmanov, A. Vazhentsev, A. Petiushko, and M. Panov. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323, 2022.

A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features from Tiny Images. 2009.

M. Kull, M. Perello-Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

E. Ledda, G. Scodeller, D. Angioni, G. Piras, A. E. Cinà, G. Fumera, B. Biggio, and F. Roli. On the robustness of adversarial training against uncertainty attacks. *Pattern Recognition*, 2026. Also available as arXiv:2410.21952.

A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

E. Medina. Rite aid's a.i. facial recognition wrongly tagged people of color. *The New York Times*, December 2023. URL `https://www.nytimes.com/2023/12/21/business/rite-aid-ai-facial-recognition.html`. Accessed: 2026-03-25.

National Highway Traffic Safety Administration. Pe 16-007: Tesla crash preliminary evaluation report. Technical report, U.S. Department of Transportation, National Highway Traffic Safety Administration, January 2017. PE 16-007.

S. Rabanser, A. Shahin Shamsabadi, O. Franzese, X. Wang, A. Weller, and N. Papernot. Confidential guardian: Cryptographically prohibiting the abuse of model abstention. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50949–50971. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/rabanser25a.html`.

Y. Shoeb, A. Nowzad, and H. Gottschalk. Out-of-distribution segmentation in autonomous driving: Problems and state of the art. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4310–4320. IEEE, 2025.

C.-J. Simon-Gabriel, Y. Ollivier, L. Bottou, B. Schölkopf, and D. Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In *International conference on machine learning*, pages 5809–5817. PMLR, 2019.

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*, 31, 2018.

O. F. Tuna, F. O. Catak, and M. T. Eskil. Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples. *Multimedia Tools and Applications*, 2021. Also available as arXiv:2102.04150.

O. F. Tuna, F. O. Catak, and M. T. Eskil. Uncertainty as a swiss army knife: new adversarial attack and defense ideas based on epistemic uncertainty. *Complex & Intelligent Systems*, 9(4):3739–3757, 2023.

P. Walley. Statistical reasoning with imprecise probabilities. 1991.

K. Wang, G. A. Faza, F. Cuzzolin, S. L. Chau, D. Moens, and H. Hallez. Learning credal ensembles via distributionally robust optimization. *arXiv preprint arXiv:2602.08470*, 2026.

Z. Yan, Y. Guo, and C. Zhang. Adversarial margin maximization networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1129–1139, 2019.

S. Zagoruyko and N. Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://arxiv.org/abs/2010.01736`.

Appendix Contents

1 Introduction **1**

2 Related Work **3**

3 Problem Setup **3**

4 Baseline Epistemic Disclosure and Vulnerability **4**

5 First-order condition for local attackability **5**

6 Experimental Setup and Results **6**

7 Summary of Limitations and Next Steps **10**

A Proof of Theoretical Claims **13**

B Empirical predictive object and Operational uncertainty functionals **16**

C Implementation Details, Hardware, and Hyperparameter Settings **16**

D Additional Analysis **17**

E Relation of Definition 4.2 to feasibility formulations. **17**

F Additional Discussion on Blessing of Apparent Ignorance of EU **19**

G Baseline Attack Details **22**

H Non-monotonic attackability under attack variants **22**

I Details of the Imagenet Experiment **23**

J Alternative Uncertainty Decompositions **24**

K Extension to Regression **25**

L Limitations **25**

M Broader Impact **26**

A Proof of Theoretical Claims

Theorem A.1 (Envelope Characterization of Feasible Damage(Theorem 4.3))**.** *Fix $\delta > 0$ and assume $EU, AU : \mathfrak{P} \to \overline{\mathbb{R}}$ with $EU(\mathcal{P}) \geq 0$ for all $\mathcal{P} \in \mathfrak{P}$. Let $x_0, x_1 \in \mathcal{X}$ satisfy*

$$EU(\mathcal{P}_{x_0}) = 0, \qquad EU(\mathcal{P}_{x_1}) =: e > 0.$$

Define the following two (nonempty) envelope values,

$$A_0(\delta) := \sup_{\tilde{x} \in \mathcal{X} : EU(\mathcal{P}_{\tilde{x}}) \leq \delta} AU(\mathcal{P}_{\tilde{x}}), \qquad A_1(e, \delta) := \sup_{\tilde{x} \in \mathcal{X} : \max(0, e-\delta) \leq EU(\mathcal{P}_{\tilde{x}}) \leq e+\delta} AU(\mathcal{P}_{\tilde{x}}).$$

Then the following statements are equivalent:

(i) $\Gamma(x_0, \delta) < \Gamma(x_1, \delta)$.

(ii) The strict envelope inequality holds:

$$A_1(e, \delta) - A_0(\delta) > AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0}). \tag{5}$$

(iii) There exists $\hat{x} \in \mathcal{X}$ *such that* $|EU(\mathcal{P}_{\hat{x}}) - e| \leq \delta$ *and*

$$AU(\mathcal{P}_{\hat{x}}) - AU(\mathcal{P}_{x_1}) > \sup_{\tilde{x} \in \mathcal{X}: EU(\mathcal{P}_{\tilde{x}}) \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_{x_0}) \Big]. \tag{6}$$

In particular, (3) *gives a necessary and sufficient condition for the maximum feasible damage budget (i.e., the largest* γ *for which the constraints in* (2) *are feasible) to be strictly smaller when* $EU(\mathcal{P}_x) = 0$ *than when* $EU(\mathcal{P}_x) = e > 0$*.*

Proof of Theorem 4.3. We first rewrite $\Gamma(x, \delta)$ in a convenient way. By the equivalent definition,

$$\Gamma(x, \delta) = \sup_{\tilde{x}: |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \Big] = \Big[\sup_{\tilde{x}: |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta} AU(\mathcal{P}_{\tilde{x}}) \Big] - AU(\mathcal{P}_x).$$

Now consider x_0 with $EU(\mathcal{P}_{x_0}) = 0$. Since $EU(\cdot) \geq 0$, the stealth constraint becomes

$$|EU(\mathcal{P}_{\tilde{x}}) - 0| \leq \delta \quad \Longleftrightarrow \quad EU(\mathcal{P}_{\tilde{x}}) \leq \delta,$$

so

$$\Gamma(x_0, \delta) = \Big[\sup_{\tilde{x}: EU(\mathcal{P}_{\tilde{x}}) \leq \delta} AU(\mathcal{P}_{\tilde{x}}) \Big] - AU(\mathcal{P}_{x_0}) = A_0(\delta) - AU(\mathcal{P}_{x_0}). \tag{7}$$

Next consider x_1 with $EU(\mathcal{P}_{x_1}) = e > 0$. The stealth constraint is

$$|EU(\mathcal{P}_{\tilde{x}}) - e| \leq \delta \quad \Longleftrightarrow \quad e - \delta \leq EU(\mathcal{P}_{\tilde{x}}) \leq e + \delta.$$

Using again that $EU(\cdot) \geq 0$, we may replace the lower bound $e - \delta$ by $\max(0, e - \delta)$ without changing the feasible set (because $EU(\mathcal{P}_{\tilde{x}}) \geq 0$ automatically). Hence,

$$\Gamma(x_1, \delta) = \Big[\sup_{\tilde{x}: \max(0, e-\delta) \leq EU(\mathcal{P}_{\tilde{x}}) \leq e+\delta} AU(\mathcal{P}_{\tilde{x}}) \Big] - AU(\mathcal{P}_{x_1}) = A_1(e, \delta) - AU(\mathcal{P}_{x_1}). \tag{8}$$

(i) $\Longleftrightarrow$ (ii). Combining (7) and (8), we have

$$\Gamma(x_0, \delta) < \Gamma(x_1, \delta) \quad \Longleftrightarrow \quad A_0(\delta) - AU(\mathcal{P}_{x_0}) < A_1(e, \delta) - AU(\mathcal{P}_{x_1}),$$

which is algebraically equivalent to (3).

