EXPECTED ADVERSARIAL RISK: A DATA-CENTRIC VIEW OF ADVERSARIAL PURIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial purification has emerged as a promising defense strategy to mitigate adversarial attacks. Despite its effectiveness, the underlying mechanism remains unclear, making researchers resort to empirical algorithms rather than principled ones. In this paper, we propose a probabilistic, data-centric view of adversarial purification through the lens of Expected Adversarial Risk (EAR), which measures the probability that an example is assigned a wrong label across randomly sampled empirical datasets. We provide theoretical analysis for the effectiveness of adversarial purification, and attribute its success to alignment of purification directions with negative gradients of EAR. We further show that such alignment may not hold when severe class overlap or imbalance exist locally. Accordingly, we construct an EAR-inspired objective for purification, and demonstrate its effectiveness through experiments on synthetic and image datasets.

1 INTRODUCTION

The vulnerability of deep neural networks to adversarial perturbations remains a fundamental challenge, especially for classification tasks: small, often imperceptible input changes can drastically alter model predictions (Szegedy et al., 2014; Goodfellow et al., 2015). Essentially, this arises from certain disagreements between model predictions and ground truth labels. Classical L_p-norm attackers assume label invariance around nearby area, providing a proxy for searching of such disagreements. Though useful in practice, this view gives limited insight for the success of adversarial purification, which is an effective defense strategy against adversarial perturbations.

Adversarial purification is a promising test-time defense by preprocessing an input before classification, typically using generative models to drive the input towards regions supported by the natural data distribution (Yoon et al., 2021; Nie et al., 2022; Xiao et al., 2023). In general, adversarial purification does not guarantee recovery of clean data. Instead, a more acknowledged explanation for the success is the probability density improvement of input data Chen et al. (2024); Nan et al. (2025). However, it is still unclear why probability improvement is effective for defense and under what circumstances it may fail. This unresolved problem hinders researchers from seeking principled solutions during algorithm design.

In this paper, we aim to answer the above question regarding success of adversarial purification through theoretical analysis in a data-centric view. The core idea is, we attribute the adversarial vulnerability to the randomness during empirical data sampling from ground truth data distribution. In other words, if a dominant class in a local area unluckily samples less examples than another class, the predicted label would easily flip no matter how to choose the function space for a classification model, leading to an universal adversarial example. An intuition is, updating an input sample towards higher probability density area is likely to reduce adversarial vulnerability in the following way: the sample moves closer to dominant class center and farther to others, thus lowers label flip chances.

Specifically, we introduce *Expected Adversarial Risk (EAR)*, a data-induced risk $P_{\mathrm{adv}}(\mathbf{x})$ that measures how likely finite samples around $\mathbf{x}$ allow a competing class to overturn the locally dominant class under a local nonparametric decision rule. This reframes adversarial vulnerability as a local distributional instability: risk is high when weak local class margins can be overwhelmed by finite-sample fluctuations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We show through theoretical analysis that, EAR is mostly determined by the difference between conditional probability $p(\mathbf{x}|y)$ from dominant class and other classes. High risk areas appear near decision boundaries, where at least one class has close conditional probability to the dominant class. Density-guided purification reduces EAR under most cases due to first-order alignment with $-\nabla P_{\mathrm{adv}}(\mathbf{x})$. Meanwhile, failure cases exist when there are significant class overlap, class imbalance, or heterogeneous class geometry in the local area around $\mathbf{x}$. Finally, we demonstrate through simulation that EAR can be estimated empirically, using which as purification direction achieves better robustness compared with density-guided purification.

Our main contributions are summarized as follows:

- We propose *Expected Adversarial Risk (EAR)*, a new data-centric risk criterion that characterizes the emergence of adversarial vulnerability. We also provide theoretical analysis on how EAR is influenced by the sample size n, the density margin $\mathbf{\Delta}(\mathbf{x})$, and the local covariance structure (Theorem 1).
- We provide theoretic analysis about why adversarial purification works and when it does not. We prove that density-guided purification directions align well with negative EAR gradients in general, but may also fail under certain local distribution properties (Theorem 2). Motivated by this analysis, we construct an EAR-inspired local-margin objective for purification, showing that it outperforms density-guided objectives in the evaluated synthetic and real-world data settings.

2 PRELIMINARIES

Data distribution and local class geometry. Consider an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and label space $\mathcal{Y} = \{1, \ldots, K\}$. Let P denote the joint distribution of (X, Y), where $Y \sim \mathrm{Categorical}(\alpha_1, \ldots, \alpha_K)$ and $q_k(\mathbf{x})$ is the density of $X \mid Y = k$. The marginal density of X is $p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k q_k(\mathbf{x})$. We denote the population-optimal class at $\mathbf{x}$ by $k^\star(\mathbf{x}) = \arg\max_{k \in \mathcal{Y}} p(Y = k \mid X = \mathbf{x}) = \arg\max_{k \in \mathcal{Y}} \alpha_k q_k(\mathbf{x})$. For a small radius $r > 0$, let $V_d(r)$ denote the volume of the d-dimensional ball of radius r. The local probability mass within $B(\mathbf{x}, r)$ is $p_{k,r}(\mathbf{x}) = \mathbb{P}(X \in B(\mathbf{x}, r), Y = k) \approx \alpha_k q_k(\mathbf{x}) V_d(r)$ and the locally dominant class is $k_r^\star(\mathbf{x}) = \arg\max_k p_{k,r}(\mathbf{x})$. For sufficiently small r, the local approximation implies $k_r^\star(\mathbf{x}) = k^\star(\mathbf{x})$ whenever $k^\star(\mathbf{x})$ is unique. We refer to any $j \neq k^\star(\mathbf{x})$ as a *trap class*.

Generalized concept for adversarial vulnerability. Adversarial vulnerability is originally described as small perturbations on clean data can change model prediction Szegedy et al. (2014). Later this concept is extended to perceptible physical perturbations, such as adversarial patches Brown et al. (2017) and optical distortions Zhang et al. (2020). The underlying consensus is, the model prediction should not change as long as the data semantic keeps. For classification tasks, this means the classifier should report the same results if the ground truth label does not change, which essentially require equivalence between the model and ground truth.

We adopt this generalized concept for adversarial vulnerability in this paper, i.e., an example $\mathbf{x}$ is adversarial for a given classifier f if $f(\mathbf{x}) \neq k^\star(\mathbf{x})$. Under this view, whether an example is adversarial is intrinsic and irrelevant to attack methods such as FGSM (Goodfellow et al., 2015) and PGD (Madry et al., 2018). Clearly, adversarial vulnerability is now determined by empirical training data, function space of model, and training algorithm. We consider non-parametric models for theoretical analysis, so as to ignore the influence of function space and training algorithm. The only factor left is empirical training data, which is what we focus on in this paper.

Purification as density-guided movement. Adversarial purification is a test-time defense that preprocesses an adversarial input $\mathbf{x}_{\mathrm{adv}}$ through a generative model before feeding it into the classifier. The goal is to remove malicious perturbations while preserving the label semantics. In our analysis, we use this view to focus on two density-guided reference directions: the marginal score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ and the maximum conditional score $\nabla_{\mathbf{x}} \log M(\mathbf{x})$, where $M(\mathbf{x}) = \max_y p(\mathbf{x} \mid y)$. Since the logarithm preserves gradient directions up to a positive scalar, we use score and density-gradient directions interchangeably for alignment.

The marginal-score direction is instantiated by RDC (Chen et al., 2024), which performs likelihood maximization in the input space with respect to the unconditional likelihood, $\max_{\mathbf{x}} \log p(\mathbf{x})$. This is achieved by gradient ascent on $\mathbf{x}$ in the input space, pushing the adversarial example toward re-

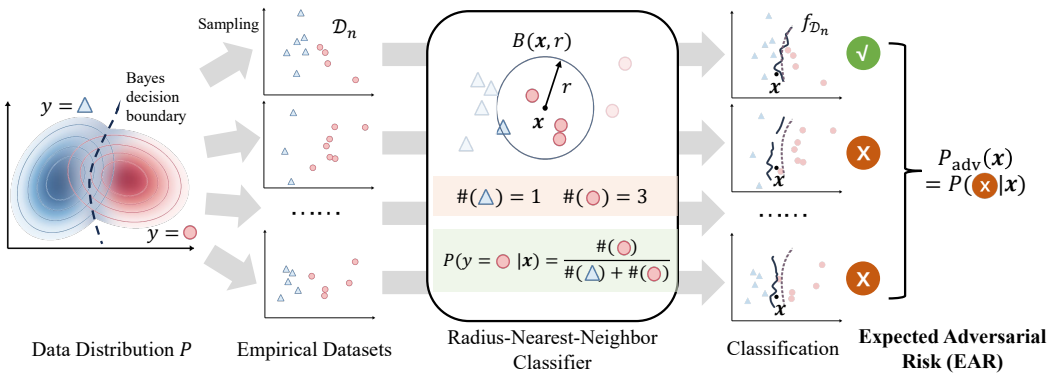


Figure 1: **EAR as sampling-induced local decision instability.** Training sets $\mathcal{D}_n \sim P^n$ are drawn from the underlying data distribution. We use the fixed-radius nearest-neighbor rule as a transparent nonparametric realization of the intuitive local-majority principle, making sampling-induced variation in local evidence explicit. EAR is the probability that finite-sample fluctuations allow the count of the trap class to match or exceed that of the locally dominant class.

gions of higher data density. However, COUP (Zhang et al., 2024) motivates class-aware guidance, since marginal density ascent may cause label shift under class overlap. To address this, COUP introduces classifier-confidence guidance into the diffusion purification process, motivating the maximum conditional-score direction $\nabla_{\mathbf{x}} \log M(\mathbf{x})$, which guides samples toward class-consistent high-density regions.

3 Expected Adversarial Risk from Finite-Sample Local Geometry

This section develops EAR as a data-centric measure of adversarial vulnerability. We first define EAR for a general model family, and then instantiate it with a specific nonparametric model family. This formulation enables explicit characterization of finite-sample risk, beginning with a binary example and extending to high-dimensional K-class settings.

3.1 Expected Adversarial Risk

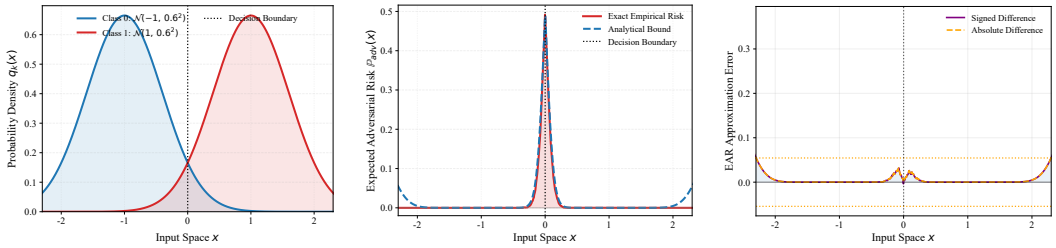
As motivated in Section 2, we characterize adversarial vulnerability through the disagreement between a classifier learned from a finite sample and the population-optimal decision. We first define this quantity for a general model family and then instantiate it with a radius-nearest-neighbor rule to expose its dependence on local class geometry.

Definition 3.1 (Expected Adversarial Risk)**.** *Let $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim P^n$ be an i.i.d. training set, and let $\mathcal{F}$ denote a model family equipped with a fixed learning rule. For each $\mathcal{D}_n$, define the empirical classifier as $f_{\mathcal{D}_n} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \neq y_i\}$, with a fixed tie-breaking rule. For a fixed query $\mathbf{x}$, the event $f_{\mathcal{D}_n}(\mathbf{x}) \neq k^\star(\mathbf{x})$ corresponds to $\mathbf{x}$ being adversarial under the classifier learned from $\mathcal{D}_n$. Define the probability of such event over the randomness of the sampled training set as Expected Adversarial Risk (EAR):*

$$\begin{aligned} P_{\text{adv}}(\mathbf{x}; \mathcal{F}, P, n) &= \mathbb{P}_{\mathcal{D}_n \sim P^n} \left(f_{\mathcal{D}_n}(\mathbf{x}) \neq k^\star(\mathbf{x}) \right) \\ &= \mathbb{E}_{\mathcal{D}_n \sim P^n} \left[\mathbb{I}\{ f_{\mathcal{D}_n}(\mathbf{x}) \neq k^\star(\mathbf{x}) \} \right]. \end{aligned} \tag{1}$$

EAR measures the probability that a classifier learned from a random finite dataset disagrees with the population-optimal label at a fixed query. Unlike attack success rate (ASR), which fixes a trained classifier and searches for a successful perturbation, EAR fixes the query and averages over training-set randomness. It therefore provides a data-centric view of finite-sample vulnerability within a given function space $\mathcal{F}$. For discussion of the relation between EAR and ASR, please refer to Appendix B.

To enable local analysis, we instantiate $\mathcal{F}$ with a fixed-radius nearest-neighbor reference rule. Define the empirical local class count $N_{r,n,k}(\mathbf{x}) = \sum_{i=1}^n \mathbb{I}\{\mathbf{x}_i \in B(\mathbf{x}, r),\ y_i = k\}$, and the correspond-



(a) Conditional distributions. (b) Gaussian EAR vs. Monte Carlo. (c) Approximation error.

Figure 2: Finite-sample EAR in a toy example of binary Gaussian mixture. The Gaussian approximation closely tracks the Monte Carlo flip probability, with deviations mainly in sparse tails.

ing non-parametric classifier $\hat{f}_r(\mathbf{x}) = \arg\max_{k\in\mathcal{Y}} N_{r,n,k}(\mathbf{x})$. Let $\mathcal{F}_r$ denote this radius-nearest-neighbor model family. We denote its EAR by

$$P_{\mathrm{adv}}(\mathbf{x}) := P_{\mathrm{adv}}(\mathbf{x}; \mathcal{F}_r, P, n) = \mathbb{P}_{\mathcal{D}_n\sim P^n}\left(\exists j \neq k^\star(\mathbf{x}) : N_{r,n,j}(\mathbf{x}) \geq N_{r,n,k^\star(\mathbf{x})}(\mathbf{x})\right). \quad (2)$$

where ties are conservatively counted as overturns. For brevity, with $\mathcal{F}_r$, P, and n fixed throughout the subsequent local analysis, we use the shorthand $P_{\mathrm{adv}}(\mathbf{x})$. Thus, $P_{\mathrm{adv}}(\mathbf{x})$ is the probability that finite-sample fluctuations allow a trap class to overturn the locally dominant class.

Figure 1 shows the calculation process of EAR, illustrating how finite-sample fluctuations induce adversarial risk under the fixed-radius nearest-neighbor rule on a toy example.

3.2 Warm-up: Binary Local Density Model

We first specialize the radius-nearest-neighbor formulation to the binary setting $\mathcal{Y} = \{0, 1\}$ with arbitrary class-conditional distributions $p(\mathbf{x} \mid y)$. The local counts $N_{r,n,1}(\mathbf{x})$ and $N_{r,n,0}(\mathbf{x})$ marginally follow binomial distributions with parameters n and local masses $p_{1,r}(\mathbf{x})$ and $p_{0,r}(\mathbf{x})$, respectively. Assume class 1 is locally dominant, i.e., $p_{1,r}(\mathbf{x}) > p_{0,r}(\mathbf{x})$. Define the difference between local counts as $D(\mathbf{x}) = N_{r,n,1}(\mathbf{x}) - N_{r,n,0}(\mathbf{x})$, the binary EAR is $P_{\mathrm{adv}}(\mathbf{x}) = \mathbb{P}(D(\mathbf{x}) \leq 0)$.

Proposition 1 (Binary EAR approximation)**.** *Given n finite samples, in the small-radius regime, let $\mu(\mathbf{x}) := p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})$, $\sigma^2(\mathbf{x}) := p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x}) - \mu^2(\mathbf{x})$, $n_{\mathrm{eff}}(\mathbf{x}) := n\big(p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x})\big)$. Then, there exists an absolute constant $C > 0$,*

$$\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) := \Phi\left(-\frac{\sqrt{n}\,\mu(\mathbf{x})}{\sigma(\mathbf{x})}\right), \qquad \left|P_{\mathrm{adv}}(\mathbf{x}) - \widetilde{P}_{\mathrm{adv}}(\mathbf{x})\right| \leq \frac{C}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}. \quad (3)$$

Here Φ is the standard normal cumulative distribution function.

Remark (Interpretation of EAR). Proposition 1 shows that the leading Gaussian EAR $\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ is governed by the local gap $p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})$, standardized by the local variance $\sigma^2 = p_{1,r} + p_{0,r} - (p_{1,r} - p_{0,r})^2$. For fixed local mass $q = p_{1,r} + p_{0,r}$, the variance is maximized when the two classes contribute equally, $p_{1,r} = p_{0,r} = q/2$, and decreases as one class becomes increasingly dominant, attaining its minimum $\sigma^2 = q(1-q)$ when the other class has zero local mass. The sample size n further scales the gap by $\sqrt{n}$, so the probability of a finite-sample reversal decreases as more samples become available. Meanwhile, the Gaussian approximation guarantee becomes weaker in sparse regions. The proof of Proposition 1 is provided in Appendix A.1.

Figure 2 visualizes the finite-sample instability predicted by EAR: near the decision boundary, small count fluctuations can make the trap class match or exceed the dominant class. This 1D example provides the core intuition for the high-dimensional analysis: adversarial risk can arise from finite local data rather than solely from model architecture or attack optimization.