(i) $\Longleftrightarrow$ (iii). Let

$$B_0 := \sup_{\tilde{x}: EU(\mathcal{P}_{\tilde{x}}) \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_{x_0}) \Big] = \Gamma(x_0, \delta),$$

$$B_1 := \sup_{\tilde{x}: |EU(\mathcal{P}_{\tilde{x}}) - e| \leq \delta} \Big[AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_{x_1}) \Big] = \Gamma(x_1, \delta).$$

Then (i) is exactly $B_0 < B_1$. If $B_0 < B_1$, by the definition of supremum there exists $\hat{x}$ with $|EU(\mathcal{P}_{\hat{x}}) - e| \leq \delta$ and

$$AU(\mathcal{P}_{\hat{x}}) - AU(\mathcal{P}_{x_1}) > \frac{B_0 + B_1}{2} > B_0,$$

which is precisely (4). Conversely, if (4) holds for some $\hat{x}$, then $B_1 \geq AU(\mathcal{P}_{\hat{x}}) - AU(\mathcal{P}_{x_1}) > B_0$, i.e. $\Gamma(x_1, \delta) > \Gamma(x_0, \delta)$. This proves (i) $\Longleftrightarrow$ (iii). The theorem follows. $\square$

Proposition A.2 (First-Order vulnerability Condition (Proposition 5.1))**.** *Let $a(x) := AU(\mathcal{P}_x)$ and $e(x) := EU(\mathcal{P}_x)$ be continuously differentiable in a neighborhood of x. Define*

$$\Gamma_{\text{lin}}(x, \varepsilon) := \sup_{\|r\|_2 \le \varepsilon,\ \nabla e(x)^\top r = 0} \nabla a(x)^\top r.$$

If $\nabla e(x) \neq 0$, then

$$\Gamma_{\text{lin}}(x, \varepsilon) = \varepsilon \left\| \Pi_{\nabla e(x)^\perp} \nabla a(x) \right\|_2,$$

where Π_m denotes orthogonal projection onto the subspace m. If $\nabla e(x) = 0$, then

$$\Gamma_{\text{lin}}(x, \varepsilon) = \varepsilon \|\nabla a(x)\|_2.$$

Proof. By first-order Taylor expansion,

$$a(x + r) = a(x) + \nabla a(x)^\top r + o(\|r\|_2), \qquad e(x + r) = e(x) + \nabla e(x)^\top r + o(\|r\|_2)$$

as $\|r\|_2 \to 0$. Under exact first-order epistemic stealth, $\nabla e(x)^\top r = 0$, so feasible directions lie in the hyperplane $\nabla e(x)^\perp$.

Decompose $\nabla a(x) = \Pi_{\nabla e(x)^\perp} \nabla a(x) + \Pi_{\text{span}\{\nabla e(x)\}} \nabla a(x)$. The second term is orthogonal to any feasible r, and Cauchy–Schwarz yields

$$\nabla a(x)^\top r = \left(\Pi_{\nabla e(x)^\perp} \nabla a(x)\right)^\top r \le \left\|\Pi_{\nabla e(x)^\perp} \nabla a(x)\right\|_2 \|r\|_2 \le \varepsilon \left\|\Pi_{\nabla e(x)^\perp} \nabla a(x)\right\|_2,$$

with equality attained at

$$r^\star = \varepsilon \frac{\Pi_{\nabla e(x)^\perp} \nabla a(x)}{\|\Pi_{\nabla e(x)^\perp} \nabla a(x)\|_2}$$

whenever the projected gradient is nonzero. This establishes the first claim.

If $\nabla e(x) = 0$, the stealth constraint is vacuous to first order, and $\Gamma_{\text{lin}}(x, \varepsilon) = \varepsilon \|\nabla a(x)\|_2$ follows again by Cauchy–Schwarz. □

Proposition A.3 (Moderate-Margin Sensitivity of a Binary Top-Two AU Surrogate (Proposition 5.2))**.** *Let*

$$m(x) = z_{(1)}(x) - z_{(2)}(x) > 0$$

denote the top-two logit margin of a single classifier, and define the binary top-two uncertainty surrogate

$$a_{\text{bin}}(x) := h(\sigma(m(x))),$$

where

$$\sigma(t) = \frac{1}{1 + e^{-t}}, \qquad h(p) = -p \log p - (1 - p) \log(1 - p).$$

Suppose that m is differentiable in a neighborhood of x. Then

$$\nabla a_{\text{bin}}(x) = -\phi(m(x)) \nabla m(x), \qquad \phi(t) := t\, \sigma(t) \big(1 - \sigma(t)\big).$$

Moreover,

$$\phi(0) = 0, \qquad \lim_{t \to \infty} \phi(t) = 0,$$

and ϕ attains its unique maximum at the positive solution $t^\star \approx 1.5434$ of

$$t \tanh(t/2) = 1.$$

Consequently,

$$\|\nabla a_{\text{bin}}(x)\|_2 = \phi(m(x)) \|\nabla m(x)\|_2.$$

For a fixed margin-gradient norm, the local sensitivity of the binary surrogate is therefore largest at a moderate margin.

If $e(x) := EU(\mathcal{P}_x)$ is also differentiable, then under exact first-order epistemic stealth,

$$\Gamma^{\text{bin}}_{\text{lin}}(x, \epsilon) = \epsilon\, \phi(m(x)) \left\|\Pi_{\nabla e(x)^\perp} \nabla m(x)\right\|_2.$$

Proof. Let $p = \sigma(m)$. Direct differentiation gives

$$\frac{dp}{dm} = \sigma(m)\big(1 - \sigma(m)\big) = p(1-p)$$

and

$$\frac{dh}{dp} = \log\left(\frac{1-p}{p}\right) = -m.$$

It follows that

$$\frac{d}{dm} h(\sigma(m)) = -m\,\sigma(m)\big(1 - \sigma(m)\big) = -\phi(m).$$

Applying the multivariate chain rule gives

$$\nabla a_{\mathrm{bin}}(x) = -\phi(m(x))\nabla m(x).$$

To analyze ϕ, observe that $\phi(0) = 0$ and $\phi(t) \to 0$ as $t \to \infty$. Differentiating $\phi(t) = t\sigma(t)(1-\sigma(t))$ and setting the derivative to zero gives

$$t \tanh(t/2) = 1.$$

This equation has a unique positive solution $t^\star \approx 1.5434$, which is the global maximizer because ϕ vanishes at both boundaries.

Finally, applying Proposition 5.1 to a_{bin} gives

$$\Gamma^{\mathrm{bin}}_{\mathrm{lin}}(x, \epsilon) = \epsilon \left\|\Pi_{\nabla e(x)^\perp} \nabla a_{\mathrm{bin}}(x)\right\|_2 = \epsilon\, \phi(m(x)) \left\|\Pi_{\nabla e(x)^\perp} \nabla m(x)\right\|_2 .$$

□

B Empirical predictive object and Operational uncertainty functionals

In the experiments, the predictor object is a CreDRO ensemble with member predictive distributions $\{p_k(\cdot \mid x)\}_{k=1}^K$. We associate to x the empirical predictive credal object $\mathcal{P}_x := \mathrm{conv}\{p_1(\cdot \mid x), \ldots, p_K(\cdot \mid x)\}$, that is, the convex hull of the ensemble member predictions. Our attack and analysis use operational uncertainty functionals derived from these member predictions. Let $H(p) := -\sum_y p(y) \log p(y)$ denote Shannon entropy. Then the experimental aleatoric and epistemic quantities are defined as $AU(\mathcal{P}_x) := \min_{1 \le k \le K} H(p_k(\cdot \mid x))$, and $EU(\mathcal{P}_x) := \max_{1 \le k \le K} H(p_k(\cdot \mid x)) - \min_{1 \le k \le K} H(p_k(\cdot \mid x))$. In our implementation, AU is the smallest member entropy and EU is the range of member entropies. These are the quantities used throughout the experiments, including the attack objective, the feasibility constraint, and the reported correlations.

C Implementation Details, Hardware, and Hyperparameter Settings

Hardware Setup All experiments were conducted on NVIDIA A100 GPUs (80GB memory). Depending on the experiment, we used up to 4 GPUs in parallel. The total compute used for the experiments reported in the paper is approximately 1500 GPU hours. Individual training runs typically require several hours to a few days depending on the dataset and model architecture. In addition to the reported experiments, we conducted preliminary and exploratory runs requiring comparable compute, which are not included in the final results.

Training configuration The runtime configuration in the repository uses SGD with momentum 0.9, weight decay 5×10^{-4}, initial learning rate 0.1, and a multistep schedule with drops at epochs 100 and 150 over 200 epochs. CIFAR training uses random crop and horizontal flip; MNIST uses normalization only. The global CreDRO parameter is 0.5, and the five ensemble members span different DRO fractions.

Attack and evaluation protocol For the main attackability experiments, we approximate

$$\Gamma_\epsilon(x, \delta) = \max_{\|\tilde{x} - x\|_\infty \le \epsilon} \big(AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x)\big)$$

Table 6: Fraction of attacked samples admitting stealthy attack for optimization diagnostics.

Dataset	Arch	Attempts	Feasible rate	Mean $\lvert\Delta EU\rvert$	Max $\lvert\Delta EU\rvert$
MNIST	ResNet56	3	1.000	0.008	0.050
MNIST	DenseNet	3	1.000	0.009	0.050
MNIST	WideResNet	3	1.000	0.015	0.050
CIFAR-10	ResNet56	3	1.000	0.030	0.050
CIFAR-10	DenseNet	3	1.000	0.031	0.050
CIFAR-10	WideResNet	3	1.000	0.032	0.050
CIFAR-100	ResNet56	3	1.000	0.021	0.050
CIFAR-100	DenseNet	3	1.000	0.022	0.050
CIFAR-100	WideResNet	3	1.000	0.025	0.050

subject to epistemic stealth using a penalty-based projected gradient method over an ℓ_∞ ball. The default reported attack parameters are $\epsilon = 0.12549$, step size 0.007843137, 100 PGD steps, stealth budget $\delta = 0.05$, and penalty weight $\lambda = 5.0$. Operationally, the attack maximizes AU while penalizing violations of $|EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \leq \delta$.

The implementation uses up to three adaptive PGD attempts per sample rather than a single run. If an early attempt does not yield a feasible adversarial example, the solver retries with a reduced step size and an increased penalty weight, and it stops once a feasible candidate is found. This provides a lightweight optimization safeguard against declaring low attackability purely because of an unfavorable step size or insufficient penalty strength.

As optimization diagnostics, we report post-attack feasibility under the target EU-stealth constraint together with the empirical distribution of absolute epistemic drift $|EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)|$. In the currently stored runs, the default configuration achieves a feasible rate of 1.0 across the main MNIST, CIFAR-10, and CIFAR-100 dataset–architecture pairs at $\delta = 0.05$, with observed absolute epistemic drifts remaining below the target tolerance. These diagnostics do not replace a full ablation over λ or a projection-based feasibility solver, but they do show that the reported AU gains in the current experiments are not being obtained by overt violations of the intended stealth constraint.

Unless otherwise noted, the reported aggregate analyses are computed on the first 5000 test samples of each dataset–architecture pair in order to keep the complexity manageable while preserving per-sample resolution. For correlation-based claims, we additionally report bootstrapped confidence intervals from 200 resamples.

D Additional Analysis

Optimization Diagnostics From Table 6, we can see that for all our settings, our attack finds the achievable maximum AU within the budgeted EU change.

Robustness Across Attack Budgets. To verify that the qualitative picture is not tied to a single attack setting, we run a targeted budget sweep on MNIST ResNet56 over the first 1000 test samples, varying $\delta \in \{0.02, 0.05, 0.10\}$ and $\epsilon \in \{0.062745, 0.125490\}$. Table 7 shows the main pattern compactly. Increasing either budget increases both mean and maximum achievable damage, but the geometric signatures remain stable: the projected-gradient correlation stays uniformly high, the $\phi(m)$ correlation remains positive in every setting, and the raw-margin correlation remains negative. This indicates that the mechanism identified in the main experiments is not an artifact of a single hyperparameter choice.

Bootstrap Confidence Intervals for Correlation Tables The corresponding 95% bootstrap confidence intervals are reported in Appendix Tables 8 and 9.

E Relation of Definition 4.2 to feasibility formulations.

The quantity $\Gamma_{\epsilon,\delta}(x)$ admits an equivalent feasibility (dual) interpretation. Instead of maximizing the achievable damage, one may ask whether a target increase γ is feasible under the same constraints,

Table 7: Budget-sweep summary on MNIST ResNet56 over the first 1000 test samples. Larger ϵ and δ increase achievable damage, while the projected-gradient and $\phi(m)$ associations remain stable across budgets.

ϵ	δ	Mean Γ	Max Γ	$\rho_s(\Gamma, \|\Pi_{\nabla EU^\perp} \nabla AU\|_2)$	$\rho_s(\Gamma, \phi(m))$	$\rho_s(\Gamma, m)$
0.062745	0.02	0.0054	0.5762	0.8642	0.6552	-0.6551
0.125490	0.02	0.0173	0.7609	0.7576	0.5462	-0.5462
0.062745	0.05	0.0092	0.7877	0.8767	0.6688	-0.6687
0.125490	0.05	0.0270	1.1045	0.7885	0.5784	-0.5784
0.062745	0.10	0.0110	0.8540	0.8772	0.6696	-0.6695
0.125490	0.10	0.0386	1.3896	0.8017	0.5928	-0.5927

Table 8: Spearman correlations between vulnerability Γ and first-order attackability descriptors on the representative 1000-sample subsets, with 95% bootstrap confidence intervals from 200 resamples. The table compares the raw aleatoric gradient norm $\|\nabla AU\|_2$ against its EU-stealth-compatible projected counterpart $\|\Pi_{\nabla EU^\perp} \nabla AU\|_2$. Small gains in the main text can thus be interpreted relative to their sampling uncertainty.