3.3 GENERALIZATION TO HIGH-DIMENSIONAL K-CLASS MANIFOLDS

We now extend the binary intuition to high-dimensional K-class settings using the local quantities defined in Section 2. By Definition 3.1, EAR can be written as

$$P_{\mathrm{adv}}(\mathbf{x}) = \mathbb{P}\left(\bigcup_{j\neq k^\star}\{N_{r,n,j}(\mathbf{x}) \geq N_{r,n,k^\star}(\mathbf{x})\}\right), \tag{4}$$

This is the probability that at least one trap class has a local empirical count no smaller than that of the dominant class. Since exact evaluation involves a union of dependent count events, we analyze the differences $N_{r,n,k^\star}(\mathbf{x}) - N_{r,n,j}(\mathbf{x})$ for $j \neq k^\star$. The non-flip event is an intersection of half-spaces and is therefore convex, allowing us to apply the high-dimensional Berry–Esseen theorem for convex sets (Bentkus, 2003).

Theorem 1 (High-dimensional EAR approximation)**.** *Let* $\mathbf{D}_n(\mathbf{x}) = \big(N_{r,n,j}(\mathbf{x}) - N_{r,n,k^\star}(\mathbf{x})\big)_{j\neq k^\star}$, $p_{\mathrm{loc}}(\mathbf{x}) = \sum_{k=1}^K p_{k,r}(\mathbf{x})$*, and* $n_{\mathrm{eff}}(\mathbf{x}) = np_{\mathrm{loc}}(\mathbf{x})$*. Assume that, for each* $\mathbf{x}$*, there exists* $\lambda_0(\mathbf{x}) > 0$ *such that* $\lambda_{\min}\left(\frac{\mathbf{\Sigma}_V(\mathbf{x})}{p_{\mathrm{loc}}(\mathbf{x})}\right) \geq \lambda_0(\mathbf{x})$*. Let* $\mathcal{C}_{K-1}$ *be the class of measurable convex subsets of* $\mathbb{R}^{K-1}$*, and let* $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K-1})$*. Then*

$$\sup_{C\in\mathcal{C}_{K-1}}\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\big(\mathbf{D}_n(\mathbf{x}) + n\mathbf{\Delta}(\mathbf{x})\big) \in C\right) - \mathbb{P}(\mathbf{G}\in C)\right| = \mathcal{O}\left(\frac{K^{3/2}(K-1)^{1/4}}{\lambda_0(\mathbf{x})^{3/2}\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right) \tag{5}$$

Here $\Delta_j(\mathbf{x}) = p_{k^\star,r}(\mathbf{x}) - p_{j,r}(\mathbf{x})$*,, and*

$$\mathbf{\Sigma}_V(\mathbf{x}) = \mathrm{diag}(\mathbf{p}_{-k^\star,r}) - \mathbf{p}_{-k^\star,r}\mathbf{p}_{-k^\star,r}^\top + p_{k^\star,r}(\mathbf{x})\big(1-p_{k^\star,r}(\mathbf{x})\big)\mathbf{1}\mathbf{1}^\top + p_{k^\star,r}(\mathbf{x})\left(\mathbf{p}_{-k^\star,r}\mathbf{1}^\top + \mathbf{1}\mathbf{p}_{-k^\star,r}^\top\right),$$

where $\mathbf{p}_{-k^\star,r} = \big(p_{j,r}(\mathbf{x})\big)_{j\neq k^\star}$.

Taking C to be the whitened non-flip lower orthant in the multivariate Berry–Esseen bound above gives

$$\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) := 1 - \Phi_{K-1}\big(\sqrt{n}\,\mathbf{\Delta}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\big). \tag{6}$$

where $\Phi_{K-1}(\cdot; \mathbf{\Sigma}_V)$ denotes the CDF of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_V)$. In the subsequent gradient analysis, $\widetilde{P}_{\mathrm{adv}}$ denotes this leading Gaussian approximation.

The Gaussian approximation shows that EAR is large when the dominant and competing class masses are close, while its approximation error is controlled by the effective local sample size $n_{\mathrm{eff}}(\mathbf{x})$. Thus, adversarial vulnerability can arise from local data ambiguity rather than from a particular classifier architecture. For $K = 2$, the result reduces to Proposition 1; see Appendix A.4. The proof is provided in Appendix A.3.

Empirical validation. On the real-world CIFAR-10 dataset, the Gaussian EAR approximation closely matches the empirical estimate; see Appendix A.5. Moreover, the radius-nearest-neighbor rule provides a tractable reference model for analyzing EAR, while EAR itself is defined for a general model family. The Gaussian EAR approximation implied by Theorem 1 is further compared with the empirical flip risk of MLPs trained on the same 2D Gaussian mixture with $K = 3$. The two exhibit similar spatial risk patterns, with high-risk regions concentrated near class boundaries. This agreement suggests that the risk structure identified by the local nonparametric analysis is not specific to the reference rule. The comparison and additional configurations are reported in Appendix A.6.

4 ADVERSARIAL PURIFICATION THROUGH THE LENS OF EAR GRADIENTS

We now analyze test-time purification through the spatial gradient of EAR. Since EAR quantifies the adversarial event defined in Section 2, its negative gradient $-\nabla\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ gives a local risk-reduction direction. We compare this direction with the marginal score $\nabla \log p(\mathbf{x})$ and maximum conditional score $\nabla \log M(\mathbf{x})$, yielding a first-order alignment criterion for density-guided purification.

4.1 WARM-UP: 1D GRADIENT ALIGNMENT

To trace the spatial direction induced by purification, we instantiate the local data geometry with a symmetric Gaussian mixture under equal class priors: $p_+(x) = \mathcal{N}(x; 1, \sigma^2)$, $p_-(x) = \mathcal{N}(x; -1, \sigma^2)$, $p(x) = \frac{1}{2}p_+(x) + \frac{1}{2}p_-(x)$. We distinguish the separated regime $\sigma^2 < 1$ from the overlapping regime $\sigma^2 > 1$. In the small-radius limit, the equal class priors give $p_{1,r}(x) \approx rp_+(x)$, $p_{0,r}(x) \approx rp_-(x)$. Let $D = p_+ - p_-$ and $S = p_+ + p_-$. Proposition 1 then gives $\widetilde{P}_{\mathrm{adv}}(x) \approx \Phi\left(-\sqrt{nr}\,\frac{D(x)}{\sqrt{S(x)}}\right)$. Since the derivative of Φ is positive, the spatial gradient satisfies, up to a strictly positive scalar factor, $\nabla\widetilde{P}_{\mathrm{adv}}(x) \propto -\left(2S(x)\nabla D(x) - D(x)\nabla S(x)\right)$.

For $x > 0$, the dominant class is $Y = 1$, so $M(x) = p_+(x)$. Comparing the signs of the marginal, conditional, and EAR directions yields the following alignment criterion.

Proposition 2 (Directional alignment criterion in 1D)**.** *Under the symmetric Gaussian mixture with equal class priors: If $\sigma^2 < 1$ (well-separated components), there exists a boundary-adjacent interval $x \in (0, x_p^*)$ such that $-\nabla\widetilde{P}_{\mathrm{adv}}(x) > 0,\ \nabla \log p(x) > 0,\ \nabla \log M(x) > 0$. If $\sigma^2 > 1$ (overlapping components), for sufficiently small $x > 0$ we have $-\nabla\widetilde{P}_{\mathrm{adv}}(x) > 0,\ \nabla \log M(x) > 0,\ \nabla \log p(x) < 0$.*

In the separated regime ($\sigma^2 < 1$), Proposition 2 shows that both marginal and conditional scores align with $-\nabla\widetilde{P}_{\mathrm{adv}}(x)$, explaining why either guidance reduces EAR in regular local geometry. Under overlap ($\sigma^2 > 1$), the marginal score becomes anti-aligned, whereas the conditional score remains aligned, showing the benefit of class-aware guidance. Detailed proofs are provided in Appendix C.1; we next extend the analysis to high-dimensional K-class settings.

4.2 GENERALIZATION TO HIGH-DIMENSIONAL K-CLASS RISK GRADIENTS

EAR gradient and T_1 dominance. The 1D gradient decomposition extends naturally to high-dimensional K-class settings. Applying the matrix chain rule to $\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$, its gradient decomposes into two structural components:

$$\nabla\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \approx \underbrace{-\frac{\partial\Phi_{K-1}}{\partial\mathbf{a}}\frac{\partial\mathbf{a}}{\partial\mathbf{x}}}_{T_1\text{: margin term}} \underbrace{-\operatorname{tr}\left(\frac{\partial\Phi_{K-1}}{\partial\boldsymbol{\Sigma}_V}\frac{\partial\boldsymbol{\Sigma}_V}{\partial\mathbf{x}}\right)}_{T_2\text{: covariance term}}, \tag{7}$$

where $\mathbf{a} = \sqrt{n}\boldsymbol{\Delta}(\mathbf{x})$. For each competing class j, $\partial\Phi_{K-1}/\partial a_j > 0$ measures how sensitive the Gaussian approximation of the correct-selection probability is to the standardized margin $a_j = \sqrt{n}\,\Delta_j$. Define the sensitivity-weighted margin direction as $\mathbf{h}(\mathbf{x}) = \sum_{j\neq k^\star} \frac{\partial\Phi_{K-1}}{\partial a_j}\nabla\Delta_j(\mathbf{x})$. Then $T_1 = -\sqrt{n}\,\mathbf{h}$ and $-\nabla\widetilde{P}_{\mathrm{adv}} \approx \sqrt{n}\,\mathbf{h} - T_2$.

As n increases, points with fixed positive class margins become stable and their EAR approaches zero. Nontrivial EAR therefore concentrates in a shrinking neighborhood of the decision boundary, where the closest class-mass gap is of order $n^{-1/2}$. In this regime, if $\mathbf{h}$ is non-degenerate while the covariance term T_2 remains bounded, then $\frac{\|T_2(\mathbf{x}_n,n)\|_2}{\|T_1(\mathbf{x}_n,n)\|_2} = O(n^{-1/2})$. Thus, the EAR descent direction in this boundary-local region is asymptotically governed by the sensitivity-weighted margin direction $\mathbf{h}$; see Appendix C.2, Lemma 6.

4.3 CHARACTERIZING DENSITY-GUIDED PURIFICATION THROUGH EAR

The preceding analysis identifies $-\nabla\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ as the local EAR-descent direction. We use this direction to characterize when marginal and maximum-conditional density guidance can reduce EAR, and how local data geometry can break this correspondence. To obtain interpretable conditions for density-guided purification, we approximate the local data density $p(\mathbf{x})$ with a Gaussian mixture (Silverman, 2018): $p(\mathbf{x}) \approx \sum_i \pi_i q_i(\mathbf{x})$, where q_i approximates the local class-conditional density of class i, and π_i is its effective mixture weight. Within this approximation, $k^\star = \arg\max_i \pi_i q_i(\mathbf{x})$ is the posterior-dominant class, while $k_M = \arg\max_i q_i(\mathbf{x})$ indexes the component selected by $M(\mathbf{x})$. Because EAR descent is defined relative to $k^\star$, we consider fixed-$k^\star$ neighborhoods; Appendix C.3 gives the corresponding local label-preservation result.

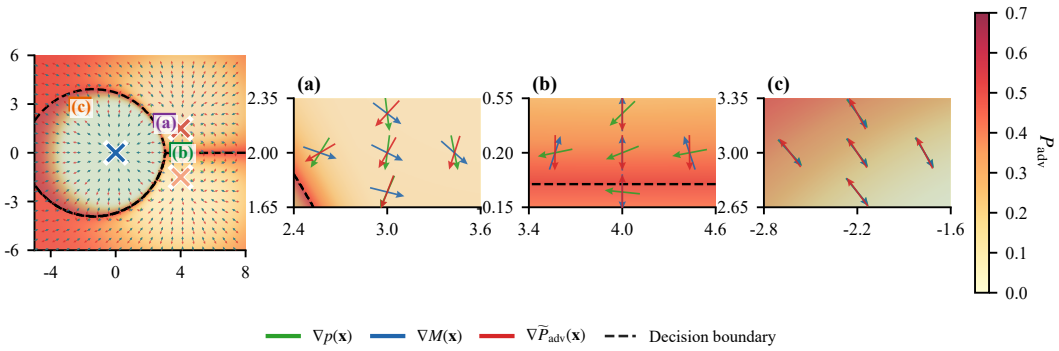


Figure 3: **Analytic alignment of density guidance and EAR descent.** Panels (a)(b)(c) shows different alignment patterns in certain regions. EAR descent points opposite to the red arrows.

Theorem 2 (Local alignment conditions for adversarial purification)**.** *Let* $\mathbf{x}$ *lie in a neighborhood where* $k^\star$ *is unique and locally fixed. Under* T_1 *dominance and the local class-conditional dominance condition in Appendix C.5, each condition below is sufficient for the corresponding direction* $\mathbf{g} \in \{\nabla p(\mathbf{x}), \nabla M(\mathbf{x})\}$ *to satisfy* $\langle \mathbf{g}, -\nabla \widetilde{P}_{\text{adv}}(\mathbf{x}) \rangle > 0$ *and hence align with the leading-order EAR-descent direction.*

Marginal density guidance $(\mathbf{g} = \nabla p(\mathbf{x}))$ *is aligned with EAR descent when*

$$\pi_{k^\star}^2 q_{k^\star}^2(\mathbf{x}) \|s_{k^\star}(\mathbf{x})\|^2 > \sum_{j \neq k^\star} \tilde{w}_j \pi_j^2 q_j^2(\mathbf{x}) \|s_j(\mathbf{x})\|^2. \tag{8}$$

Maximum conditional density guidance $(\mathbf{g} = \nabla M(\mathbf{x}))$*. Since* $M(\mathbf{x}) = q_{k_M}(\mathbf{x})$*, its local gradient is* $\nabla M(\mathbf{x}) = q_{k_M}(\mathbf{x}) s_{k_M}(\mathbf{x})$*. This direction aligns with EAR descent when*

$$\pi_{k^\star} q_{k_M}(\mathbf{x}) q_{k^\star}(\mathbf{x}) \langle s_{k_M}(\mathbf{x}), s_{k^\star}(\mathbf{x}) \rangle > \sum_{j \neq k^\star} \tilde{w}_j \pi_j q_{k_M}(\mathbf{x}) q_j(\mathbf{x}) \langle s_{k_M}(\mathbf{x}), s_j(\mathbf{x}) \rangle. \tag{9}$$

With equal priors, $k_M = k^\star$*, and Eq. 9 reduces to*

$$\pi_{k^\star} q_{k^\star}^2(\mathbf{x}) \|s_{k^\star}(\mathbf{x})\|^2 > \sum_{j \neq k^\star} \tilde{w}_j \pi_j q_{k^\star}(\mathbf{x}) q_j(\mathbf{x}) \langle s_{k^\star}(\mathbf{x}), s_j(\mathbf{x}) \rangle. \tag{10}$$

Here $q_j(\mathbf{x}) = \phi_d(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$*,* $s_j(\mathbf{x}) = \nabla \log q_j(\mathbf{x}) = -\boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)$*, and* $\tilde{w}_j = \frac{\partial \Phi_{K-1}/\partial a_j}{\sum_{\ell \neq k^\star} \partial \Phi_{K-1}/\partial a_\ell}$ *is the normalized margin-sensitivity weight inherited from* $\mathbf{h}$*, with* $\sum_{j \neq k^\star} \tilde{w}_j = 1$*.*

The proof is provided in Appendix C.5. Theorem 2 gives a sufficient condition for density guidance to reduce the Gaussian EAR approximation: its directional increase in the dominant-class mass must exceed the sensitivity-weighted changes in the competing-class masses. The two conditions differ because ∇p combines all class components, whereas ∇M follows a single selected component.

From density guidance to EAR-guided purification. The alignment analysis explains why existing density-guided purification works in many cases. Under regular local geometry, the posterior-dominant class controls the marginal-density gradient $\nabla p(\mathbf{x})$, while $k_M = k^\star$ makes $\nabla M(\mathbf{x})$ ascend the same class-conditional density. Both directions then increase the class margins that control EAR and approximately follow EAR descent.

These density directions are proxies for EAR descent. Marginal density guidance combines contributions from all classes and can be distorted by class overlap, prior imbalance, or heterogeneous local class geometry such as variance imbalance. Maximum conditional guidance reduces this mixture interference, but follows only one selected class and can miss influence from other competing classes. The conditions in Theorem 2 characterize when this correspondence can break.

The analytic field in Figure 3 and the KDE estimates in Appendix C.6 support this interpretation. Globally, 81.9% of $\nabla \hat{p}$ directions and 90.7% of $\nabla \hat{M}$ directions form acute angles with empirical EAR descent, showing that both density-guided directions are usually effective. These ratios decrease to 61.6% and 65.9%, respectively, over the 15% of grid points with highest empirical EAR.

Table 1: Synthetic-data conditional ASR and mean EAR, reported as mean ± standard deviation over four seeds. EAR_B and EAR_A denote EAR before and after the attack, respectively.

Metric	Transfer Attack						Adaptive Attack		
	$p(\mathbf{x})$	$M(\mathbf{x})$	$S_{\mathrm{EAR}}(\mathbf{x})$	Distilled $p(\mathbf{x})$	Distilled $M(\mathbf{x})$	Distilled $S_{\mathrm{EAR}}(\mathbf{x})$	Distilled $p(\mathbf{x})$	Distilled $M(\mathbf{x})$	Distilled $S_{\mathrm{EAR}}(\mathbf{x})$
ASR	39.65 ± 3.52	7.61 ± 0.96	$\mathbf{2.01 \pm 0.55}$	39.74 ± 3.55	9.21 ± 0.67	$\mathbf{6.43 \pm 0.60}$	41.26 ± 3.23	9.55 ± 0.30	$\mathbf{7.39 \pm 1.06}$
EAR_B	46.82 ± 1.46	46.82 ± 1.46	46.82 ± 1.46	46.82 ± 1.46	46.82 ± 1.46	46.82 ± 1.46	46.17 ± 1.41	46.49 ± 1.45	46.45 ± 1.82
EAR_A	38.11 ± 1.95	31.49 ± 1.53	$\mathbf{31.08 \pm 1.53}$	38.09 ± 1.92	32.52 ± 1.58	$\mathbf{32.14 \pm 1.61}$	37.37 ± 1.96	32.05 ± 1.57	$\mathbf{31.64 \pm 2.04}$

5 EXPERIMENTS

We perform experiments on synthetic and image dataset, to show that EAR-guided purification is a better choice. Note that our main aim is to illustrate the advantage of EAR for purification over $p(\mathbf{x})$ or $M(\mathbf{x})$, rather than proposing an rigorously validated, practically usable purification algorithm.