Dataset	Arch	$\rho_s(\Gamma, \|\nabla AU\|_2)$	$\rho_s(\Gamma, \|\Pi_{\nabla EU^\perp} \nabla AU\|_2)$	Gain
MNIST	ResNet56	0.782 [0.752, 0.802]	0.789 [0.758, 0.809]	0.007
	WideResNet	0.593 [0.547, 0.638]	0.650 [0.611, 0.695]	0.056
	DenseNet	0.679 [0.640, 0.712]	0.717 [0.681, 0.753]	0.038
CIFAR-10	ResNet56	0.127 [0.066, 0.189]	0.127 [0.066, 0.189]	-0.000
	WideResNet	0.289 [0.224, 0.364]	0.289 [0.223, 0.364]	-0.000
	DenseNet	0.234 [0.175, 0.304]	0.233 [0.174, 0.302]	-0.001
CIFAR-100	ResNet56	-0.327 [-0.388, -0.252]	-0.325 [-0.388, -0.248]	0.002
	WideResNet	0.026 [-0.050, 0.103]	0.037 [-0.034, 0.115]	0.012
	DenseNet	-0.159 [-0.220, -0.080]	-0.156 [-0.217, -0.079]	0.004

Table 9: Spearman correlations between vulnerability Γ and classifier top-two margin on the representative 1000-sample subsets, with 95% bootstrap confidence intervals from 200 resamples. These intervals quantify the stability of the margin–vulnerability relation across datasets and architectures.

Dataset	Arch	$\rho_s(\Gamma, \text{margin})$ with 95% CI
MNIST	ResNet56	-0.590 [-0.630, -0.540]
	WideResNet	-0.457 [-0.509, -0.406]
	DenseNet	-0.589 [-0.631, -0.545]
CIFAR-10	ResNet56	-0.169 [-0.228, -0.108]
	WideResNet	-0.317 [-0.393, -0.247]
	DenseNet	-0.255 [-0.317, -0.196]
CIFAR-100	ResNet56	0.313 [0.241, 0.368]
	WideResNet	-0.060 [-0.138, 0.019]
	DenseNet	0.140 [0.064, 0.201]

Table 10: Spearman correlations between vulnerability Γ and the margin-modulation factor $\phi(m) = m\,\sigma(m)(1-\sigma(m))$ on the representative 1000-sample subsets, with 95% bootstrap confidence intervals from 200 resamples. We report both the full subset and the correctly classified subset to show how the moderate-margin signal changes after excluding misclassified examples.

Dataset	Arch	Full subset	Correct-only subset
MNIST	ResNet56	0.578 [0.524, 0.616]	0.577 [0.523, 0.620]
	WideResNet	0.335 [0.273, 0.396]	0.337 [0.277, 0.389]
	DenseNet	0.453 [0.397, 0.499]	0.448 [0.385, 0.499]
CIFAR-10	ResNet56	0.202 [0.132, 0.266]	0.249 [0.190, 0.320]
	WideResNet	0.338 [0.273, 0.410]	0.374 [0.307, 0.450]
	DenseNet	–	–
CIFAR-100	ResNet56	-0.159 [-0.213, -0.086]	0.039 [-0.038, 0.109]
	WideResNet	0.105 [0.036, 0.180]	0.220 [0.137, 0.299]
	DenseNet	-0.023 [-0.087, 0.046]	0.132 [0.049, 0.203]

i.e., whether there exists $\tilde{x}$ such that

$$\|\tilde{x} - x\|_\infty \le \epsilon, \quad |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \le \delta, \quad AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \ge \gamma.$$

From this perspective, $\Gamma_\epsilon(x, \delta)$ is the supremum over all feasible damage levels γ. This establishes the equivalence between the maximum-damage formulation used in experiments and the minimum-distance or feasibility-based viewpoints commonly used in adversarial analysis.

F Additional Discussion on Blessing of Apparent Ignorance of EU

We discuss additional interpretation of the results from Theorem 4.3. A possible interpretation of condition (3) is as follows. Following [Walley, 1991], we may posit that in many imprecise probability learning settings EU does not vanish with a finite sample size, so that typically $EU(\mathcal{P}_x) > 0$. In that case, if the defender explicitly computes $EU(\mathcal{P}_x)$ (and this baseline value is accessible to the attacker), then the EU-based stealth constraint in (2) centers the attacker's admissible perturbations around a positive level $e = EU(\mathcal{P}_x)$, potentially granting access to credal sets with substantially higher achievable AU (and therefore damage) than those available under the low-epistemic cap $EU(\mathcal{P}_x) \le \delta$.

Condition (3) states precisely that this "opportunity gap" $A_1(e, \delta) - A_0(\delta)$ dominates any baseline difference $AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0})$, so that the attacker has strictly more room to increase AU at x_1 than at x_0. In this sense, *knowledge or disclosure* (or accurate estimation) of the baseline EU can increase vulnerability under EU-based monitoring, while treating the system as "epistemically certain" (or not revealing EU) can shrink the attacker's stealthy action set and reduce the maximal achievable damage.

An Extended Interpretation of Theorem 4.3 We can provide an even easier way to interpret our necessary and sufficient condition for the "blessing of apparent ignorance" phenomenon.

Identifying the Room Available. For a fixed $x \in \mathcal{X}$ and $\delta > 0$, define the maximal feasible increase in AU under EU-based stealth as

$$\Gamma(x, \delta) := \sup_{\tilde{x} \in \mathcal{X} : |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \le \delta} \Big(AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x) \Big).$$

This can be rewritten as a "ceiling minus baseline" identity:

$$\Gamma(x, \delta) = \Big(\sup_{\tilde{x} : |EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \le \delta} AU(\mathcal{P}_{\tilde{x}}) \Big) - AU(\mathcal{P}_x).$$

Hence $\Gamma(x, \delta)$ measures the *room* to increase AU while remaining stealthy.

Two Different Ceilings. Assume $EU(\mathcal{P}) \ge 0$ for all predictive credal sets $\mathcal{P}$. If $EU(\mathcal{P}_{x_0}) = 0$, then the stealth constraint reduces to

$$|EU(\mathcal{P}_{\tilde{x}}) - 0| \le \delta \quad \Longleftrightarrow \quad EU(\mathcal{P}_{\tilde{x}}) \le \delta,$$

so the best attainable AU under stealth is

$$A_0(\delta) := \sup_{\tilde{x}: EU(\mathcal{P}_{\tilde{x}}) \le \delta} AU(\mathcal{P}_{\tilde{x}}), \qquad \text{and thus} \qquad \Gamma(x_0, \delta) = A_0(\delta) - AU(\mathcal{P}_{x_0}).$$

If instead $EU(\mathcal{P}_{x_1}) = e > 0$, then stealth requires

$$|EU(\mathcal{P}_{\tilde{x}}) - e| \le \delta \quad \iff \quad e - \delta \le EU(\mathcal{P}_{\tilde{x}}) \le e + \delta,$$

and the best attainable AU under stealth is

$$A_1(e, \delta) := \sup_{\tilde{x}: e-\delta \le EU(\mathcal{P}_{\tilde{x}}) \le e+\delta} AU(\mathcal{P}_{\tilde{x}}), \qquad \text{and thus} \qquad \Gamma(x_1, \delta) = A_1(e, \delta) - AU(\mathcal{P}_{x_1}).$$

Opportunity Gap. The difference

$$A_1(e, \delta) - A_0(\delta)$$

is the *opportunity gap*: it compares the maximal AU attainable in the EU-band $[e - \delta, e + \delta]$ (relevant when the baseline EU is $e > 0$) to the maximal AU attainable under the low-epistemic cap $EU \le \delta$ (relevant when the baseline EU is 0).