To instantiate EAR-guided purification, we first construct a tractable objective for optimization. Directly optimizing the $\widetilde{P}_{\mathrm{adv}}(x)$ is computationally costly, as it requires estimating the local covariance matrix $\mathbf{\Sigma}_V$ and differentiating through a multivariate Gaussian CDF. Our solution include two parts: first, only considering the gradient direction and ignore irrelative constant factors; second, replacing the less important covariance matrix and weights with constants. The detailed process is provided in Appendix D.2, together with performance comparisons among each approximation steps. The final surrogate purification objective retains the core idea of maximizing margins w.r.t. all non-dominant classes in EAR minimization:

$$S_{\mathrm{EAR}}(\mathbf{x}) := \frac{1}{K-1} \sum_{j \neq \widehat{k}^\star} \left[p_{\widehat{k}^\star}(\mathbf{x}) - p_j(\mathbf{x}) \right], \tag{11}$$

where $\widehat{k}^\star = \widehat{k}^\star(\mathbf{x})$ is an approximation of $k^\star(\mathbf{x})$ using models. EAR-guided purification is performed through gradient ascending w.r.t S_{EAR}.

5.1 SIMULATIONS ON SYNTHETIC MANIFOLDS

Setup. We use a 2D synthetic Gaussian-mixture model with five classes and four components per class as data distribution. Full settings are provided in Appendix D.1. Each purification objective defines a 10-step iterative purifier g. To enable adaptive attack, we additionally train distilled one-step purifiers as approximations $g_\theta^{\mathrm{dist}}(\mathbf{x}) = \mathbf{x} + R_\theta(\mathbf{x})$, where the MLP R_θ predicts $g(\mathbf{x}) - \mathbf{x}$. Under transfer attack, we evaluate both purifiers to assess the effect of distillation; adaptive attack evaluation uses g_θ^{dist} for efficient differentiation. Within each attack setting, we compare EAR guidance against $p(\mathbf{x})$- and $M(\mathbf{x})$-based guidance.

For evaluation, both attacks are implemented using PGD (Madry et al., 2018). Transfer adversarial examples are generated against f, whereas adaptive examples are generated against $f \circ g_\theta^{\mathrm{dist}}$. To better illustrate the purification effect, we filter 2,000 clean samples for evaluation, which are correctly classified by f but successfully attacked by PGD over f. Attack, filtering, distillation, and radius-selection details are provided in Appendix D.1.

Main Result. Table 1 shows that EAR-guided purification is effective under both attack settings. Under transfer PGD, it achieves the lowest post-purification EAR and ASR both before and after distillation, showing that distillation preserves its relative advantage. Under adaptive PGD, the distilled EAR-guided purifier also outperforms the two density-guided purifiers. Within each evaluation setting, objectives with lower post-purification EAR also yield lower conditional ASR, supporting the expected relationship between EAR reduction and purification effectiveness.

5.2 EVALUATION ON REAL-WORLD IMAGES

Having verified EAR-guided purification on synthetic manifolds, we next evaluate whether it remains effective on real images. Since there are currently no reliable adaptive attack methods for diffusion-based purification, we evaluate the first 512 CIFAR-10 test images using transferred APGD-CE examples with an ℓ_∞ budget of $8/255$. The examples are generated against a pretrained

Table 2: CIFAR-10 accuracy on clean inputs and transferred APGD-CE adversarial examples.

Input	No purification	$p(\mathbf{x})$	$M(\mathbf{x})$	EAR-guided
Clean	**96.68**	83.59	89.45	94.53
APGD-CE (8/255)	0.00	73.83	78.32	**91.99**

WideResNet-70-16-dropout, which is also used as the down-stream classifier. A conditional diffusion classifier supplies the fixed $\widehat{k}(\mathbf{x})$ for $M(\mathbf{x})$ and EAR guidance; full implementation details are provided in Appendix D.3

As shown in Table 2, EAR-guided purification achieves 91.99% accuracy on transferred APGD-CE examples, exceeding $p(\mathbf{x})$- and $M(\mathbf{x})$-guided purification by 18.16 and 13.67 percentage points, respectively. It also retains 94.53% clean accuracy, only 2.15 points below the model without purification. These results show that the advantage of EAR guidance extends from controlled synthetic distributions to natural images under the evaluated transfer attack.

6 Related work

Adversarial purification. Adversarial purification preprocesses adversarial inputs before classification, typically by moving them toward regions supported by the natural data distribution (Samangouei et al., 2018; Yoon et al., 2021; Nie et al., 2022). A line of label-free methods relies on unconditional likelihood,diffusion denoising, or other label-free guidance signals (Chen et al., 2024; Nie et al., 2022; Xiao et al., 2023; Bai et al., 2024). These methods demonstrate that density or score dynamics can improve robustness, but they do not characterize when the induced purification direction reduces intrinsic adversarial risk. Class-aware purification methods introduce classifier- or label-dependent guidance to mitigate label shift under class overlap (Zhang et al., 2024; Lin et al., 2025). Our analysis provides a data-centric explanation for this distinction: purification succeeds when its direction aligns with the EAR descent direction, and can fail when overlap or imbalance breaks this alignment.

Theoretical explanations of purification. Existing theory mainly explains purification through specific generative mechanisms. Diffusion-based analyses study distributional contraction or reverse-time sampling toward high-density regions (Nie et al., 2022; Xiao et al., 2023); other works analyze contrastive guidance, manifold reconstruction, or diffusion-bridge purification (Bai et al., 2024; Yang et al., 2024; Li et al., 2025). These results clarify why particular purification procedures can recover clean samples, but the guarantees are tied to the corresponding generative dynamics. In contrast, EAR gives a mechanism-agnostic risk criterion: it compares any purification direction with a local intrinsic-risk descent field, thereby identifying when density-guided recovery is aligned with robustness improvement.

Theoretical views of adversarial vulnerability. Prior theory explains adversarial vulnerability through robust optimization, robustness–accuracy trade–offs, data concentration, boundary geometry, data sparsity, and non-robust features (Goodfellow et al., 2015; Madry et al., 2018; Tsipras et al.; Pal et al., 2023; Tanay & Griffin, 2016; Paknezhad et al., 2022; Ilyas et al., 2019; Li & Li, 2025). These works explain why adversarial vulnerability arises from classifiers, features, or distributional geometry. A more local line of work studies adversarial evaluation and the geometric structure of adversarial regions (Uesato et al., 2018; Ma et al., 2018). These works are closest in spirit to EAR, but focus on evaluation or adversarial-region characterization rather than purification.

7 Conclusion

We provide a probabilistic, data-centric explanation of when adversarial purification works through Expected Adversarial Risk (EAR), a local finite-sample decision-flip probability determined by class masses, density margins, and sample size. Using the negative EAR gradient as the intrinsic risk-reduction direction, we show that density-guided updates reduce risk when aligned with this direction, while local class overlap, mixture-weight imbalance, or variance imbalance may break the alignment. Experiments on synthetic data and CIFAR-10 demonstrate the superior performance of EAR-guided purification over density-guided objectives.

AI USE STATEMENT

Generative AI tools were used only to improve the clarity and concision of the manuscript. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content

REFERENCES

Mingyuan Bai, Wei Huang, Tenghui Li, Andong Wang, Junbin Gao, Cesar F Caiafa, and Qibin Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. In *International Conference on Machine Learning (ICML)*, pp. 2375–2391, 2024.

V. Bentkus. On the dependence of the berry–esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning (ICML)*, pp. 6643–6665, 2024.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Binghui Li and Yuanzhi Li. Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 73687–73722, 2025.

Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yingzhe He, Jie Shi, and Xiaolin Hu. Adbm: Adversarial diffusion bridge model for reliable adversarial purification. In *International Conference on Learning Representations (ICLR)*, 2025.

Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Adversarial guided diffusion models for adversarial purification. *Neural Networks*, 191:107705, 2025.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi N. R. Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)*, 2018.

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

Junjie Nan, Jianing Li, Wei Chen, Mingkun Zhang, and Xueqi Cheng. Nappure: Adversarial purification for robust image classification under non-additive perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2260–2269, 2025.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Mahsa Paknezhad, Cuong Phuc Ngo, Amadeus Aristo Winarto, Alistair Cheong, Chuen Yang Beh, Jiayang Wu, and Hwee Kuan Lee. Explaining adversarial vulnerability with a data sparsity hypothesis. *Neurocomputing*, 495:178–193, 2022.

Ambar Pal, Jeremias Sulam, and René Vidal. Adversarial examples might be avoidable: The role of data concentration in adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*.

Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning (ICML)*, pp. 5025–5034, 2018.

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models for adversarial robustness. In *International Conference on Learning Representations, (ICLR)*, 2023.

Zhaoyuan Yang, Zhiwei Xu, Jing Zhang, Richard I. Hartley, and Peter H. Tu. Adversarial purification with the manifold hypothesis. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 16379–16387, 2024.

Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pp. 12062–12072, 2021.

Mingkun Zhang, Jianing Li, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Classifier guidance enhances diffusion-based adversarial purification by preserving predictive information. In *European Conference on Artificial Intelligence (ECAI)*, pp. 2234–2241. IOS Press, 2024.

Y. H. Zhang, W. J. Ruan, F. Wang, and X. W. Huang. Generalizing universal adversarial attacks beyond additive perturbations. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 1412–1417. IEEE, 2020.

A Proofs of section 3

A.1 Proof of proposition 1: Binary EAR approximation

Notation

Throughout this subsection, we fix a test point $\mathbf{x}$, a radius $r > 0$, and the local neighborhood $B_r(\mathbf{x}) := B(\mathbf{x}, r)$ with volume V_r. The symbols $p_+(\mathbf{x})$ and $p_-(\mathbf{x})$ denote the two global Gaussian class-conditional densities. Under the balanced two-Gaussian model, the global marginal density is $p(\mathbf{x}) = \frac{1}{2}p_+(\mathbf{x}) + \frac{1}{2}p_-(\mathbf{x})$.

The symbols $p_{1,r}(\mathbf{x})$ and $p_{0,r}(\mathbf{x})$ denote the local probability masses inside $B_r(\mathbf{x})$, namely $p_{1,r}(\mathbf{x}) := \mathbb{P}\big(\mathbf{X} \in B_r(\mathbf{x}), Y = 1\big)$, $p_{0,r}(\mathbf{x}) := \mathbb{P}\big(\mathbf{X} \in B_r(\mathbf{x}), Y = 0\big)$. Their total local probability mass is

$$p_{\mathrm{loc}}(\mathbf{x}) := p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x}) = \mathbb{P}\big(\mathbf{X} \in B_r(\mathbf{x})\big) = \int_{B_r(\mathbf{x})} p(\mathbf{u})\, d\mathbf{u}.$$

Accordingly,

$$n_{\mathrm{eff}}(\mathbf{x}) = np_{\mathrm{loc}}(\mathbf{x}) = n \int_{B_r(\mathbf{x})} p(\mathbf{u})\, d\mathbf{u}. \tag{12}$$

To relate this quantity to the pointwise density, assume that p is $L_{\mathbf{x}}$-Lipschitz on $B_r(\mathbf{x})$. Then

$$|p_{\text{loc}}(\mathbf{x}) - V_r p(\mathbf{x})| \leq \int_{B_r(\mathbf{x})} |p(\mathbf{u}) - p(\mathbf{x})| \, d\mathbf{u} \leq L_{\mathbf{x}} r V_r.$$

Consequently,

$$|n_{\text{eff}}(\mathbf{x}) - n V_r p(\mathbf{x})| \leq n L_{\mathbf{x}} r V_r. \tag{13}$$

For one-dimensional inputs, $B_r(x) = [x - r, x + r]$ and $V_r = 2r$. In this case,

$$|p_{\text{loc}}(x) - 2rp(x)| \leq L_x \int_{x-r}^{x+r} |u - x| \, du = L_x r^2, \tag{14}$$

and therefore

$$|n_{\text{eff}}(x) - 2nrp(x)| \leq n L_x r^2. \tag{15}$$

PROOF OF PROPOSITION 1

Proof. For each training sample, define

$$Z_i(\mathbf{x}) = \mathbb{I}\{\mathbf{X}_i \in B_r(\mathbf{x}), Y_i = 1\} - \mathbb{I}\{\mathbf{X}_i \in B_r(\mathbf{x}), Y_i = 0\}.$$

Then

$$D(\mathbf{x}) = \sum_{i=1}^{n} Z_i(\mathbf{x}),$$

with

$$\mathbb{P}(Z_i = 1) = p_{1,r}(\mathbf{x}), \ \mathbb{P}(Z_i = -1) = p_{0,r}(\mathbf{x}),$$

and

$$\mathbb{P}(Z_i = 0) = 1 - p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x}).$$

Let

$$\mu(\mathbf{x}) = p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x}), \ q(\mathbf{x}) = p_{\text{loc}}(\mathbf{x}) = p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x}).$$

Then

$$\mathbb{E}[Z_i(\mathbf{x})] = \mu(\mathbf{x}), \ \mathbb{E}[Z_i^2(\mathbf{x})] = q(\mathbf{x}),$$

and hence

$$\sigma^2(\mathbf{x}) := \text{Var}(Z_i(\mathbf{x})) = q(\mathbf{x}) - \mu^2(\mathbf{x}). \tag{16}$$

Let

$$\rho(\mathbf{x}) = \mathbb{E}\left|Z_i(\mathbf{x}) - \mu(\mathbf{x})\right|^3.$$

Since $Z_i(\mathbf{x}) \in \{-1, 0, 1\}$,

$$\rho(\mathbf{x}) = p_{1,r}(\mathbf{x})|1 - \mu(\mathbf{x})|^3 + p_{0,r}(\mathbf{x})|-1 - \mu(\mathbf{x})|^3 + \big(1 - q(\mathbf{x})\big)|\mu(\mathbf{x})|^3.$$

Using $|\mu(\mathbf{x})| \leq q(\mathbf{x}) \leq 1$, we have

$$|1 - \mu(\mathbf{x})| \leq 2, \qquad |-1 - \mu(\mathbf{x})| \leq 2.$$

Therefore,

$$\begin{aligned} \rho(\mathbf{x}) &\leq 8\big(p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x})\big) + \big(1 - q(\mathbf{x})\big)|\mu(\mathbf{x})|^3 \\ &\leq 8q(\mathbf{x}) + q^3(\mathbf{x}) \\ &\leq 9q(\mathbf{x}), \end{aligned} \tag{17}$$

where the last inequality follows from $q(\mathbf{x}) \leq 1$.

Moreover,

$$\sigma^2(\mathbf{x}) = q(\mathbf{x}) - \mu^2(\mathbf{x}) \geq q(\mathbf{x}) - q^2(\mathbf{x}) = q(\mathbf{x})\big(1 - q(\mathbf{x})\big).$$

In the small-radius regime, $q(\mathbf{x}) \leq 1/2$ for sufficiently small r, so

$$\sigma^2(\mathbf{x}) \geq \frac{1}{2} q(\mathbf{x}). \tag{18}$$

Therefore,

$$\frac{\rho(\mathbf{x})}{\sigma^3(\mathbf{x})} \le \frac{9 \cdot 2^{3/2}}{\sqrt{q(\mathbf{x})}}. \tag{19}$$

By the classical Berry–Esseen theorem,

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\frac{D(\mathbf{x}) - n\mu(\mathbf{x})}{\sqrt{n}\sigma(\mathbf{x})} \le z \right) - \Phi(z) \right| \le \frac{C_{\mathrm{BE}}\rho(\mathbf{x})}{\sigma^3(\mathbf{x})\sqrt{n}} \le \frac{9 \cdot 2^{3/2} C_{\mathrm{BE}}}{\sqrt{nq(\mathbf{x})}} = \frac{9 \cdot 2^{3/2} C_{\mathrm{BE}}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}.$$

This establishes the one-dimensional Berry–Esseen bound underlying the Gaussian approximation in Proposition 1. Equivalently, substituting $\mu(\mathbf{x}) = p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})$ and $\sigma^2(\mathbf{x}) = p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x}) - \big(p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})\big)^2$ into Eq. 20 gives

$$\sup_{z\in\mathbb{R}} \left| \mathbb{P}\left(\frac{D(\mathbf{x}) - n\big(p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})\big)}{\sqrt{n\left[p_{1,r}(\mathbf{x}) + p_{0,r}(\mathbf{x}) - \big(p_{1,r}(\mathbf{x}) - p_{0,r}(\mathbf{x})\big)^2\right]}} \le z \right) - \Phi(z) \right| = \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}} \right). \tag{20}$$

Taking

$$z := -\frac{\sqrt{n}\mu(\mathbf{x})}{\sigma(\mathbf{x})}$$

in Eq. 20 and using

$$P_{\mathrm{adv}}(\mathbf{x}) := \mathbb{P}\big(D(\mathbf{x}) \le 0\big)$$

gives

$$\left| P_{\mathrm{adv}}(\mathbf{x}) - \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \right| \le \frac{9 \cdot 2^{3/2} C_{\mathrm{BE}}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}},$$

where

$$\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) := \Phi\left(-\frac{\sqrt{n}\mu(\mathbf{x})}{\sigma(\mathbf{x})} \right).$$

Thus, Proposition 1 holds with $C = 9 \cdot 2^{3/2} C_{\mathrm{BE}}$. □

A.2 Error term analysis: breakdown of the approximation at the tails

Define the finite-sample Berry–Esseen error by $\Delta_{\mathrm{BE}}(\mathbf{x}) := \sup_{z\in\mathbb{R}} \left| \mathbb{P}\big(S_n(\mathbf{x}) \le z\big) - \Phi(z) \right|$. By Eq. 23,

$$\Delta_{\mathrm{BE}}(\mathbf{x}) \le \frac{9 \cdot 2^{3/2} C_{\mathrm{BE}}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}. \tag{21}$$

For one-dimensional inputs, Eq. 15 gives $n_{\mathrm{eff}}(x) \ge n\big(2rp(x) - L_x r^2\big)$. Whenever $2rp(x) > L_x r^2$, it follows that

$$\Delta_{\mathrm{BE}}(x) \le \frac{9 \cdot 2^{3/2} C_{\mathrm{BE}}}{\sqrt{n\big(2rp(x) - L_x r^2\big)}}. \tag{22}$$

Thus, for a fixed finite sample size n, the approximation guarantee becomes weaker in sparse regions because fewer observations are expected to fall inside the local neighborhood. If the displayed upper bound exceeds one, it becomes non-informative; the actual probability approximation error remains bounded by one.

A.3 Proof of theorem 1: analytic bound of local adversarial probability

In this section, we prove Theorem 1. We first express the local counts as sums of independent indicator vectors, then establish moment and covariance lemmas that allow the application of the high-dimensional Berry–Esseen theorem over convex sets.

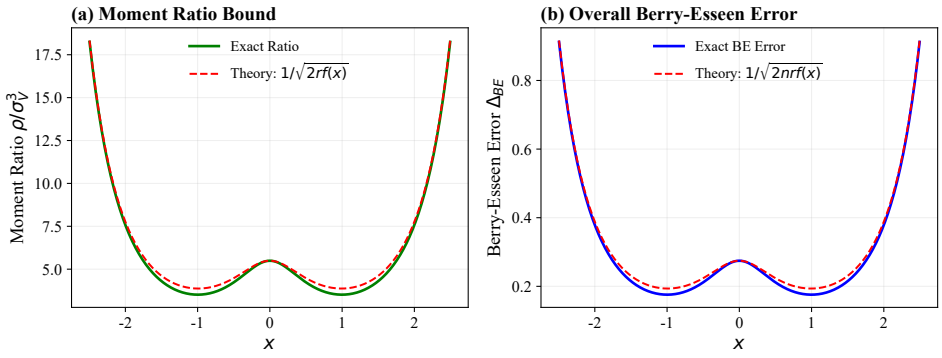


Figure 4: **Numerical breakdown of the Berry–Esseen error bound.** Subplot a shows that the exact standardized third-moment ratio follows the local-density scaling $1/\sqrt{2rp(x)}$. Subplot b illustrates that the normal-approximation error becomes larger in sparse regions, where the effective local sample size is smaller.

A.3.1 Auxiliary lemmas

We first state a fundamental result from high-dimensional probability theory, which bounds the normal approximation error over convex sets.

Lemma 1 (High-dimensional Berry–Esseen Theorem for Convex Sets (Bentkus, 2003))**.** *Let $\mathcal{C}$ be the class of all measurable convex sets in $\mathbb{R}^m$. Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be independent and identically distributed random vectors in $\mathbb{R}^m$ such that $\mathbb{E}[\mathbf{Z}_1] = \mathbf{0}$ and the covariance matrix is the identity $\mathbf{I}_m$. Denote the third absolute moment as $\beta = \mathbb{E}\|\mathbf{Z}_1\|^3$. Let $\mathbf{S}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{Z}_i$, and let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ be a standard Gaussian vector. Then, the distribution of $\mathbf{S}_n$ satisfies:*

$$\sup_{C \in \mathcal{C}} |\mathbb{P}(\mathbf{S}_n \in C) - \mathbb{P}(\mathbf{Y} \in C)| \leq C_B m^{1/4} \frac{\beta}{\sqrt{n}}, \tag{23}$$

where C_B is an absolute constant independent of the dimension m and sample size n. For non-identity covariance, we apply the theorem after whitening; equivalently, the Gaussian approximation below is written with the corresponding covariance matrix.

To apply Lemma 1 to our local density framework, we define the indicator vector $\mathbf{Z}_i(\mathbf{x}) \in \mathbb{R}^K$ for the i-th sample, where $Z_i^{(k)}(\mathbf{x}) = \mathbb{I}\{\mathbf{X}_i \in B(\mathbf{x}, r), Y_i = k\}$. The expected vector is $\boldsymbol{\mu}(\mathbf{x}) = \mathbb{E}[\mathbf{Z}_i(\mathbf{x})] = \mathbf{p}(\mathbf{x})$. To apply Lemma 1, we must control the third absolute moment after whitening by the local difference covariance. The following lemma shows that this moment scales as $p_{\text{loc}}(\mathbf{x})^{-1/2}$.

Lemma 2 (Whitened Third Absolute Moment)**.** *Let $J = \{1, \dots, K\} \setminus \{k^\star\} = \{j_1, \dots, j_{K-1}\}$, where the elements of J are given a fixed ordering. Define the difference matrix $A \in \mathbb{R}^{(K-1)\times K}$ by*

$$A_{\ell k} = \mathbb{I}\{k = j_\ell\} - \mathbb{I}\{k = k^\star\}, \qquad \ell = 1, \dots, K-1, \quad k = 1, \dots, K. \tag{24}$$

Equivalently, the ℓ-th row of A is $e_{j_\ell}^\top - e_{k^\star}^\top$, where $e_k \in \mathbb{R}^K$ is the k-th standard basis vector. Thus, for any $\mathbf{v} \in \mathbb{R}^K$, $A\mathbf{v} = \big(v_j - v_{k^\star}\big)_{j \in J}$.

Define the centered one-sample difference vector $\mathbf{U}_i(\mathbf{x}) = A\big(\mathbf{Z}_i(\mathbf{x}) - \mathbf{p}(\mathbf{x})\big)$. Then $\mathrm{Cov}\big(\mathbf{U}_i(\mathbf{x})\big) = \boldsymbol{\Sigma}_V(\mathbf{x})$, and its whitened third absolute moment satisfies

$$\mathbb{E}\left\|\boldsymbol{\Sigma}_V(\mathbf{x})^{-1/2}\mathbf{U}_i(\mathbf{x})\right\|_2^3 = \mathcal{O}\left(\frac{K^{3/2}}{\lambda_0(\mathbf{x})^{3/2}\sqrt{p_{\text{loc}}(\mathbf{x})}}\right). \tag{25}$$

Proof. Since $\mathbb{E}\mathbf{Z}_i(\mathbf{x}) = \mathbf{p}(\mathbf{x})$, we have $\mathbb{E}\mathbf{U}_i(\mathbf{x}) = A\left(\mathbb{E}\mathbf{Z}_i(\mathbf{x}) - \mathbf{p}(\mathbf{x})\right) = \mathbf{0}$. Moreover,

$$\mathrm{Cov}\big(\mathbf{U}_i(\mathbf{x})\big) = A\,\mathrm{Cov}\big(\mathbf{Z}_i(\mathbf{x})\big)A^\top = \boldsymbol{\Sigma}_V(\mathbf{x}),$$

where the last equality follows from Lemma 3

It remains to bound the third absolute moment after whitening. Denote $\rho = p_{\mathrm{loc}}(\mathbf{x}) = \sum_{k=1}^{K} p_{k,r}(\mathbf{x})$. Since $\mathbf{p}(\mathbf{x})$ has nonnegative entries summing to ρ, $\|\mathbf{p}(\mathbf{x})\|_2 \le \|\mathbf{p}(\mathbf{x})\|_1 = \rho$. If the sample is outside $B(\mathbf{x}, r)$, then $\mathbf{Z}_i(\mathbf{x}) = \mathbf{0}$, which occurs with probability $1-\rho$. If it falls inside $B(\mathbf{x}, r)$ and belongs to class k, then $\mathbf{Z}_i(\mathbf{x}) = e_k$, which occurs with probability $p_{k,r}(\mathbf{x})$. Therefore,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{Z}_i(\mathbf{x}) - \mathbf{p}(\mathbf{x})\|_2^3 &= (1-\rho)\|\mathbf{p}(\mathbf{x})\|_2^3 + \sum_{k=1}^{K} p_{k,r}(\mathbf{x})\,\|e_k - \mathbf{p}(\mathbf{x})\|_2^3 \\ &\le (1-\rho)\rho^3 + \sum_{k=1}^{K} p_{k,r}(\mathbf{x})\,\left(1 + \|\mathbf{p}(\mathbf{x})\|_2\right)^3 \\ &\le \rho^3 + 8\rho \\ &\le 9\rho. \end{aligned} \tag{26}$$

Next, the difference matrix satisfies $AA^\top = \mathbf{I}_{K-1} + \mathbf{1}\mathbf{1}^\top$. The eigenvalues of $\mathbf{I}_{K-1} + \mathbf{1}\mathbf{1}^\top$ are 1, with multiplicity $K-2$, and K, with multiplicity one. Consequently, $\|A\|_{\mathrm{op}} = \sqrt{\lambda_{\max}(AA^\top)} = \sqrt{K}$. It follows from $\mathbf{U}_i = A(\mathbf{Z}_i - \mathbf{p})$ that

$$\begin{aligned}\mathbb{E}\|\mathbf{U}_i(\mathbf{x})\|_2^3 &\le \|A\|_{\mathrm{op}}^3 \mathbb{E}\,\|\mathbf{Z}_i(\mathbf{x}) - \mathbf{p}(\mathbf{x})\|_2^3 \\ &\le 9K^{3/2}\rho. \end{aligned} \tag{27}$$

By the non-degeneracy assumption in Theorem 1, $\lambda_{\min}\left(\frac{\mathbf{\Sigma}_V(\mathbf{x})}{\rho}\right) \ge \lambda_0(\mathbf{x})$, and hence $\lambda_{\min}\big(\mathbf{\Sigma}_V(\mathbf{x})\big) \ge \lambda_0(\mathbf{x})\rho$. Therefore, $\left\|\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\right\|_{\mathrm{op}} = \frac{1}{\sqrt{\lambda_{\min}(\mathbf{\Sigma}_V(\mathbf{x}))}} \le \frac{1}{\sqrt{\lambda_0(\mathbf{x})\rho}}$. Combining this inequality with Eq. 27 gives

$$\begin{aligned}\mathbb{E}\left\|\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\mathbf{U}_i(\mathbf{x})\right\|_2^3 &\le \left\|\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\right\|_{\mathrm{op}}^3 \mathbb{E}\|\mathbf{U}_i(\mathbf{x})\|_2^3 \\ &\le \frac{9K^{3/2}\rho}{(\lambda_0(\mathbf{x})\rho)^{3/2}} = \frac{9K^{3/2}}{\lambda_0(\mathbf{x})^{3/2}\sqrt{\rho}} = \mathcal{O}\!\left(\frac{1}{\sqrt{\rho}}\right) = \mathcal{O}\!\left(\frac{1}{\sqrt{p_{\mathrm{loc}}(\mathbf{x})}}\right). \end{aligned} \tag{28}$$

This proves Eq. 25. □

Next, we establish the explicit covariance structure in the difference space where the decision boundary is evaluated.

Lemma 3 (Covariance Structure of the Difference Space)**.** *Under the small-radius relation in Section 2, $k^\star(\mathbf{x}) = k_r^\star(\mathbf{x}) = \arg\max_k p_{k,r}(\mathbf{x})$. Define the standardized sum $\mathbf{S}_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{Z}_i(\mathbf{x}) - \boldsymbol{\mu}(\mathbf{x}))$. Let $J = \{1,\dots,K\} \setminus \{k^\star\}$. Define the difference vector $\mathbf{V}(\mathbf{x}) \in \mathbb{R}^{K-1}$ by*

$$\mathbf{V}(\mathbf{x}) = \big(S_{n,j}(\mathbf{x}) - S_{n,k^\star}(\mathbf{x})\big)_{j\in J}.$$

Equivalently, $\mathbf{V}(\mathbf{x}) = A\mathbf{S}_n(\mathbf{x})$, where the row of $A \in \mathbb{R}^{(K-1)\times K}$ corresponding to class $j \in J$ is $e_j^\top - e_{k^\star}^\top$. Let $\mathbf{p}_{-k^\star,r} = (p_{j,r}(\mathbf{x}))_{j\in J} \in \mathbb{R}^{K-1}$ be the vector of trap probabilities, ordered consistently with $\mathbf{V}(\mathbf{x})$. The covariance matrix $\mathbf{\Sigma}_V(\mathbf{x}) = \mathrm{Cov}(\mathbf{V}(\mathbf{x}))$ is explicitly given by:

$$\mathbf{\Sigma}_V(\mathbf{x}) = \mathrm{diag}(\mathbf{p}_{-k^\star,r}) - \mathbf{p}_{-k^\star,r}\mathbf{p}_{-k^\star,r}^\top + p_{k^\star,r}(1-p_{k^\star,r})\mathbf{1}\mathbf{1}^\top + p_{k^\star,r}(\mathbf{p}_{-k^\star,r}\mathbf{1}^\top + \mathbf{1}\mathbf{p}_{-k^\star,r}^\top). \tag{29}$$

Proof. Let $\mathbf{\Sigma}(\mathbf{x}) = \mathrm{Cov}(\mathbf{Z}_1(\mathbf{x})) \in \mathbb{R}^{K\times K}$, and write $\Sigma_{k\ell}$ for its (k,ℓ)-th entry. Since $Z_1^{(k)}(\mathbf{x})$ is an indicator variable with mean $p_{k,r}(\mathbf{x})$, its variance is

$$\Sigma_{kk} = \mathrm{Var}\left(Z_1^{(k)}(\mathbf{x})\right) = p_{k,r}(\mathbf{x})(1 - p_{k,r}(\mathbf{x})).$$

For $k \ne \ell$, the two indicators $Z_1^{(k)}(\mathbf{x})$ and $Z_1^{(\ell)}(\mathbf{x})$ cannot both equal one for the same sample. Hence

$$\mathbb{E}\big[Z_1^{(k)}(\mathbf{x})Z_1^{(\ell)}(\mathbf{x})\big] = 0,$$

and therefore

$$\Sigma_{k\ell} = \mathrm{Cov}\left(Z_1^{(k)}(\mathbf{x}), Z_1^{(\ell)}(\mathbf{x})\right) = -p_{k,r}(\mathbf{x})p_{\ell,r}(\mathbf{x}).$$

Thus $\mathbf{\Sigma}(\mathbf{x})$ has the standard multinomial-like covariance structure. Since $\mathbf{V}(\mathbf{x}) = A\mathbf{S}_n(\mathbf{x})$, its covariance satisfies

$$\mathbf{\Sigma}_V(\mathbf{x}) = A\mathbf{\Sigma}(\mathbf{x})A^\top.$$

Equivalently, for $j, \ell \in J$, the entries of $\mathbf{\Sigma}_V(\mathbf{x})$ are

$$(\mathbf{\Sigma}_V)_{j\ell} = \mathrm{Cov}(S_{n,j} - S_{n,k^\star}, S_{n,\ell} - S_{n,k^\star}) = \Sigma_{j\ell} - \Sigma_{jk^\star} - \Sigma_{k^\star\ell} + \Sigma_{k^\star}. \tag{30}$$

For diagonal elements where $j = \ell \neq k^\star$:

$$(\mathbf{\Sigma}_V)_{jj} = p_{j,r} - p_{j,r}^2 + 2p_{j,r}p_{k^\star,r} + p_{k^\star,r}(1 - p_{k^\star,r}). \tag{31}$$

For off-diagonal elements where $j \neq \ell$ and $j, \ell \neq k^\star$:

$$(\mathbf{\Sigma}_V)_{j\ell} = -p_{j,r}p_{\ell,r} + p_{j,r}p_{k^\star,r} + p_{k^\star,r}p_{\ell,r} + p_{k^\star,r}(1 - p_{k^\star,r}). \tag{32}$$

Let $q = \mathbf{p}_{-k^\star,r}$ and $p_* = p_{k^\star,r}$. For $j, \ell \in J$,

$$\begin{aligned}(\mathbf{\Sigma}_V)_{j\ell} &= \Sigma_{j\ell} - \Sigma_{jk^\star} - \Sigma_{k^\star\ell} + \Sigma_{k^\star} \\ &= \mathbb{I}\{j = \ell\}p_{j,r} - p_{j,r}p_{\ell,r} + p_{j,r}p_* + p_*p_{\ell,r} + p_*(1 - p_*) \\ &= \left[\mathrm{diag}(q) - qq^\top + p_*(1 - p_*)\mathbf{1}\mathbf{1}^\top + p_*\left(q\mathbf{1}^\top + \mathbf{1}q^\top\right)\right]_{j\ell}. \end{aligned} \tag{33}$$

Hence,

$$\mathbf{\Sigma}_V(\mathbf{x}) = \mathrm{diag}(\mathbf{p}_{-k^\star,r}) - \mathbf{p}_{-k^\star,r}\mathbf{p}_{-k^\star,r}^\top + p_{k^\star,r}(1 - p_{k^\star,r})\mathbf{1}\mathbf{1}^\top + p_{k^\star,r}\left(\mathbf{p}_{-k^\star,r}\mathbf{1}^\top + \mathbf{1}\mathbf{p}_{-k^\star,r}^\top\right). \tag{34}$$

□

A.3.2 Proof of the main theorem

With the auxiliary lemmas established, we now complete the proof of Theorem 1.