Dominating Baseline Differences. Saying that the attacker has strictly more room to increase AU at x_1 than at x_0 means

$$\Gamma(x_1, \delta) > \Gamma(x_0, \delta).$$

Substituting the expressions above yields

$$A_1(e, \delta) - AU(\mathcal{P}_{x_1}) > A_0(\delta) - AU(\mathcal{P}_{x_0}),$$

which is equivalent (by rearranging terms) to the envelope condition

$$A_1(e, \delta) - A_0(\delta) > AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0}).$$

Thus, when the opportunity gap $A_1(e, \delta) - A_0(\delta)$ dominates the baseline difference $AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0})$, the attacker has strictly more room to increase AU under EU-based stealth monitoring at x_1 than at x_0.

A simplified exampled: We show the blessing of (apparent) ignorance of EU with an example in binary classification Let $\mathcal{Y} = \{0, 1\}$ and identify each $P \in \Delta_{\mathcal{Y}}$ with a Bernoulli distribution $\mathrm{Bern}(p)$, $p \in [0, 1]$. Consider a parametric family of credal sets given by intervals of Bernoulli parameters:

$$\mathcal{P}(m, u) := \left\{ \mathrm{Bern}(p) : p \in \left[m - \tfrac{u}{2}, m + \tfrac{u}{2} \right] \cap [0, 1] \right\}, \qquad m \in [0, 1], u \in [0, 1].$$

Define EU as *imprecision width*

$$EU(\mathcal{P}(m, u)) := u,$$

and AU as *lower entropy*

$$AU(\mathcal{P}(m, u)) := \inf_{p \in [m-u/2, m+u/2] \cap [0,1]} h(p),$$

where $h(p) := -p \log_2 p - (1 - p) \log_2 (1 - p)$ is the binary Shannon entropy.

We now impose a restriction on the set of reachable credal sets $\mathfrak{P}$ (as the input varies), capturing the idea that low EU is encountered only in "easy" regions with strongly biased predictions, whereas moderate EU is encountered near decision boundaries where predictions are closer to $1/2$. Concretely, fix $\delta = 0.1$ and $e = 0.4$ and assume:

- For all reachable credal sets with $EU = u \in [0, \delta]$, the center is $m_0 = 0.05$, i.e. $\mathcal{P}_{\tilde{x}} = \mathcal{P}(0.05, u)$.
- For all reachable credal sets with $EU = u \in [e - \delta, e + \delta] = [0.3, 0.5]$, the center is $m_1 = 0.5$, i.e. $\mathcal{P}_{\tilde{x}} = \mathcal{P}(0.5, u)$.

Choose baseline points x_0, x_1 such that

$$\mathcal{P}_{x_0} = \mathcal{P}(0.05, 0) \quad \big(EU(\mathcal{P}_{x_0}) = 0\big), \qquad \mathcal{P}_{x_1} = \mathcal{P}(0.5, 0.4) \quad \big(EU(\mathcal{P}_{x_1}) = e\big).$$

We compute the envelopes in (3). For $u \in [0, 0.1]$ centered at 0.05, the lower entropy equals $h(0.05 - u/2)$, which decreases with u; hence the maximum over this family is attained at $u = 0$, yielding

$$A_0(\delta) = \sup_{u \in [0,\delta]} AU(\mathcal{P}(0.05, u)) = AU(\mathcal{P}(0.05, 0)) = h(0.05) \approx 0.2864.$$

For the band $u \in [0.3, 0.5]$ centered at 0.5, the lower entropy equals $h(0.5 - u/2)$, which is larger when u is smaller; thus

$$A_1(e, \delta) = \sup_{u \in [0.3,0.5]} AU(\mathcal{P}(0.5, u)) = AU(\mathcal{P}(0.5, 0.3)) = h(0.35) \approx 0.9341.$$

Moreover,

$$AU(\mathcal{P}_{x_0}) = h(0.05) \approx 0.2864, \qquad AU(\mathcal{P}_{x_1}) = AU(\mathcal{P}(0.5, 0.4)) = h(0.30) \approx 0.8813.$$

Therefore,

$$A_1(e, \delta) - A_0(\delta) \approx 0.9341 - 0.2864 = 0.6477 \quad > \quad 0.8813 - 0.2864 = 0.5949 = AU(\mathcal{P}_{x_1}) - AU(\mathcal{P}_{x_0}),$$

and condition (3) is satisfied.

The above construction illustrates the same structural mechanism behind (3):

- In the low-epistemic region $\{EU(\mathcal{P}) \leq \delta\}$, the reachable credal sets have comparatively small maximal AU, so $A_0(\delta)$ is small.
- In the band $\{EU(\mathcal{P}) \in [e - \delta, e + \delta]\}$ around a positive baseline $e > 0$, the reachable credal sets include substantially larger AU, so $A_1(e, \delta)$ is much larger.

When this "opportunity gap" $A_1(e, \delta) - A_0(\delta)$ dominates any baseline difference in $AU(\mathcal{P}_x)$, the attacker has strictly more room to increase AU under EU-based stealth monitoring.

Fixed-sample disclosure. Disclosure concerns the information available to the attacker for a fixed sample x. Let the true baseline be $e := EU(\mathcal{P}_x) \in \mathcal{E}_x$, where $\mathcal{E}_x \subseteq \mathbb{R}_+$ represents the attacker's uncertainty about that baseline. Define the exact-knowledge and robust feasible sets

$$\mathcal{F}_{\text{exact}}(x, e) := \{\tilde{x} : \|\tilde{x} - x\|_\infty \leq \epsilon,\ |EU(\mathcal{P}_{\tilde{x}}) - e| \leq \delta\}, \tag{9}$$

$$\mathcal{F}_{\text{rob}}(x, \mathcal{E}_x) := \left\{\tilde{x} : \|\tilde{x} - x\|_\infty \leq \epsilon,\ \sup_{e' \in \mathcal{E}_x} |EU(\mathcal{P}_{\tilde{x}}) - e'| \leq \delta\right\}. \tag{10}$$

Because $e \in \mathcal{E}_x$, every robustly feasible perturbation is also feasible under exact knowledge:

$$\mathcal{F}_{\text{rob}}(x, \mathcal{E}_x) \subseteq \mathcal{F}_{\text{exact}}(x, e). \tag{11}$$

Define $\Gamma_{\text{exact}}(x, e, \delta)$ and $\Gamma_{\text{rob}}(x, \mathcal{E}_x, \delta)$ by taking the supremum of $AU(\mathcal{P}_{\tilde{x}}) - AU(\mathcal{P}_x)$ over $\mathcal{F}_{\text{exact}}(x, e)$ and $\mathcal{F}_{\text{rob}}(x, \mathcal{E}_x)$, respectively. The set inclusion in (11) therefore gives

$$\Gamma_{\text{rob}}(x, \mathcal{E}_x, \delta) \leq \Gamma_{\text{exact}}(x, e, \delta). \tag{12}$$

More generally, if $\mathcal{E}_1 \subseteq \mathcal{E}_2$, then $\mathcal{F}_{\text{rob}}(x, \mathcal{E}_2) \subseteq \mathcal{F}_{\text{rob}}(x, \mathcal{E}_1)$, so increasing uncertainty about the baseline EU cannot increase the robust feasible damage budget.

For the interval case $\mathcal{E}_x = [\underline{e}_x, \overline{e}_x]$, writing $u := EU(\mathcal{P}_{\tilde{x}})$, the robust stealth constraint is equivalent to

$$u \in [\underline{e}_x - \delta, \underline{e}_x + \delta] \cap [\overline{e}_x - \delta, \overline{e}_x + \delta] = \left[\overline{e}_x - \delta, \underline{e}_x + \delta\right]. \tag{13}$$

This band contracts as $\overline{e}_x - \underline{e}_x$ increases and is empty whenever $\overline{e}_x - \underline{e}_x > 2\delta$. Equations (11)–(12) establish the non-strict disclosure result. A strict reduction additionally requires an AU-improving point that is feasible under exact knowledge but excluded by the robust constraint; Figure 4 shows that this occurs in our evaluated setting. Appendix F separately gives an illustrative construction in which the strict inequality characterized by Theorem 4.3 occurs.

Table 11: Appendix baseline summary for ResNet56. Each entry reports mean ΔAU / mean $|\Delta EU|$ / EU-violation rate / label-flip rate / accuracy drop.

Attack objective	MNIST	CIFAR-10	CIFAR-100
Proposed AU+EU stealth	0.0286 / 0.0081 / 0.000 / 0.002 / 0.001	0.0853 / 0.0300 / 0.000 / 0.179 / 0.147	0.1568 / 0.0211 / 0.000 / 0.329 / 0.234
Classification PGD	0.0042 / 0.4110 / 0.841 / 0.073 / 0.073	-0.0093 / 0.1532 / 0.307 / 0.953 / 0.950	-0.1438 / 0.7000 / 0.871 / 0.846 / 0.768
Entropy maximization	0.0200 / 0.7498 / 0.858 / 0.018 / 0.015	0.9191 / 0.6494 / 0.980 / 0.795 / 0.747	3.0427 / 0.6403 / 0.954 / 0.935 / 0.704
Confidence minimization	0.0077 / 0.4283 / 0.843 / 0.040 / 0.035	0.0330 / 0.4585 / 0.718 / 0.745 / 0.695	1.2905 / 0.5771 / 0.947 / 0.929 / 0.700
Total uncertainty maximization	0.0029 / 0.6297 / 0.755 / 0.001 / 0.000	0.0494 / 2.0063 / 1.000 / 0.204 / 0.163	0.2878 / 3.1235 / 1.000 / 0.380 / 0.206
AU-only maximization	0.0587 / 0.0401 / 0.116 / 0.011 / 0.010	1.4375 / 0.2286 / 0.909 / 0.659 / 0.610	3.0497 / 0.6384 / 0.932 / 0.854 / 0.629
Random ℓ_∞	0.0000 / 0.0032 / 0.016 / 0.000 / 0.000	0.0053 / 0.1378 / 0.379 / 0.051 / 0.026	0.0544 / 0.4399 / 0.845 / 0.231 / 0.095

G Baseline Attack Details

The baseline-comparison experiment in Section 5 uses the same perturbation budget, step size, and number of PGD iterations as the main proposed attack, but replaces the optimization objective. For an ensemble with member predictions $\{p_k(\cdot \mid x)\}_{k=1}^K$, let $\bar{p}(\cdot \mid x) = \frac{1}{K}\sum_{k=1}^K p_k(\cdot \mid x)$ denote the ensemble-mean predictive distribution, let $H(\bar{p})$ denote predictive entropy, and let $\hat{y}(x) = \arg\max_c \bar{p}_c(\cdot \mid x)$.

The compared attacks are:

- **Proposed AU+EU stealth**: projected-gradient attack maximizing $AU(\mathcal{P}_{\tilde{x}})$ with a penalty for violating $|EU(\mathcal{P}_{\tilde{x}}) - EU(\mathcal{P}_x)| \le \delta$.
- **Classification PGD**: untargeted PGD maximizing the per-sample negative log-likelihood $-\log \bar{p}_y(\cdot \mid \tilde{x})$ of the true label.
- **Entropy maximization**: PGD maximizing $H(\bar{p}(\cdot \mid \tilde{x}))$.
- **Confidence minimization**: PGD minimizing $\max_c \bar{p}_c(\cdot \mid \tilde{x})$, equivalently maximizing its negative.
- **Total uncertainty maximization**: PGD maximizing $AU(\mathcal{P}_{\tilde{x}}) + EU(\mathcal{P}_{\tilde{x}})$. In the current repository implementation, this equals the maximum member entropy.
- **AU-only maximization**: PGD maximizing $AU(\mathcal{P}_{\tilde{x}})$ without any epistemic-stealth penalty.
- **Random ℓ_∞**: non-adaptive random perturbation sampled uniformly from the same ℓ_∞ ball.

For all baselines, clean and adversarial labels are computed from the ensemble-mean prediction. The uncertainty decomposition used in evaluation matches the repository implementation: AU is the minimum member entropy, and EU is the difference between maximum and minimum member entropy across ensemble members. The summary statistics reported in the main text and below are computed over 5000 test samples for ResNet56 on MNIST, CIFAR-10, and CIFAR-100. We report mean ΔAU, mean absolute ΔEU, epistemic-constraint violation rate, label-flip rate, and accuracy drop.

Table 11 shows that the proposed method is the only attack that consistently combines non-trivial aleatoric damage with negligible epistemic drift. The remaining objectives each fail in a different way: classification PGD mostly damages accuracy rather than AU, entropy/confidence/total-uncertainty baselines substantially distort EU, and the AU-only baseline attains high raw AU gain at the cost of widespread stealth violations on CIFAR-10 and CIFAR-100. This is why the main text evaluates the baselines through the joint lens of AU gain and EU-feasibility rather than raw uncertainty distortion alone.

H Non-monotonic attackability under attack variants

To verify that the main non-monotonicity phenomenon is not specific to our original credal decomposition or to attacks that allow prediction changes, we repeated the damage-versus-baseline-EU analysis for two additional attack settings. First, we reran the proposed attack using the standard deep-ensemble uncertainty decomposition, where AU is the ensemble-averaged entropy and EU is the mutual-information term. Second, we evaluated a label-preserving variant of the proposed attack, which enforces $\hat{y}(\tilde{x}) = \hat{y}(x)$ in addition to the ℓ_∞ and epistemic-stealth constraints. For each dataset and architecture, we binned test samples by baseline epistemic uncertainty EU_{base} and plotted the

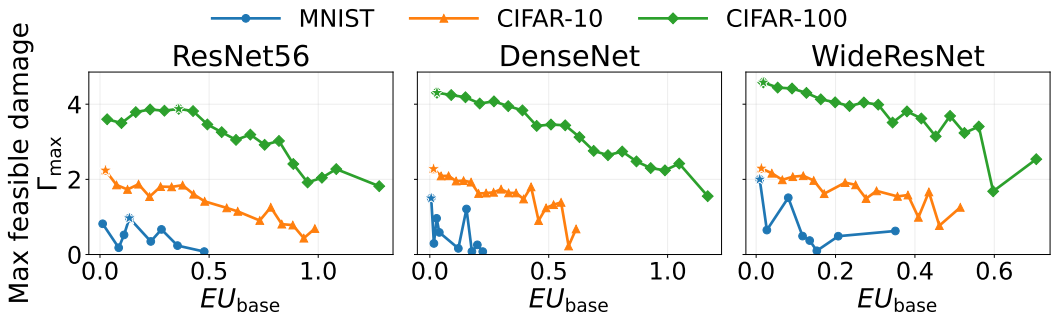


Figure 6: Maximum feasible aleatoric damage as a function of baseline epistemic uncertainty under the standard deep-ensemble uncertainty decomposition. For each dataset and architecture, we bin samples by EU_{base} and report the upper envelope of feasible damage within each bin. The peak damage typically occurs away from both extremes, corroborating the intermediate-EU vulnerability pattern under the standard expected-entropy / mutual-information decomposition.