Proof of Theorem 1. Fix $\mathbf{x}$, let $J = \{1, \ldots, K\} \setminus \{k^\star\}$, and let $A \in \mathbb{R}^{(K-1)\times K}$ be the difference matrix defined in Eq. 24. Define

$$\mathbf{U}_i(\mathbf{x}) = A\big(\mathbf{Z}_i(\mathbf{x}) - \mathbf{p}(\mathbf{x})\big).$$

By Lemma 2,

$$\mathbb{E}\mathbf{U}_i(\mathbf{x}) = \mathbf{0}, \qquad \mathrm{Cov}\big(\mathbf{U}_i(\mathbf{x})\big) = \mathbf{\Sigma}_V(\mathbf{x}).$$

Moreover,

$$A\mathbf{p}(\mathbf{x}) = -\mathbf{\Delta}(\mathbf{x}),$$

and therefore

$$\begin{aligned}\sum_{i=1}^n \mathbf{U}_i(\mathbf{x}) &= A\left(\sum_{i=1}^n \mathbf{Z}_i(\mathbf{x}) - n\mathbf{p}(\mathbf{x})\right) \\ &= \mathbf{D}_n(\mathbf{x}) + n\mathbf{\Delta}(\mathbf{x}). \end{aligned} \tag{35}$$

Define

$$\mathbf{W}_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\big(\mathbf{D}_n(\mathbf{x}) + n\mathbf{\Delta}(\mathbf{x})\big).$$

Equivalently,

$$\mathbf{W}_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\mathbf{U}_i(\mathbf{x}),$$

where the whitened summands $\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}\mathbf{U}_i(\mathbf{x})$ are independent, centered, and have identity covariance. Hence, Lemmas 1 and 2 give

$$\sup_{C\in\mathcal{C}_{K-1}} \left|\mathbb{P}\big(\mathbf{W}_n(\mathbf{x})\in C\big) - \mathbb{P}(\mathbf{G}\in C)\right| = \mathcal{O}\left(\frac{(K-1)^{1/4}}{\sqrt{np_{\mathrm{loc}}(\mathbf{x})}}\right) = \mathcal{O}\left(\frac{(K-1)^{1/4}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right). \tag{36}$$

This proves Eq. 5.

For completeness, define the lower orthant

$$L_n(\mathbf{x}) = \left\{\mathbf{v}\in\mathbb{R}^{K-1} : \mathbf{v} < \sqrt{n}\,\mathbf{\Delta}(\mathbf{x})\right\}.$$

The non-flip event satisfies

$$\{\mathbf{D}_n(\mathbf{x}) < \mathbf{0}\} = \left\{\frac{\mathbf{D}_n(\mathbf{x}) + n\mathbf{\Delta}(\mathbf{x})}{\sqrt{n}} \in L_n(\mathbf{x})\right\}.$$

Since $\mathbf{\Sigma}_V(\mathbf{x})^{-1/2}L_n(\mathbf{x})$ is convex, Eq. 5 implies

$$\mathbb{P}\left(\mathbf{D}_n(\mathbf{x}) < \mathbf{0}\right) = \Phi_{K-1}\left(\sqrt{n}\,\mathbf{\Delta}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\right) + \mathcal{O}\left(\frac{(K-1)^{1/4}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right).$$

Taking complements yields

$$P_{\mathrm{adv}}(\mathbf{x}) = 1 - \Phi_{K-1}\left(\sqrt{n}\,\mathbf{\Delta}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\right) + \mathcal{O}\left(\frac{(K-1)^{1/4}}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right).$$

□

A.4 Reduction of Theorem 1 to the binary case

We verify that the Gaussian approximation in Theorem 1 reduces to the binary form in Proposition 1. When $K = 2$, there is only one competing class $j \neq k^\star(\mathbf{x})$. Hence,

$$\Delta(\mathbf{x}) = p_{k^\star,r}(\mathbf{x}) - p_{j,r}(\mathbf{x}),$$

and the difference covariance reduces to

$$\Sigma_V(\mathbf{x}) = p_{k^\star,r}(\mathbf{x}) + p_{j,r}(\mathbf{x}) - \big(p_{k^\star,r}(\mathbf{x}) - p_{j,r}(\mathbf{x})\big)^2.$$

Moreover,

$$n_{\mathrm{eff}}(\mathbf{x}) = n\big(p_{k^\star,r}(\mathbf{x}) + p_{j,r}(\mathbf{x})\big).$$

Let

$$D_n^{\mathrm{trap}}(\mathbf{x}) = N_{r,n,j}(\mathbf{x}) - N_{r,n,k^\star}(\mathbf{x}).$$

Specializing Eq. 5 to $K = 2$ gives

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left(\frac{D_n^{\mathrm{trap}}(\mathbf{x}) + n\Delta(\mathbf{x})}{\sqrt{n\Sigma_V(\mathbf{x})}} \le z\right) - \Phi(z)\right| = \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right). \tag{37}$$

To match the notation of Proposition 1, define the dominant-minus-trap difference

$$D_n^{\mathrm{dom}}(\mathbf{x}) = N_{r,n,k^\star}(\mathbf{x}) - N_{r,n,j}(\mathbf{x}) = -D_n^{\mathrm{trap}}(\mathbf{x}).$$

Since the standard normal distribution is symmetric, Eq. 37 is equivalently

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left(\frac{D_n^{\mathrm{dom}}(\mathbf{x}) - n\Delta(\mathbf{x})}{\sqrt{n\Sigma_V(\mathbf{x})}} \le z\right) - \Phi(z)\right| = \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right). \tag{38}$$

This recovers the one-dimensional Berry–Esseen bound after identifying $p_{k^\star,r} = p_{1,r}$, $p_{j,r} = p_{0,r}$, and $D_n^{\mathrm{dom}} = D$.

Finally, evaluating Eq. 38 at

$$z = -\frac{\sqrt{n}\,\Delta(\mathbf{x})}{\sqrt{\Sigma_V(\mathbf{x})}}$$

gives the binary EAR approximation

$$P_{\mathrm{adv}}(\mathbf{x}) = \Phi\left(-\sqrt{n}\,\frac{p_{k^\star,r}(\mathbf{x}) - p_{j,r}(\mathbf{x})}{\sqrt{p_{k^\star,r}(\mathbf{x}) + p_{j,r}(\mathbf{x}) - \big(p_{k^\star,r}(\mathbf{x}) - p_{j,r}(\mathbf{x})\big)^2}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{eff}}(\mathbf{x})}}\right).$$

A.5 Validation on the CIFAR-10 Dataset

To evaluate the Gaussian EAR approximation beyond synthetic settings, we conduct an empirical validation on CIFAR-10. We use a frozen CNN feature extractor and compute channel-wise spatial means from four convolutional blocks, obtaining 960-dimensional features. We then select 32 dimensions by Fisher score, robust-standardize them using training-set statistics, and fit class-wise marginal Gaussian KDEs in the resulting feature space. The product-box radius is selected on a held-out validation set as the 90th percentile of the nearest-training-sample ℓ_∞ distance, yielding $r = 1.17$. The fitted KDEs are then used to estimate the local class masses $p_{j,r}(x)$ in Eq. 4.

A continuity correction is needed because the discrete local-count event in Eq. 4 is defined by discrete local counts, whereas the Gaussian approximation is continuous. For each competing class $j \neq k^\star$, the count difference $N_{k^\star}(x) - N_j(x)$ can take only integer values. In particular, $N_{k^\star}(x) - N_j(x) = 0$ is a tie and is conservatively counted as an error, whereas $N_{k^\star}(x) - N_j(x) = 1$ is the smallest strictly winning outcome. The relevant transition is therefore

$$\underbrace{N_{k^\star}(x) - N_j(x) = 0}_{\text{tie: error}} < \underbrace{c}_{\text{Gaussian boundary}} < \underbrace{N_{k^\star}(x) - N_j(x) = 1}_{\text{smallest strict win}}.$$

Without a continuity correction, the Gaussian boundary is placed directly at zero and does not account for the probability mass represented by the discrete tie. This mismatch is most consequential near class boundaries, where the EAR is governed by small differences between local class counts. The continuity correction instead places the continuous boundary between the tie and the smallest strict win.

Using the notation of Theorem 1, where $\Delta_j(x) = p_{k^\star,r}(x) - p_{j,r}(x)$ and $\mathbf{\Sigma}_V(x)$ is defined in Theorem 1, the continuity-corrected Gaussian approximation is

$$P_{\mathrm{adv}}(x) \approx 1 - \Phi_{K-1}\left(\sqrt{n}\,\Delta(x) - \frac{c}{\sqrt{n}}\mathbf{1}; \boldsymbol{\Sigma_V(x)}\right).$$

The conventional midpoint correction uses $c = 0.5$. We evaluate nearby correction values and use $c = 0.6$, which gives the closest agreement with the Monte Carlo reference on CIFAR-10. This value slightly shifts the Gaussian boundary from the midpoint toward the smallest strict win while preserving the distinction between a tie and a correct local majority.

On 10,000 CIFAR-10 test images, the Gaussian approximation with $c = 0.6$ achieves an MAE of 4.55×10^{-3}, an RMSE of 1.19×10^{-2}, a 95th-percentile absolute error of 3.28×10^{-2}, and a correlation of 0.999 against Monte Carlo simulation of the same fixed-radius local-count event. As shown in Fig. 5, the estimated EAR values closely follow the diagonal over the full risk range.

A.6 Validation of EAR on neural networks

The radius-nearest-neighbor rule provides a tractable reference model for analyzing EAR, while EAR itself is defined for a general model family. We examine whether the risk geometry predicted by EAR is also reflected by neural classifiers trained on finite samples. Specifically, we compare the analytical EAR with the empirical flip risk of MLPs across one base configuration and four randomly generated data geometries.

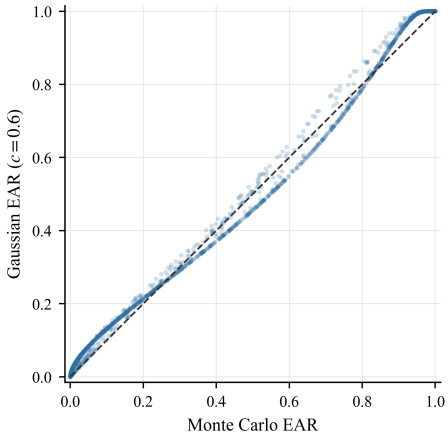


Figure 5: Continuity-corrected Gaussian EAR estimates with $c = 0.6$ versus Monte Carlo estimates of $P_{\mathrm{adv}}(x)$ on 10,000 CIFAR-10 test images. Both estimates use the same fixed-radius local-count event. The dashed diagonal denotes exact agreement, and $r = 1.17$ is selected using only the held-out validation set.

Data configurations. All experiments use 3-class 2D Gaussian mixtures with uniform class priors, neighborhood radius $r = 0.5$, and training size $n = 1000$. For the base configuration (Exp 1), the class means are $\boldsymbol{\mu}_1 = (0, 0)$, $\boldsymbol{\mu}_2 = (2.8, 1.1)$, and $\boldsymbol{\mu}_3 = (2.8, -1.1)$. The covariance matrices are $\boldsymbol{\Sigma}_1 = 0.55\mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.75 & 0.18 \\ 0.18 & 0.55 \end{bmatrix}$, $\boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.75 & -0.18 \\ -0.18 & 0.55 \end{bmatrix}$. For Exp 2–5, the first two components are fixed at $(2, 0)$ and $(-2, 0)$ with covariance $0.8\mathbf{I}_2$. The third component is sampled as $\boldsymbol{\mu}_3 \sim \mathrm{Uniform}([-2, 2]^2)$, $\boldsymbol{\Sigma}_3 = \sigma^2\mathbf{I}_2$, $\sigma^2 \sim \mathrm{Uniform}(0.4, 1.2)$.

MLP training and empirical risk. We use an MLP with two hidden layers of width 64 and GELU activations. Each model is trained on a newly sampled dataset using Adam with learning rate 10^{-3}, weight decay 10^{-3}, cross-entropy loss, and early stopping on a 20% validation split. At each grid point, empirical risk is the misclassification frequency over 800 independently sampled training sets and MLP training runs, measured relative to the Bayes-optimal decision.

Results. Across all five configurations, the analytical EAR and empirical MLP risk exhibit similar spatial patterns, with high-risk regions concentrated near class boundaries. This correspondence persists as the location and variance of the third component change, suggesting that the risk geometry identified by the local nonparametric analysis is also reflected by neural classifiers.

The visual comparisons below are presented in the order Exp 1–5. Exp 1 is the base configuration described above. Exp 2 uses seed 49, $\boldsymbol{\mu}_3 = (-0.796, -1.012)$, and $\sigma^2 = 1.141$; Exp 3 uses seed 91, $\boldsymbol{\mu}_3 = (-1.196, -0.684)$, and $\sigma^2 = 0.637$; Exp 4 uses seed 133, $\boldsymbol{\mu}_3 = (-0.322, -1.481)$, and $\sigma^2 = 0.488$; and Exp 5 uses seed 175, $\boldsymbol{\mu}_3 = (1.208, -0.936)$, and $\sigma^2 = 0.843$.

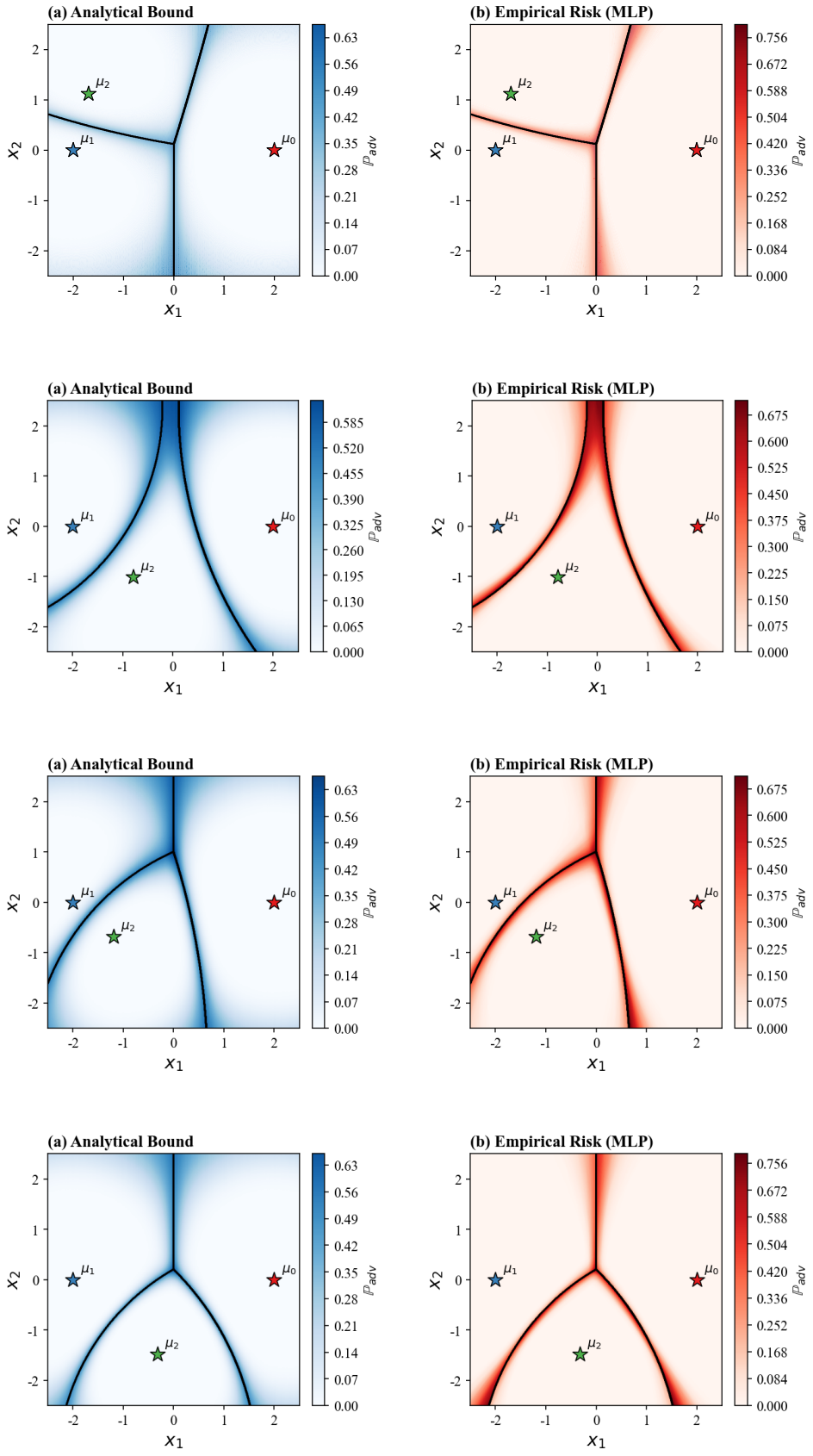
In each image, the left panel shows the analytical EAR from Theorem 1, and the right panel shows empirical MLP risk averaged over 800 independent training runs. Darker colors indicate higher risk, and stars denote class means. The images follow the configuration order given above.