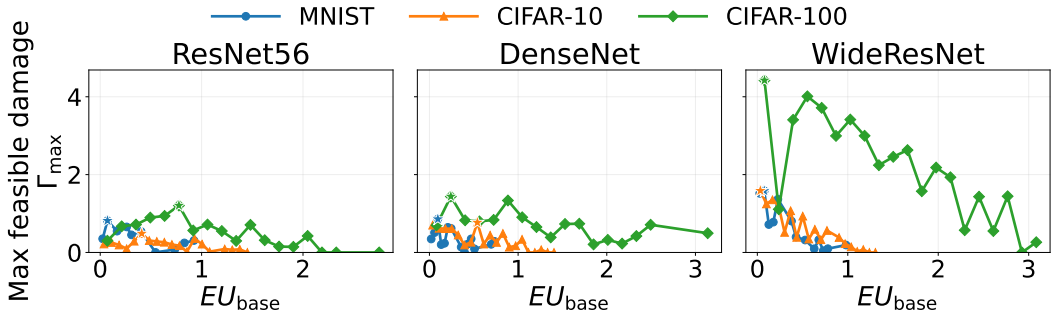


Figure 7: Maximum feasible aleatoric damage as a function of baseline epistemic uncertainty for the label-preserving attack variant. Here the adversarial example is constrained to keep the ensemble prediction unchanged, $\hat{y}(\tilde{x}) = \hat{y}(x)$. The resulting curves remain non-monotone across datasets and architectures, indicating that the intermediate-EU damage peak is not an artifact of label changes.

upper envelope of feasible aleatoric damage, i.e., the maximum observed ΔAU within each bin among attack outcomes satisfying the prescribed stealth constraint. Figures 6 and 7 show that the same qualitative pattern persists in both settings: the largest feasible damage typically does not occur at the lowest or highest baseline epistemic levels, but rather at intermediate values of EU_{base}.

I Details of the Imagenet Experiment

We construct the evaluation subset deterministically by sorting the validation-image paths lexicographically and selecting the first 10,000 images. This produces a class-balanced subset containing 50 images from each of the first 200 ImageNet classes.

We construct a finite predictive set from four predictors available through `torchvision` – three ViT-B/16 weight or training-recipe variants (ImageNet-1K V1, SWAG end-to-end, and SWAG linear) and one ViT-B/32 ImageNet-1K V1 predictor. Thus, this predictive set captures ambiguity arising from different pretrained weights, training recipes, and one architectural variation. It should not be interpreted as an ensemble of four independently retrained models.

For consistency with the main experiments, we define

$$AU(\mathcal{P}_x) = \min_{1 \le k \le 4} H(p_k(\cdot \mid x))$$

and

$$EU(\mathcal{P}_x) = \max_{1 \le k \le 4} H(p_k(\cdot \mid x)) - \min_{1 \le k \le 4} H(p_k(\cdot \mid x)),$$

Table 12: Alignment between the input gradients of AU and EU over 5,000 samples per setting. Here c is their signed cosine similarity and $r = \|\Pi_{\nabla EU^\perp} \nabla AU\|_2 / \|\nabla AU\|_2$ is the per-sample fraction of the AU-gradient norm retained after projection.

Dataset	Architecture	Mean c	Median c	Mean $\lvert c\rvert$	Mean r	Median r
MNIST	ResNet56	-0.105	-0.113	0.189	0.971	0.987
	DenseNet	-0.329	-0.352	0.355	0.914	0.935
	WideResNet	-0.260	-0.300	0.336	0.921	0.942
CIFAR-10	ResNet56	0.083	0.115	0.171	0.973	0.989
	DenseNet	0.089	0.132	0.201	0.964	0.985
	WideResNet	0.197	0.253	0.290	0.938	0.959
CIFAR-100	ResNet56	-0.170	-0.057	0.271	0.914	0.986
	DenseNet	-0.124	-0.001	0.286	0.912	0.980
	WideResNet	0.022	0.183	0.344	0.896	0.955

where $p_k(\cdot \mid x)$ is the predictive distribution of the k-th ViT and H denotes Shannon entropy. We optimize the attack for 40 steps with step size $0.25/255$ and penalty weight $\lambda = 5$. We evaluate $\epsilon \in \{4/255, 8/255\}$ and $\delta \in \{0.02, 0.05, 0.10\}$.

I.1 AU/EU Gradient-Alignment Diagnostic

To explain the modest difference between the raw and projected gradient descriptors in Table 1, we measure the directional alignment between

$$g_A(x) := \nabla_x AU(\mathcal{P}_x) \qquad \text{and} \qquad g_E(x) := \nabla_x EU(\mathcal{P}_x).$$

For each sample, we compute the cosine similarity

$$c(x) := \frac{\langle g_A(x), g_E(x) \rangle}{\|g_A(x)\|_2 \|g_E(x)\|_2}$$

and the fraction of the AU-gradient norm retained after removing its component parallel to $g_E(x)$:

$$r(x) := \frac{\|\Pi_{g_E(x)^\perp} g_A(x)\|_2}{\|g_A(x)\|_2}.$$

Whenever both gradients are nonzero,

$$r(x) = \sqrt{1 - c(x)^2}.$$

Table 12 reports these quantities over 5,000 samples for each dataset–architecture pair. We include both the signed cosine c and its absolute value because the signed mean can be small through cancellation between positively and negatively aligned gradients, even when their directional dependence is non-negligible.

Across the nine settings, the mean absolute cosine similarity ranges from 0.171 to 0.355. The mean per-sample retained norm ranges from 0.896 to 0.973, while the median ranges from 0.935 to 0.989. Projection therefore removes only a limited fraction of the AU-gradient norm for most samples. As a result, the raw and projected norms tend to rank samples similarly, explaining why their correlations with realized attackability in Table 1 differ only slightly in most settings.

These results support the projected gradient as the correct first-order quantity under the local EU-stealth condition. They do not show a large or uniform empirical predictive advantage over the raw gradient norm, because the projection commonly leaves the AU gradient nearly unchanged.

J Alternative Uncertainty Decompositions

The constrained attack is not tied to the credal uncertainty functionals used in the main experiments. For any predictive mechanism that supplies differentiable input-dependent functionals $AU(x)$ and $EU(x)$, Proposition 5.1 applies locally whenever

$$\left\|\Pi_{\nabla EU(x)^\perp} \nabla AU(x)\right\|_2 > 0.$$

Table 13: Label-preserving attack on an MC-dropout ResNet-18 over the complete CIFAR-10 test set.

Model and data	T	ϵ	δ	Mean / median / max ΔAU	Positive ΔAU	Mean $\lvert\Delta EU\rvert$	EU violations	Label flips
MC-dropout ResNet-18, CIFAR-10 ($n = 10{,}000$)	10	$8/255$	0.05	0.996 / 0.990 / 2.129	99.71%	0.0346	0%	0%

Table 14: Regression attack results over five independently trained seeds. "Positive" is the percentage of test cases with $\Delta AU > 0$.

Dataset	Cases/seed	ϵ	δ	Mean ΔAU	Median ΔAU	Max ΔAU	Positive	Mean $\lvert\Delta EU\rvert$	EU violations
Diabetes	111	0.10	0.002	0.2285 ± 0.0508	0.0446 ± 0.0092	5.036 ± 2.382	$93.87\% \pm 2.95$ pp	0.001436 ± 0.000085	0%
YearPredictionMSD	51,630	0.05	0.002	0.1491 ± 0.0033	0.0602 ± 0.0006	2.977 ± 0.400	$88.35\% \pm 0.30$ pp	0.001270 ± 0.000006	0%

This is a geometric condition, not a condition on the statistical accuracy or calibration of the uncertainty estimates. It guarantees a locally AU-increasing direction with zero first-order EU change, but does not by itself guarantee a particular finite-budget damage level or transfer the credal-set opportunity analysis to every UQ method. We therefore evaluate two specific alternatives below and do not claim coverage of Bayesian UQ methods generally.