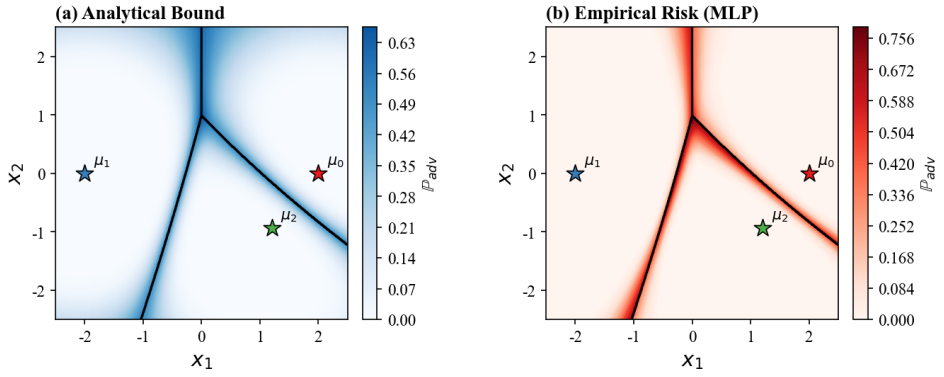
B Relation between EAR and ASR

For a classifier $f_{\mathcal{D}_n}$ trained on $\mathcal{D}_n$, its attack success rate (ASR) is defined as

$$\mathrm{ASR}(f_{\mathcal{D}_n}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\sup_{\hat{\mathbf{x}}\in B_\epsilon(\mathbf{x})} \mathbb{I}\left\{f_{\mathcal{D}_n}(\hat{\mathbf{x}}) \neq k^\star(\hat{\mathbf{x}})\right\}\right]. \tag{39}$$

Intuitively, there is a close relation between ASR and EAR: sample $\mathbf{x}$ can be successfully attacked if and only if there is an adversarial example in $B_\epsilon(\mathbf{x})$, so if for all $\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})$ the corresponding EAR $P_{\mathrm{adv}}(\hat{\mathbf{x}}; \mathcal{F}, P, n)$ decreases, $\mathbf{x}$ would be less likely to be attacked. However, it is non-trivial to build a strict connection between them without making assumptions on the joint distribution of nearby


(a) Analytical Bound
(b) Empirical Risk (MLP)
μ_0
μ_1
μ_2
x_1
x_2
$\mathbb{P}_{adv}$



points being adversarial. In an extreme case where all points are independently to be adversarial or not, the ASR over $\mathbf{x}$ would either be 0 or 1, which is almost unaffected by EAR, since there are infinite points in $B_\epsilon(\mathbf{x})$ and only require one point to be adversarial.

Therefore, we propose an assumption that depicts a practically reasonable joint distribution of nearby points being adversarial. That is, if point $\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})$ is an adversarial example for $f_{\mathcal{D}_n}$, all points in $B_\epsilon(\mathbf{x})$ with larger EAR than it are also adversarial examples for $f_{\mathcal{D}_n}$. The formal definition is given in Assumption 1.

Assumption 1. *For any input $\mathbf{x}$ and training set $\mathcal{D}_n \sim P^n$, the occurrence of an adversarial example within $B_\epsilon(\mathbf{x})$ is monotone with respect to EAR. Specifically, for any $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \in B_\epsilon(\mathbf{x})$,*

$$\left.\begin{array}{c} P_{\text{adv}}(\hat{\mathbf{x}}_1; \mathcal{F}, P, n) \le P_{\text{adv}}(\hat{\mathbf{x}}_2; \mathcal{F}, P, n), \\ a_{\mathcal{D}_n}(\hat{\mathbf{x}}_1; \mathbf{x}) = 1 \end{array}\right\} \quad \Longrightarrow \quad a_{\mathcal{D}_n}(\hat{\mathbf{x}}_2; \mathbf{x}) = 1. \tag{40}$$

where $a_{\mathcal{D}_n}(\hat{\mathbf{x}}; \mathbf{x}) = \mathbb{I}\left[\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x}) \wedge f_{\mathcal{D}_n}(\hat{\mathbf{x}}) \neq k^\star(\hat{\mathbf{x}})\right]$.

Under this assumption, we derive the relation between EAR and ASR in Proposition 3. An important intermediate conclusion is given in Eq. 52, which says the probability of $\mathbf{x}$ being successfully attacked equals the maximum EAR value in $B_\epsilon(\mathbf{x})$. This supports our finding that EAR minimization through purification also achieves decreased ASR.

Proposition 3 (Relationship between EAR and ASR)**.** *Under Assumption 1, averaging ASR over training sets sampled from P^n yields expected maximum EAR within the perturbation set:*

$$\mathbb{E}_{\mathcal{D}_n \sim P^n}\left[\text{ASR}(f_{\mathcal{D}_n})\right] = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}\left[\sup_{\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})} P_{\text{adv}}(\hat{\mathbf{x}}; \mathcal{F}, P, n)\right]. \tag{41}$$

Proof. Fix an input $\mathbf{x}$. For every $\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})$, Definition 3.1 and the definition of $a_{\mathcal{D}_n}(\hat{\mathbf{x}}; \mathbf{x})$ give

$$\mathbb{E}_{\mathcal{D}_n \sim P^n}\left[a_{\mathcal{D}_n}(\hat{\mathbf{x}}; \mathbf{x})\right] = P_{\text{adv}}(\hat{\mathbf{x}}; \mathcal{F}, P, n). \tag{42}$$

Define the maximum EAR within the perturbation set as

$$M_\epsilon(\mathbf{x}) = \sup_{\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})} P_{\text{adv}}(\hat{\mathbf{x}}; \mathcal{F}, P, n). \tag{43}$$

Choose a sequence $\{\hat{\mathbf{x}}_m\}_{m=1}^{\infty} \subseteq B_\epsilon(\mathbf{x})$ such that

$$P_{\text{adv}}(\hat{\mathbf{x}}_1; \mathcal{F}, P, n) \le P_{\text{adv}}(\hat{\mathbf{x}}_2; \mathcal{F}, P, n) \le \cdots \longrightarrow M_\epsilon(\mathbf{x}). \tag{44}$$

If the supremum is attained, the sequence can be chosen to be constant at a maximizer.

For each m, define

$$\mathcal{E}_m(\mathbf{x}) = \left\{\mathcal{D}_n : a_{\mathcal{D}_n}(\hat{\mathbf{x}}_m; \mathbf{x}) = 1\right\}. \tag{45}$$

For any $\mathcal{D}_n \in \mathcal{E}_m(\mathbf{x})$, we have

$$a_{\mathcal{D}_n}(\hat{\mathbf{x}}_m; \mathbf{x}) = 1.$$

Since

$$P_{\text{adv}}(\hat{\mathbf{x}}_m; \mathcal{F}, P, n) \leq P_{\text{adv}}(\hat{\mathbf{x}}_{m+1}; \mathcal{F}, P, n),$$

Assumption 1 implies

$$a_{\mathcal{D}_n}(\hat{\mathbf{x}}_{m+1}; \mathbf{x}) = 1.$$

Thus, $\mathcal{E}_m(\mathbf{x}) \subseteq \mathcal{E}_{m+1}(\mathbf{x})$, and consequently

$$\mathcal{E}_1(\mathbf{x}) \subseteq \mathcal{E}_2(\mathbf{x}) \subseteq \cdots . \tag{46}$$

Define the event that the perturbation set contains at least one adversarial example as

$$\mathcal{E}(\mathbf{x}) = \left\{ \mathcal{D}_n : \sup_{\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})} a_{\mathcal{D}_n}(\hat{\mathbf{x}}; \mathbf{x}) = 1 \right\}. \tag{47}$$

Because every $\hat{\mathbf{x}}_m$ belongs to $B_\epsilon(\mathbf{x})$, any training set in $\mathcal{E}_m(\mathbf{x})$ also belongs to $\mathcal{E}(\mathbf{x})$. Hence,

$$\bigcup_{m=1}^{\infty} \mathcal{E}_m(\mathbf{x}) \subseteq \mathcal{E}(\mathbf{x}). \tag{48}$$

For the reverse direction, take any $\mathcal{D}_n \in \mathcal{E}(\mathbf{x})$. By the definition of $\mathcal{E}(\mathbf{x})$, there exists $\hat{\mathbf{x}}' \in B_\epsilon(\mathbf{x})$ such that

$$a_{\mathcal{D}_n}(\hat{\mathbf{x}}'; \mathbf{x}) = 1. \tag{49}$$

If $M_\epsilon(\mathbf{x})$ is attained at $\hat{\mathbf{x}}^\star$, then

$$P_{\text{adv}}(\hat{\mathbf{x}}'; \mathcal{F}, P, n) \leq P_{\text{adv}}(\hat{\mathbf{x}}^\star; \mathcal{F}, P, n).$$

Assumption 1 and Eq. 49 imply

$$a_{\mathcal{D}_n}(\hat{\mathbf{x}}^\star; \mathbf{x}) = 1.$$

Since the sequence can be chosen to be constant at $\hat{\mathbf{x}}^\star$, we obtain $\mathcal{D}_n \in \mathcal{E}_m(\mathbf{x})$.

If $M_\epsilon(\mathbf{x})$ is not attained, then

$$P_{\text{adv}}(\hat{\mathbf{x}}'; \mathcal{F}, P, n) < M_\epsilon(\mathbf{x}).$$

The convergence in Eq. 44 ensures that there exists an m such that

$$P_{\text{adv}}(\hat{\mathbf{x}}'; \mathcal{F}, P, n) \leq P_{\text{adv}}(\hat{\mathbf{x}}_m; \mathcal{F}, P, n).$$

Assumption 1 and Eq. 49 then imply

$$a_{\mathcal{D}_n}(\hat{\mathbf{x}}_m; \mathbf{x}) = 1,$$

and hence $\mathcal{D}_n \in \mathcal{E}_m(\mathbf{x})$.

In both cases, $\mathcal{D}_n \in \bigcup_{m=1}^{\infty} \mathcal{E}_m(\mathbf{x})$. Therefore,

$$\mathcal{E}(\mathbf{x}) \subseteq \bigcup_{m=1}^{\infty} \mathcal{E}_m(\mathbf{x}). \tag{50}$$

Combining Eqs. 48 and 50 yields

$$\mathcal{E}(\mathbf{x}) = \bigcup_{m=1}^{\infty} \mathcal{E}_m(\mathbf{x}). \tag{51}$$

The random variable

$$\sup_{\hat{\mathbf{x}} \in B_\epsilon(\mathbf{x})} a_{\mathcal{D}_n}(\hat{\mathbf{x}}; \mathbf{x})$$

is the indicator of $\mathcal{E}(\mathbf{x})$. Moreover, Eq. 46 shows that $\{\mathcal{E}_m(\mathbf{x})\}_{m\geq 1}$ is an increasing sequence of events. Continuity of probability from below therefore gives

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_n\sim P^n}\left[\sup_{\hat{\mathbf{x}}\in B_\epsilon(\mathbf{x})} a_{\mathcal{D}_n}(\hat{\mathbf{x}};\mathbf{x})\right] \\
&\quad= \mathbb{P}_{\mathcal{D}_n\sim P^n}\left(\mathcal{E}(\mathbf{x})\right) \\
&\quad= \lim_{m\to\infty} \mathbb{P}_{\mathcal{D}_n\sim P^n}\left(\mathcal{E}_m(\mathbf{x})\right) \\
&\quad= \lim_{m\to\infty} \mathbb{E}_{\mathcal{D}_n\sim P^n}\left[a_{\mathcal{D}_n}(\hat{\mathbf{x}}_m;\mathbf{x})\right] \\
&\quad= \lim_{m\to\infty} P_{\mathrm{adv}}(\hat{\mathbf{x}}_m;\mathcal{F},P,n) \\
&\quad= M_\epsilon(\mathbf{x}) \\
&\quad= \sup_{\hat{\mathbf{x}}\in B_\epsilon(\mathbf{x})} P_{\mathrm{adv}}(\hat{\mathbf{x}};\mathcal{F},P,n).
\end{aligned} \tag{52}$$

Finally, averaging Eq. 52 over $\mathbf{x}\sim p(\mathbf{x})$ and exchanging the order of the two expectations gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n\sim P^n}\left[\mathrm{ASR}(f_{\mathcal{D}_n})\right] &= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathcal{D}_n\sim P^n}\left[\sup_{\hat{\mathbf{x}}\in B_\epsilon(\mathbf{x})} a_{\mathcal{D}_n}(\hat{\mathbf{x}};\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\sup_{\hat{\mathbf{x}}\in B_\epsilon(\mathbf{x})} P_{\mathrm{adv}}(\hat{\mathbf{x}};\mathcal{F},P,n)\right].
\end{aligned} \tag{53}$$

This proves Eq. 41. □

C Proofs of section 4

C.1 Detailed proofs for 1D gradient dynamics

We first provide the detailed algebraic proofs for the 1D gradient dynamics discussed in Section 4.1. We consider a symmetric Gaussian mixture model where the class conditional distributions are given by $p_+(x)=\mathcal{N}(1,\sigma^2)$ and $p_-(x)=\mathcal{N}(-1,\sigma^2)$ and $p(x)=\frac{1}{2}\mathcal{N}(1,\sigma^2)+\frac{1}{2}\mathcal{N}(-1,\sigma^2)$.

Proof of Proposition 2. To prove the directional alignment, we analyze the signs of the gradients of the unconditional density $\nabla\log p(x)$, the maximum conditional density $\nabla\log M(x)$, and the negative intrinsic risk $-\nabla\tilde{P}_{\mathrm{adv}}(x)$ in the positive half-space $x>0$. We divide the proof into three steps.

Step 1: Stationary point of the unconditional density and defining x_p^*.
The unconditional density is $p(x)=\frac{1}{2}p_+(x)+\frac{1}{2}p_-(x)$. The score function, which represents the gradient of the log-density, is:

$$s(x)=\nabla_x\log p(x)=\frac{\frac{1}{2}\nabla_x p_+(x)+\frac{1}{2}\nabla_x p_-(x)}{p(x)}. \tag{54}$$

Substituting the derivatives of the Gaussian components yields:

$$s(x)=-\frac{x}{\sigma^2}+\frac{1}{2\sigma^2}\frac{p_+(x)-p_-(x)}{p(x)}=-\frac{x}{\sigma^2}+\frac{1}{\sigma^2}\tanh\left(\frac{x}{\sigma^2}\right). \tag{55}$$

For $\sigma^2<1$ (well-separated), we evaluate the boundaries. As $x\to 0^+$, let $t=x/\sigma^2\to 0^+$. Using the first-order expansion $\tanh(t)=t+o(t)$, we have $s(x)=\frac{x}{\sigma^4}(1-\sigma^2)+o(x)$. Hence, when $\sigma^2<1$, $s(x)>0$ for sufficiently small $x>0$.

Conversely, evaluating at the class center $x=1$ yields:

$$s(1)=\frac{1}{\sigma^2}\left[\tanh\left(\frac{1}{\sigma^2}\right)-1\right]<0. \tag{56}$$

Moreover,

$$s''(x) = -\frac{2}{\sigma^6}\operatorname{sech}^2\left(\frac{x}{\sigma^2}\right)\tanh\left(\frac{x}{\sigma^2}\right) < 0, \qquad x > 0. \tag{57}$$

Hence $s(x)$ is strictly concave on $(0,\infty)$. Since $s(x) > 0$ for sufficiently small $x > 0$, while $s(1) < 0$, continuity guarantees the existence of a root $x_p^* \in (0,1)$. The strict concavity of s on $(0,\infty)$ guarantees that this positive root is unique. Therefore, for $x \in (0, x_p^*)$, $\nabla \log p(x) > 0$.

For $\sigma^2 > 1$ (overlapping), we analyze the behavior near $x = 0$ using the Taylor expansion $\tanh(t) = t - \frac{1}{3}t^3 + O(t^5)$ with $t = x/\sigma^2$:

$$\begin{aligned} s(x) &= -\frac{x}{\sigma^2} + \frac{1}{\sigma^2}\left(\frac{x}{\sigma^2} - \frac{1}{3}\frac{x^3}{\sigma^6} + O(x^5)\right) \\ &= \frac{x}{\sigma^4}(1-\sigma^2) - \frac{1}{3}\frac{x^3}{\sigma^8} + O(x^5). \end{aligned} \tag{58}$$

For $\sigma^2 > 1$, the coefficient $1 - \sigma^2$ is negative, so $s(x) < 0$ for sufficiently small $x > 0$. Thus, near the boundary, the marginal score points leftward toward $x = 0$ rather than rightward toward the class center.

Step 2: Gradient of the maximum conditional density.
We define the maximum conditional density function as $M(x) = \max_{y\in\{0,1\}} p(x|Y = y)$. Given the symmetric Gaussian mixture, for $x > 0$, the maximum is achieved by the positive class $Y = 1$. Thus, we have:

$$M(x) = p(x|Y=1) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-1)^2}{2\sigma^2}\right), \quad x \geq 0. \tag{59}$$

To analyze the gradient direction, we compute the derivative of the log-likelihood $m(x) = \log M(x)$:

$$m'(x) = \nabla_x\left[\log C - \frac{(x-1)^2}{2\sigma^2}\right] = -\frac{x-1}{\sigma^2}, \quad x > 0. \tag{60}$$

Clearly, for any $x \in (0,1)$, we have $m'(x) > 0$. Since we established in Step 1 that $x_p^* < 1$, it immediately follows that for $x \in (0, x_p^*)$, $\nabla \log M(x) > 0$.

Step 3: Monotonicity of the adversarial risk.
The scaled argument of the adversarial risk is $g(x) = -\sqrt{nr}\,\frac{D(x)}{\sqrt{S(x)}}$, where $D(x) = p_+(x) - p_-(x)$ and $S(x) = p_+(x) + p_-(x)$. Therefore,

$$g'(x) = -\sqrt{nr}\,\frac{2D'(x)S(x) - D(x)S'(x)}{2S(x)^{3/2}}. \tag{61}$$

After factoring out strictly positive Gaussian terms, we obtain

$$g'(x) = -C(x)A(x), \qquad C(x) > 0,$$

where

$$A(x) = 2\cosh\left(\frac{2x}{\sigma^2}\right) - 2x\sinh\left(\frac{2x}{\sigma^2}\right) + 6. \tag{62}$$

Let $x^* > 0$ denote the first positive root of $A(x)$. Moreover,

$$A(1) = 2\cosh\left(\frac{2}{\sigma^2}\right) - 2\sinh\left(\frac{2}{\sigma^2}\right) + 6 = 2e^{-2/\sigma^2} + 6 > 0.$$

Therefore $x^* > 1$. By the definition of x^*,

$$A(x) > 0, \qquad 0 < x < x^*.$$

Since $C(x) > 0$, it follows that

$$g'(x) < 0, \qquad 0 < x < x^*.$$

Because

$$\widetilde{P}_{\mathrm{adv}}(x) = \Phi(g(x))$$

and Φ is strictly increasing, we conclude that

$$-\nabla \widetilde{P}_{\mathrm{adv}}(x) > 0, \qquad 0 < x < x^*.$$

Thus, within the monotone-decreasing regime of the intrinsic adversarial-risk approximation, the EAR descent direction points toward increasing x.