J.1 Approximate Bayesian MC Dropout

We additionally evaluate a pretrained MC-dropout ResNet-18 on the complete 10,000-image CIFAR-10 test set. The checkpoint uses dropout rate 0.3. During inference and attack optimization, dropout remains active while BatchNorm statistics remain frozen. From $T = 10$ stochastic predictions $\{p_t(\cdot \mid x)\}_{t=1}^{T}$, we define

$$AU_{\mathrm{MC}}(x) := \frac{1}{T} \sum_{t=1}^{T} H(p_t(\cdot \mid x)) \,, \tag{14}$$

$$EU_{\mathrm{MC}}(x) := H\left(\frac{1}{T} \sum_{t=1}^{T} p_t(\cdot \mid x) \right) - AU_{\mathrm{MC}}(x). \tag{15}$$

The label-preserving attack is re-optimized for these quantities using 20 steps and at most two attempts. Table 13 shows that it produces positive AU damage on 99.71% of the test set, with no label flips or measured EU violations. Because every clean MC-mean prediction is preserved, clean and adversarial accuracy are both 92.18%. This establishes vulnerability for this finite MC predictive set and this approximate Bayesian model; it does not establish the same behavior for Bayesian UQ methods generally.

K Extension to Regression

The constrained objective also applies to regression when differentiable AU and EU functionals are available. We evaluate five independently trained ensembles, each containing five heteroscedastic Gaussian regressors. If member k predicts mean $\mu_k(x)$ and noise variance $\sigma_k^2(x)$, we define

$$AU_{\mathrm{reg}}(x) := \frac{1}{K} \sum_{k=1}^{K} \sigma_k^2(x), \qquad EU_{\mathrm{reg}}(x) := \mathrm{Var}_k[\mu_k(x)] \,.$$

Table 14 reports the mean and standard deviation across five seeds, with each seed evaluated on its complete test set. The perturbation budgets are measured in training-standardized feature units. Across all 555 Diabetes evaluations and 258,150 YearPredictionMSD evaluations, no attack exceeds the EU tolerance. The mean absolute changes in the ensemble predictive mean are 0.0415 ± 0.0133 and 0.0775 ± 0.0008, respectively.

L Limitations

Our analysis relies on several assumptions that may limit its generality. First, the empirical study instantiates uncertainty using entropy-based functionals over ensemble predictions, where aleatoric

uncertainty is defined as the minimum member entropy and epistemic uncertainty as the entropy range. While this choice is consistent with the credal framework, there are alternate credal set formulation and uncertainty decomposition framework for Credal prediction (e.g., forming credal set via likelihood, decalibration based credal set formation, distance based uncertainty decomposition etc.). It would be interesting to observe if these approaches for Credal set formulation and uncertainty decomposition also hold the same vulnerabilities.

Second, the theoretical analysis is primarily first-order and local, characterizing vulnerability through gradient geometry under small perturbations. As observed empirically, this approximation is strongly predictive in simpler settings but becomes less sufficient on more complex datasets, suggesting that higher-order effects, optimization dynamics, and representation heterogeneity play an important role in harder regimes. Finally, our experiments focus on classification tasks with ensemble-based predictors. Extending the analysis to other settings, such as large-scale foundation models, structured prediction tasks, or non-ensemble uncertainty estimators, remains an open direction.

M Broader Impact

This work studies adversarial manipulation of predictive uncertainty in machine learning systems. On the positive side, our findings can help improve the reliability and safety of uncertainty-aware models by identifying previously overlooked vulnerabilities in uncertainty decomposition. In safety-critical applications such as autonomous driving, medical decision support, and risk-sensitive classification, understanding how uncertainty signals can be manipulated is essential for designing robust monitoring and defense mechanisms.

At the same time, the proposed attack formulation may introduce potential risks if misused. In particular, adversaries could exploit these insights to manipulate uncertainty estimates without triggering standard epistemic alarms, potentially causing systems to misinterpret adversarial inputs as naturally ambiguous cases. Such vulnerabilities could undermine trust in uncertainty-aware decision-making systems.

We emphasize that this work is intended as a diagnostic and analytical contribution rather than a deployment tool. We hope that exposing these vulnerabilities will encourage the development of more robust uncertainty estimation methods, improved monitoring strategies, and defenses against uncertainty-targeted attacks. Future work should investigate safeguards, detection mechanisms, and robustness guarantees for uncertainty-aware systems.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims are clearly stated in the Introduction and summarized in Section 1, including the formalization of uncertainty vulnerability, the theoretical results (Theorem 4.3, Propositions 5.1–5.2), and the empirical findings validating these claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations of the proposed approach, including assumptions on uncertainty functionals, reliance on first-order analysis, and restriction to ensemble-based predictors, are discussed in the limitations paragraph in the Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical results are formally stated in Sections 4 and 5 with clearly specified assumptions. Complete proofs are provided in Appendix A.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Experimental setup, datasets, model architectures, and evaluation protocols are described in Section 6, with full implementation details, hyperparameters, and attack procedures provided in Appendix C.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The paper provides the link to the repository where we are currently organizing the codes with complete instructions to reproduce the results.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 6 describes the datasets, model architectures, uncertainty quantities, and evaluation setup. Appendix C provides the training configuration, optimizer, hyperparameters, attack settings, PGD procedure, and evaluation protocol needed to understand and reproduce the experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The paper reports statistical evidence using Spearman rank correlations and bootstrap confidence intervals. In particular, Figure 3 includes 95% bootstrap confidence intervals, and Appendix C describes the use of 200 bootstrap resamples for correlation-based claims.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Compute resources are detailed in Appendix C, including GPU type (NVIDIA A100, 80GB), number of GPUs used (up to 4), and an estimate of total compute (approximately 1500 GPU-hours).

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work conforms to the NeurIPS Code of Ethics. The study uses standard public benchmark datasets and focuses on analyzing adversarial vulnerabilities of uncertainty-aware predictors. No human-subject experiments, private data, or personally identifying information are involved.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The work has both positive and negative societal implications. Positively, it identifies vulnerabilities in uncertainty-aware systems and can motivate stronger defenses and more reliable uncertainty monitoring in safety-critical applications. Negatively, the proposed attack formulation could inform misuse against deployed uncertainty-aware models. The paper frames the work as a diagnostic and robustness analysis rather than as a deployment tool.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained models, scraped datasets, or generative systems requiring controlled access. The experiments use standard public classification benchmarks and trained ensemble classifiers for research evaluation.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper uses standard benchmark datasets and architectures, including MNIST, CIFAR-10, CIFAR-100, ResNet56, WideResNet, and DenseNet, and cites the corresponding original works. No scraped or proprietary datasets are used.

Guidelines:

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

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, pretrained model, or benchmark asset. The codebase is planned for public release after cleanup and documentation, as noted in the open access to data and code item.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The work does not involve crowdsourcing, user studies, human-subject experiments, participant instructions, or compensation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The work does not involve human subjects, crowdsourcing, private user data, or interaction with participants; therefore IRB approval is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology, experiments, or scientific claims of this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.