For completeness, the three directions for $x > 0$, where the dominant class is $Y = 1$ and $M = p_+$, can be written explicitly as

$$\text{Marginal score:} \quad \nabla \log p(x) = -\frac{x}{\sigma^2} + \frac{1}{\sigma^2} \tanh\left(\frac{x}{\sigma^2}\right), \tag{63}$$

$$\text{Conditional score:} \quad \nabla \log M(x) = -\frac{x-1}{\sigma^2}, \tag{64}$$

$$\text{EAR gradient:} \quad \nabla \widetilde{P}_{\mathrm{adv}}(x) \propto -\left[2\cosh\left(\frac{2x}{\sigma^2}\right) - 2x \sinh\left(\frac{2x}{\sigma^2}\right) + 6\right]. \tag{65}$$

Conclusion. Combining the three steps, in the separated regime $\sigma^2 < 1$, for $x \in (0, x_p^*)$, we have $-\nabla \widetilde{P}_{\mathrm{adv}}(x) > 0, \quad \nabla \log p(x) > 0, \quad \nabla \log M(x) > 0$. Thus both marginal and conditional density guidance align with EAR descent. In the overlapping regime $\sigma^2 > 1$, for sufficiently small $x > 0$, Step 1 gives $\nabla \log p(x) < 0$, while Steps 2 and 3 give $\nabla \log M(x) > 0$ and $-\nabla \widetilde{P}_{\mathrm{adv}}(x) > 0$. Hence marginal guidance is anti-aligned near the boundary, whereas conditional guidance remains aligned. This completes the proof of Proposition 2.

□

C.2 Exact derivation of the Gaussian EAR gradient and sufficient conditions for T_1 dominance

In this section, we derive the gradient of the Gaussian EAR approximation $\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ and establish a local finite-sample sufficient condition for T_1 dominance.

C.2.1 Decomposition of the EAR gradient

Recall that $\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) = 1 - \Phi_m\big(\mathbf{a}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\big),\ \mathbf{a}(\mathbf{x}) = \sqrt{n}\,\mathbf{\Delta}(\mathbf{x})$, where $m = K - 1$.

Lemma 4 (Gradient decomposition)**.** *The gradient of the Gaussian EAR approximation with respect to $\mathbf{x} \in \mathbb{R}^d$ decomposes exactly into the margin term T_1 and the covariance term T_2:*

$$\nabla_{\mathbf{x}} \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) = T_1 + T_2. \tag{66}$$

where $T_1 \in \mathbb{R}^d$ and $T_2 \in \mathbb{R}^d$ are explicitly given by:

$$T_1 = -\sqrt{n} \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\nabla p_{k^\star, r}(\mathbf{x}) - \nabla p_{j,r}(\mathbf{x})\right), \tag{67}$$

$$(T_2)_\ell = -\operatorname{tr}\left(\frac{1}{2} \mathbf{\Sigma}_V^{-1} \big(\mathbf{M} - \Phi_m \mathbf{\Sigma}_V\big) \mathbf{\Sigma}_V^{-1} \frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell}\right), \quad \textit{for } \ell = 1, \ldots, d. \tag{68}$$

Here, $\mathbf{M} = \mathbb{E}[\mathbf{V}\mathbf{V}^\top \mathbb{I}_{\{\mathbf{V} \le \mathbf{a}\}}]$ is the truncated second moment matrix of $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_V)$.

Proof of lemma 4. By the multivariate chain rule, the total derivative of $\Phi_m\big(\mathbf{a}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\big)$ with respect to the ℓ-th spatial coordinate x_ℓ expands as:

$$\frac{\partial \Phi_m}{\partial x_\ell} = \sum_{j=1}^{m} \frac{\partial \Phi_m}{\partial a_j} \frac{\partial a_j}{\partial x_\ell} + \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial \Phi_m}{\partial (\mathbf{\Sigma}_V)_{ij}} \frac{\partial (\mathbf{\Sigma}_V)_{ij}}{\partial x_\ell}. \tag{69}$$

Expressing this in vector-matrix form for the full spatial gradient $\nabla_{\mathbf{x}} \Phi_m \in \mathbb{R}^d$ yields the decomposition:

$$\nabla_{\mathbf{x}} \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \approx -\nabla_{\mathbf{x}} \Phi_m = \underbrace{-\left(\frac{\partial \mathbf{a}}{\partial \mathbf{x}}\right)^\top \frac{\partial \Phi_m}{\partial \mathbf{a}}}_{T_1} \underbrace{- \begin{pmatrix} \operatorname{tr}\left(\frac{\partial \Phi_m}{\partial \mathbf{\Sigma}_V} \frac{\partial \mathbf{\Sigma}_V}{\partial x_1}\right) \\ \vdots \\ \operatorname{tr}\left(\frac{\partial \Phi_m}{\partial \mathbf{\Sigma}_V} \frac{\partial \mathbf{\Sigma}_V}{\partial x_d}\right) \end{pmatrix}}_{T_2}. \tag{70}$$

Derivation of T_1: Given the integration limit vector $\mathbf{a}(\mathbf{x}) = \sqrt{n}\boldsymbol{\Delta}(\mathbf{x})$, its spatial Jacobian matrix evaluates to:

$$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} = \sqrt{n}\left((\nabla p_{k^\star,r} - \nabla p_{j,r})^\top\right)_{j\in J} \in \mathbb{R}^{m\times d}. \tag{71}$$

Applying Leibniz's integral rule to the multidimensional integral $\Phi_m(\mathbf{a}) = \int_{-\infty}^{a_1}\cdots\int_{-\infty}^{a_m}\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v}$, the partial derivative with respect to a_j isolates the j-th marginal:

$$\frac{\partial \Phi_m}{\partial a_j} = \int_{-\infty}^{a_1}\cdots\int_{-\infty}^{a_{j-1}}\int_{-\infty}^{a_{j+1}}\cdots\int_{-\infty}^{a_m}\phi_m(a_j,\mathbf{v}_{-j};\boldsymbol{\Sigma}_V)\,d\mathbf{v}_{-j}, \tag{72}$$

where $\mathbf{v}_{-j} = (v_1,\ldots,v_{j-1},v_{j+1},\ldots,v_m)$ denotes the vector of integration variables excluding v_j . Multiplying the transposed Jacobian by this gradient vector directly yields T_1:

$$T_1 = -\sqrt{n}\sum_{j\neq k^\star}\frac{\partial \Phi_m}{\partial a_j}\left(\nabla p_{k^\star,r}(\mathbf{x}) - \nabla p_{j,r}(\mathbf{x})\right). \tag{73}$$

Derivation of T_2: For the zero-mean multivariate Gaussian density $\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)$, its analytical derivative with respect to the covariance matrix is given by standard matrix calculus:

$$\frac{\partial \phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)}{\partial \boldsymbol{\Sigma}_V} = \frac{1}{2}\boldsymbol{\Sigma}_V^{-1}\big(\mathbf{v}\mathbf{v}^\top - \boldsymbol{\Sigma}_V\big)\boldsymbol{\Sigma}_V^{-1}\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V). \tag{74}$$

Since the integration limit $\mathbf{a}$ is independent of $\boldsymbol{\Sigma}_V$, we can differentiate under the integral sign. Substituting the derivative of the PDF into the integral yields:

$$\begin{aligned}\frac{\partial \Phi_m}{\partial \boldsymbol{\Sigma}_V} &= \int_{\mathbf{v}\le\mathbf{a}}\frac{\partial \phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)}{\partial \boldsymbol{\Sigma}_V}\,d\mathbf{v}\\ &= \int_{\mathbf{v}\le\mathbf{a}}\left[\frac{1}{2}\boldsymbol{\Sigma}_V^{-1}\big(\mathbf{v}\mathbf{v}^\top - \boldsymbol{\Sigma}_V\big)\boldsymbol{\Sigma}_V^{-1}\right]\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v}.\end{aligned} \tag{75}$$

Because the matrices $\boldsymbol{\Sigma}_V$ and $\boldsymbol{\Sigma}_V^{-1}$ are constant with respect to the integration variable $\mathbf{v}$, we can distribute the terms and factor these constant matrices out of the integrals:

$$\begin{aligned}\frac{\partial \Phi_m}{\partial \boldsymbol{\Sigma}_V} &= \frac{1}{2}\boldsymbol{\Sigma}_V^{-1}\left(\int_{\mathbf{v}\le\mathbf{a}}\mathbf{v}\mathbf{v}^\top\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v} - \int_{\mathbf{v}\le\mathbf{a}}\boldsymbol{\Sigma}_V\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v}\right)\boldsymbol{\Sigma}_V^{-1}\\ &= \frac{1}{2}\boldsymbol{\Sigma}_V^{-1}\left(\int_{\mathbf{v}\le\mathbf{a}}\mathbf{v}\mathbf{v}^\top\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v} - \boldsymbol{\Sigma}_V\int_{\mathbf{v}\le\mathbf{a}}\phi_m(\mathbf{v};\boldsymbol{\Sigma}_V)\,d\mathbf{v}\right)\boldsymbol{\Sigma}_V^{-1}.\end{aligned} \tag{76}$$

By definition, the first integral is the truncated second uncentered moment matrix $\mathbf{M} = \mathbb{E}\big[\mathbf{V}\mathbf{V}^\top\mathbb{I}_{\{\mathbf{V}\le\mathbf{a}\}}\big]$, and the second integral is exactly the CDF Φ_m. Substituting these definitions simplifies the expression to:

$$\frac{\partial \Phi_m}{\partial \boldsymbol{\Sigma}_V} = \frac{1}{2}\boldsymbol{\Sigma}_V^{-1}\big(\mathbf{M} - \Phi_m\boldsymbol{\Sigma}_V\big)\boldsymbol{\Sigma}_V^{-1}. \tag{77}$$

Finally, we compute the spatial gradient via the matrix chain rule. Applying the trace identity $\sum_{i,j}\frac{\partial f}{\partial Y_{ij}}\frac{\partial Y_{ij}}{\partial x_\ell} = \operatorname{tr}\left(\left(\frac{\partial f}{\partial \mathbf{Y}}\right)^\top\frac{\partial \mathbf{Y}}{\partial x_\ell}\right)$ to the symmetric covariance matrix $\boldsymbol{\Sigma}_V$, the ℓ-th component of the T_2 dynamics is exactly established:

$$(T_2)_\ell = -\operatorname{tr}\left(\frac{\partial \Phi_m}{\partial \boldsymbol{\Sigma}_V}\frac{\partial \boldsymbol{\Sigma}_V}{\partial x_\ell}\right). \tag{78}$$

□

C.2.2 SPATIAL DERIVATIVE OF THE COVARIANCE STRUCTURE

To fully compute T_2, we provide the explicit spatial derivative of the multinomial covariance matrix $\mathbf{\Sigma}_V(\mathbf{x})$.

Lemma 5 (Derivatives of the Covariance Matrix)**.** *Assume that the dominant class $k^\star$ is unique and locally fixed in a neighborhood of $\mathbf{x}$. Let $J = \{1, \dots, K\} \setminus \{k^\star\}$, $m = K - 1$, and*

$$\mathbf{p}_r := \big(p_{j,r}\big)_{j \in J} \in \mathbb{R}^m,$$

and let $\mathbf{1} \in \mathbb{R}^m$ denote the all-ones vector. Rows and columns of all matrices below are indexed by J. The spatial derivative of $\mathbf{\Sigma}_V(\mathbf{x})$ decomposes over the local class masses as

$$\frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell} = \sum_{k=1}^{K} \frac{\partial p_{k,r}}{\partial x_\ell} \mathbf{A}_k, \tag{79}$$

where

$$\mathbf{A}_k(\mathbf{x}) := \frac{\partial \mathbf{\Sigma}_V}{\partial p_{k,r}}(\mathbf{x}).$$

For the dominant class $k = k^\star$,

$$\big(\mathbf{A}_{k^\star}\big)_{ij} = 1 - 2p_{k^\star,r} + p_{i,r} + p_{j,r}, \qquad i, j \in J. \tag{80}$$

For each competing class $k \in J$,

$$\big(\mathbf{A}_k\big)_{ij} = \begin{cases} 1 - 2p_{k,r} + 2p_{k^\star,r}, & i = j = k, \\ -p_{j,r} + p_{k^\star,r}, & i = k,\ j \neq k, \\ -p_{i,r} + p_{k^\star,r}, & i \neq k,\ j = k, \\ 0, & \textit{otherwise}. \end{cases} \tag{81}$$

Proof of Lemma 5. The covariance matrix can be written as

$$\mathbf{\Sigma}_V(\mathbf{x}) = \operatorname{diag}(\mathbf{p}_r) - \mathbf{p}_r\mathbf{p}_r^\top + p_{k^\star,r}(1 - p_{k^\star,r})\mathbf{1}\mathbf{1}^\top + p_{k^\star,r}\left(\mathbf{p}_r\mathbf{1}^\top + \mathbf{1}\mathbf{p}_r^\top\right). \tag{82}$$

Since $k^\star$ is locally fixed, the coordinate system of $\mathbf{\Sigma}_V$ remains unchanged in a neighborhood of $\mathbf{x}$. We first differentiate with respect to the local class masses and then apply the chain rule with respect to x_ℓ.

For $k = k^\star$,

$$\begin{aligned} \frac{\partial \mathbf{\Sigma}_V}{\partial p_{k^\star,r}} &= \frac{\partial}{\partial p_{k^\star,r}} \left[p_{k^\star,r}(1 - p_{k^\star,r})\mathbf{1}\mathbf{1}^\top + p_{k^\star,r}\left(\mathbf{p}_r\mathbf{1}^\top + \mathbf{1}\mathbf{p}_r^\top\right)\right] \\ &= (1 - 2p_{k^\star,r})\mathbf{1}\mathbf{1}^\top + \mathbf{p}_r\mathbf{1}^\top + \mathbf{1}\mathbf{p}_r^\top. \end{aligned} \tag{83}$$

For $k \in J$, let $\mathbf{e}_k \in \mathbb{R}^m$ denote the standard basis vector corresponding to class k under the indexing by J. Then

$$\begin{aligned} \frac{\partial \mathbf{\Sigma}_V}{\partial p_{k,r}} &= \frac{\partial}{\partial p_{k,r}} \left[\operatorname{diag}(\mathbf{p}_r) - \mathbf{p}_r\mathbf{p}_r^\top + p_{k^\star,r}\left(\mathbf{p}_r\mathbf{1}^\top + \mathbf{1}\mathbf{p}_r^\top\right)\right] \\ &= \mathbf{e}_k\mathbf{e}_k^\top - \left(\mathbf{e}_k\mathbf{p}_r^\top + \mathbf{p}_r\mathbf{e}_k^\top\right) + p_{k^\star,r}\left(\mathbf{e}_k\mathbf{1}^\top + \mathbf{1}\mathbf{e}_k^\top\right). \end{aligned} \tag{84}$$

Finally, applying the chain rule gives

$$\frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell} = \frac{\partial p_{k^\star,r}}{\partial x_\ell}\mathbf{A}_{k^\star}(\mathbf{x}) + \sum_{k \in J} \frac{\partial p_{k,r}}{\partial x_\ell}\mathbf{A}_k(\mathbf{x}). \tag{85}$$

This completes the proof. □

Lemma 6 (Finite-sample and boundary-local T_1 dominance)**.** *Fix a query point* $\mathbf{x}$ *and sample size* n*. Assume that* $k^\star$ *is unique and locally fixed, the local class masses are continuously differentiable near* $\mathbf{x}$*, and* $\mathbf{\Sigma}_V(\mathbf{x})$ *is positive definite. Define*

$$\beta_j(\mathbf{x}, n) = \frac{\partial \Phi_m}{\partial a_j}\left(\sqrt{n}\,\mathbf{\Delta}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\right) > 0 \tag{86}$$

and the sensitivity-weighted margin direction

$$\mathbf{h}(\mathbf{x}, n) = \sum_{j \neq k^\star} \beta_j(\mathbf{x}, n) \nabla \Delta_j(\mathbf{x}). \tag{87}$$

Suppose that

$$\left\| \frac{\partial \Phi_m}{\partial \mathbf{\Sigma}_V} \right\|_F \leq C_\Phi, \qquad \max_{\ell \in \{1,\dots,d\}} \left\| \frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell} \right\|_F \leq C_\Sigma, \tag{88}$$

where the derivative of Φ_m *is evaluated at*

$$\left(\sqrt{n}\,\mathbf{\Delta}(\mathbf{x}); \mathbf{\Sigma}_V(\mathbf{x})\right).$$

If

$$\|\mathbf{h}(\mathbf{x}, n)\|_2 > 0, \tag{89}$$

then

$$\frac{\|T_2(\mathbf{x}, n)\|_2}{\|T_1(\mathbf{x}, n)\|_2} \leq \frac{\sqrt{d}\, C_\Phi C_\Sigma}{\sqrt{n}\, \|\mathbf{h}(\mathbf{x}, n)\|_2}. \tag{90}$$

In particular, if

$$\sqrt{n}\, \|\mathbf{h}(\mathbf{x}, n)\|_2 > \sqrt{d}\, C_\Phi C_\Sigma, \tag{91}$$

then

$$\|T_2(\mathbf{x}, n)\|_2 < \|T_1(\mathbf{x}, n)\|_2. \tag{92}$$

Assume additionally that the sequence remains in the nontrivial-EAR boundary layer, so that the closest standardized margin is bounded: $0 \leq \min_{j \neq k^\star} \sqrt{n}\, \Delta_j(\mathbf{x}_n) \leq C_\Delta$ *for all sufficiently large* n*. Assume that the bounds in Eq. 88 hold uniformly along this sequence, with* C_Φ *and* C_Σ *independent of* n*, and that the sensitivity-weighted margin direction is uniformly non-degenerate:*

$$\|\mathbf{h}(\mathbf{x}_n, n)\|_2 \geq c_h \tag{93}$$

for some constant $c_h > 0$ *and all sufficiently large* n*. Then*

$$\frac{\|T_2(\mathbf{x}_n, n)\|_2}{\|T_1(\mathbf{x}_n, n)\|_2} \leq \frac{\sqrt{d}\, C_\Phi C_\Sigma}{c_h \sqrt{n}} = O(n^{-1/2}), \tag{94}$$

and consequently

$$\lim_{n \to \infty} \frac{\|T_2(\mathbf{x}_n, n)\|_2}{\|T_1(\mathbf{x}_n, n)\|_2} = 0. \tag{95}$$

Proof. From the Gaussian risk-gradient decomposition,

$$T_1(\mathbf{x}, n) = -\sqrt{n} \sum_{j \neq k^\star} \beta_j(\mathbf{x}, n) \nabla \Delta_j(\mathbf{x}) = -\sqrt{n}\, \mathbf{h}(\mathbf{x}, n). \tag{96}$$

Therefore,

$$\|T_1(\mathbf{x}, n)\|_2 = \sqrt{n}\, \|\mathbf{h}(\mathbf{x}, n)\|_2. \tag{97}$$

For each spatial coordinate x_ℓ, the covariance contribution satisfies

$$(T_2)_\ell = -\operatorname{tr}\left[\left(\frac{\partial \Phi_m}{\partial \mathbf{\Sigma}_V} \right)^\top \frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell} \right]. \tag{98}$$

By the Frobenius Cauchy–Schwarz inequality,

$$|(T_2)_\ell| \leq \left\| \frac{\partial \Phi_m}{\partial \mathbf{\Sigma}_V} \right\|_F \left\| \frac{\partial \mathbf{\Sigma}_V}{\partial x_\ell} \right\|_F \tag{99}$$

$$\leq C_\Phi C_\Sigma. \tag{100}$$

Summing over the d input coordinates gives

$$\|T_2(\mathbf{x}, n)\|_2 \le \sqrt{d}\, C_\Phi C_\Sigma. \tag{101}$$

Combining Eqs. 97 and 101, and using $\|\mathbf{h}(\mathbf{x}, n)\|_2 > 0$, yields

$$\frac{\|T_2(\mathbf{x}, n)\|_2}{\|T_1(\mathbf{x}, n)\|_2} \le \frac{\sqrt{d}\, C_\Phi C_\Sigma}{\sqrt{n}\, \|\mathbf{h}(\mathbf{x}, n)\|_2}, \tag{102}$$

which proves Eq. 90. If Eq. 91 holds, the right-hand side is strictly smaller than one, and hence

$$\|T_2(\mathbf{x}, n)\|_2 < \|T_1(\mathbf{x}, n)\|_2.$$

We next consider the boundary-local sequence $\{\mathbf{x}_n\}$. By Eq. 93,

$$\|T_1(\mathbf{x}_n, n)\|_2 = \sqrt{n}\, \|\mathbf{h}(\mathbf{x}_n, n)\|_2 \ge c_h \sqrt{n}. \tag{103}$$

The uniform covariance-sensitivity bounds similarly imply

$$\|T_2(\mathbf{x}_n, n)\|_2 \le \sqrt{d}\, C_\Phi C_\Sigma. \tag{104}$$

Consequently,

$$\frac{\|T_2(\mathbf{x}_n, n)\|_2}{\|T_1(\mathbf{x}_n, n)\|_2} \le \frac{\sqrt{d}\, C_\Phi C_\Sigma}{c_h \sqrt{n}} = O(n^{-1/2}), \tag{105}$$

which proves Eqs. 94 and 95. □

C.3 Conditional label preservation under EAR-gradient flow

The EAR descent direction is defined relative to the locally dominant class. If an update crosses a class boundary, the reference class changes, and subsequent EAR descent may reinforce assignment to the competing class. Label preservation is therefore required for interpreting EAR-gradient descent as purification.

Consider a continuously differentiable path $\mathbf{x}(s)$ such that $k^\star(\mathbf{x}(s)) = k_0$ for $s < 0$, $\mathbf{x}(0) = \mathbf{x}_c$, and the path enters a competing class region for $s > 0$. All margins and Gaussian sensitivities below are evaluated from the k_0 side of the boundary.

Definition C.1 (Active and non-active sets)**.** *The active set at $\mathbf{x}_c$, relative to the approaching class k_0, is*

$$\mathcal{M}(\mathbf{x}_c) = \{m \ne k_0 \mid p_{m,r}(\mathbf{x}_c) = p_{k_0,r}(\mathbf{x}_c)\},$$

whereas the non-active set is

$$\mathcal{N}(\mathbf{x}_c) = \{\ell \ne k_0 \mid p_{\ell,r}(\mathbf{x}_c) < p_{k_0,r}(\mathbf{x}_c)\}.$$

Thus, $\Delta_m(\mathbf{x}_c) = 0$ for $m \in \mathcal{M}$, while $\Delta_\ell(\mathbf{x}_c) > 0$ for $\ell \in \mathcal{N}$.

The crossing is transverse if, for some $m \in \mathcal{M}$,

$$\frac{d}{ds}\Delta_m(\mathbf{x}(s))\bigg|_{s=0^-} = \nabla\Delta_m(\mathbf{x}_c)^\top \dot{\mathbf{x}}(0) < 0.$$

That is, the path has a nonzero component normal to at least one active boundary and moves from the k_0 region toward the corresponding competing class.

Assumption 2 (Landscape smoothness)**.** *The local class-mass functions $p_{k,r}(\mathbf{x})$ are continuously differentiable with locally Lipschitz gradients. Within any compact neighborhood $\mathcal{U}$ where the dominant class is fixed and $\mathbf{\Sigma}_V(\mathbf{x})$ remains positive definite, the margin gradients are bounded and $\nabla\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ is locally Lipschitz. Consequently, the gradient flow*

$$\frac{d\mathbf{x}}{dt} = -\nabla\widetilde{P}_{\mathrm{adv}}(\mathbf{x}(t))$$

admits a unique local solution for every initial point in $\mathcal{U}$.

Lemma 7 (Conditional boundary repulsion)**.** *Let $\mathbf{x}(s)$ approach a decision boundary from the region of class k_0. Assume local T_1 dominance, Assumption 2, and the non-degenerate weighted crossing condition*

$$\limsup_{s\to 0^-} \sum_{m\in\mathcal{M}} \beta_m(\mathbf{x}(s), n) \frac{d\Delta_m(\mathbf{x}(s))}{ds} < 0.$$

Then, for sufficiently large n,

$$\lim_{s\to 0^-} \frac{d}{ds} \widetilde{P}_{\mathrm{adv}}(\mathbf{x}(s)) > 0.$$

For a boundary with a single active competitor, the weighted condition is equivalent to transverse crossing because $\beta_m > 0$.

Proof. For brevity, let

$$\dot{\Delta}_\ell(s) := \frac{d}{ds}\Delta_\ell(\mathbf{x}(s)).$$

By the EAR-gradient decomposition,

$$\begin{aligned}\frac{d}{ds}\widetilde{P}_{\mathrm{adv}}(\mathbf{x}(s)) = &-\sqrt{n}\sum_{m\in\mathcal{M}} \beta_m(\mathbf{x}(s), n)\dot{\Delta}_m(s) \\ &- \sqrt{n}\sum_{\ell\in\mathcal{N}} \beta_\ell(\mathbf{x}(s), n)\dot{\Delta}_\ell(s) + T_2^\top \dot{\mathbf{x}}(s). \end{aligned} \tag{106}$$

For every $\ell \in \mathcal{N}$, $\Delta_\ell(\mathbf{x}_c) > 0$. Hence

$$\sqrt{n}\,\Delta_\ell(\mathbf{x}_c) \to \infty$$

as $n \to \infty$, and the corresponding Gaussian sensitivity β_ℓ decays exponentially. Since the margin derivatives are bounded, the non-active contribution in Eq. 106 is asymptotically negligible.

For $m \in \mathcal{M}$, the active margins satisfy $\Delta_m(\mathbf{x}_c) = 0$, so their Gaussian sensitivities remain non-vanishing. Together with local T_1 dominance, this gives the compact boundary expansion

$$\frac{d}{ds}\widetilde{P}_{\mathrm{adv}}(\mathbf{x}(s)) = -\sqrt{n}\sum_{m\in\mathcal{M}} \beta_m(\mathbf{x}(s), n)\dot{\Delta}_m(s) + o(\sqrt{n}).$$

The weighted crossing condition therefore implies, for sufficiently large n,

$$\lim_{s\to 0^-} \frac{d}{ds} \widetilde{P}_{\mathrm{adv}}(\mathbf{x}(s)) > 0.$$

□

Corollary 1 (Conditional local label preservation)**.** *Suppose the conditions of Lemma 7 hold in a neighborhood of a class boundary. Then the continuous negative-gradient flow*

$$\frac{d\mathbf{x}}{dt} = -\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}(t))$$

cannot cross that boundary transversely while these conditions remain valid.

Proof. Along the negative-gradient flow,

$$\frac{d}{dt}\widetilde{P}_{\mathrm{adv}}(\mathbf{x}(t)) = -\left\|\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}(t))\right\|_2^2 \le 0.$$

By Lemma 7, a transverse crossing satisfying the weighted crossing condition would instead require the approximate EAR to increase as the trajectory approaches the boundary. These two directional behaviors are incompatible. Hence, the flow cannot cross the boundary while the stated conditions remain valid. □

Empirical Visualization of Semantic Preservation To empirically illustrate label preservation in typical settings and the pathological edge cases where label flipping occurs, we construct a 2D Gaussian Mixture Model (GMM) with $K = 3$ classes. We simulate the purification trajectories driven by the negative gradient of our intrinsic risk, $-\nabla \widetilde{P}_{\text{adv}}(\mathbf{x})$.

Experimental Setup. The components of the GMM share the following spatial means: $\boldsymbol{\mu}_0 = [2.5, 0.0]^\top$, $\boldsymbol{\mu}_1 = [-2.5, 0.0]^\top$, and $\boldsymbol{\mu}_2 = [3.0, 0.0]^\top$. Note that Class 2 is positioned very close to Class 0 to act as a potential adversarial trap. We draw $N = 25$ random samples within a predefined vulnerable bounding box ($x_1 \in [1.0, 4.5], x_2 \in [-1.5, 1.5]$) and optimize them using gradient descent with a learning rate of 0.08 for 40 steps. We compare a representative imbalanced scenario, where the local boundary geometry remains regular, with an intentionally pathological scenario in which the sufficient conditions for label preservation can fail:

- **Standard Imbalance (Figure 6a):** A case with moderate prior mismatch and variance difference. Priors: $\alpha_0 = \alpha_1 = 0.05$, $\alpha_2 = 0.90$. Covariances: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = 5.0\mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = 0.5\mathbf{I}_2$..
- **Pathological Extreme (Figure 6b):** An artificially severe distortion designed to break the semantic preservation. Priors: $\alpha_0 = \alpha_1 = 0.0005$, $\alpha_2 = 0.999$. Covariances: $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = 10.0\mathbf{I}_2$, $\boldsymbol{\Sigma}_2 = 0.2\mathbf{I}_2$..

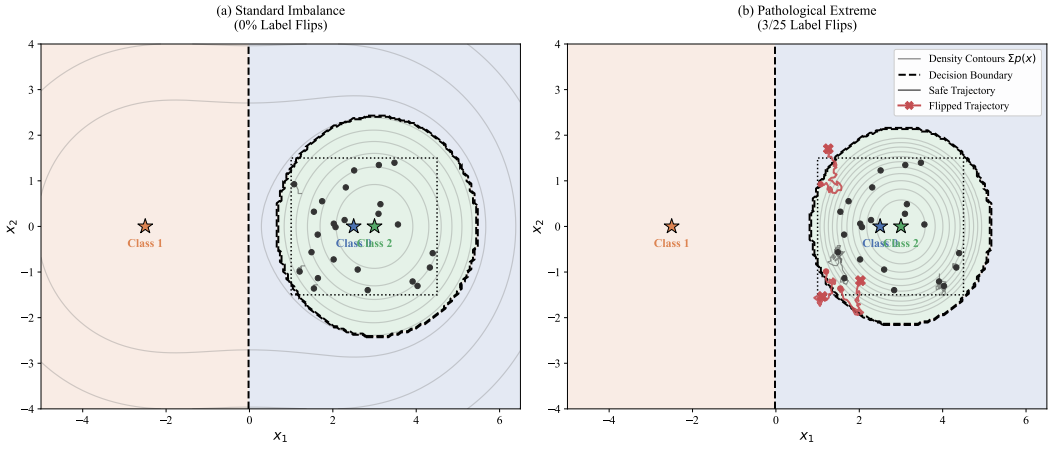


Figure 6: **Purification trajectories and density contours under standard vs. pathological imbalance.** The background colors indicate Bayesian decision regions, and the gray lines represent the global probability density contours $\sum_k p_k(\mathbf{x})$. **(a) Standard Imbalance:** The density landscape maintains distinct local modes. Trajectories safely converge without crossing the decision boundary (0% label flips). **(b) Pathological Extreme:** Extreme variance and prior distortions severely skew the global density geometry, pulling a fraction of trajectories across the boundary (marked in red with 'X'). This confirms that over-purification strictly requires highly pathological distributions.

C.4 A SUFFICIENT CONDITION FOR LOCAL SEPARATION

We use the existence of a Bayes decision boundary on the open segment between two class means as a tractable geometric proxy for local class separation. We next derive a sufficient condition for this inter-mean boundary to exist.

Step 1: The ideal decision boundary. The optimal Bayes decision boundary between class i and class j satisfies:

$$\log\left(\pi_i p(\mathbf{x}|i)\right) = \log\left(\pi_j p(\mathbf{x}|j)\right). \tag{107}$$

Expanding the log-likelihood of the multidimensional Gaussian distributions, we have:

$$\log \pi_i - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$
$$= \log \pi_j - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j).$$

Canceling the constant terms and rearranging the quadratic forms, the boundary equation becomes:

$$(\mathbf{x}-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) - (\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = 2\log\left(\frac{\pi_i}{\pi_j}\right) + \log\left(\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}\right) \triangleq C. \quad (108)$$

Step 2: Intersection on the connecting line. Let $\boldsymbol{\Delta}_{ij} = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i$ and parameterize the line segment between the two means as

$$\mathbf{x}(t) = \boldsymbol{\mu}_i + t\boldsymbol{\Delta}_{ij}, \qquad t \in [0,1].$$

Then

$$\mathbf{x}(t) - \boldsymbol{\mu}_i = t\boldsymbol{\Delta}_{ij}, \qquad \mathbf{x}(t) - \boldsymbol{\mu}_j = (t-1)\boldsymbol{\Delta}_{ij}.$$

Substituting this parameterization into Eq. 108 gives

$$(t-1)^2 \boldsymbol{\Delta}_{ij}^\top \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Delta}_{ij} - t^2 \boldsymbol{\Delta}_{ij}^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{ij} = C.$$

Let

$$D_j^2 = \boldsymbol{\Delta}_{ij}^\top \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Delta}_{ij}, \qquad D_i^2 = \boldsymbol{\Delta}_{ij}^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_{ij}.$$

Define

$$f(t) = (t-1)^2 D_j^2 - t^2 D_i^2 - C.$$

A sufficient geometric proxy for a low-density valley is that the Bayes boundary intersects the open segment between the two means, i.e., there exists $t^* \in (0,1)$ such that $f(t^*) = 0$.

We have

$$f(0) = D_j^2 - C, \qquad f(1) = -D_i^2 - C.$$

A sufficient condition for $f(0) > 0$ and $f(1) < 0$ is

$$D_j^2 > C, \qquad D_i^2 > -C,$$

which is guaranteed by

$$\min(D_i^2, D_j^2) > |C|.$$

It therefore suffices to show that the Bayes boundary intersects the open segment between the two means, i.e., that there exists $t^\star \in (0,1)$ such that $f(t^\star) = 0$.

Step 3: Bounding with the Rayleigh quotient. To express this condition in Euclidean distance, we use the Rayleigh quotient. For any component $k \in \{i, j\}$,

$$D_k^2 = \boldsymbol{\Delta}_{ij}^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\Delta}_{ij} \geq \lambda_{\min}(\boldsymbol{\Sigma}_k^{-1}) \|\boldsymbol{\Delta}_{ij}\|^2 = \frac{\|\boldsymbol{\Delta}_{ij}\|^2}{\lambda_{\max}(\boldsymbol{\Sigma}_k)}.$$

Let

$$\hat{\sigma}^2_{\max} = \max_{k \in \{i,j\}} \lambda_{\max}(\boldsymbol{\Sigma}_k).$$

Then

$$\min(D_i^2, D_j^2) \geq \frac{\|\boldsymbol{\Delta}_{ij}\|^2}{\hat{\sigma}^2_{\max}}.$$

Thus, a simple sufficient condition for the intersection is

$$\frac{\|\boldsymbol{\Delta}_{ij}\|^2}{\hat{\sigma}^2_{\max}} > |C|.$$

Equivalently,

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 > \hat{\sigma}^2_{\max} |C|.$$

Substituting the definition of C in Eq. 108 yields

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2 > \hat{\sigma}^2_{\max} \left| \log\left(\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} \frac{\pi_i^2}{\pi_j^2} \right) \right|.$$

Therefore, the derived condition guarantees that the Bayes boundary intersects the open segment between the two class means. It provides a tractable geometric proxy for local separation and motivated the local class-conditional dominance condition used below.

C.5 Proof of Theorem 2: adversarial purification conditions

This section provides the proof of Theorem 2. We derive sufficient local alignment conditions under which ascending empirical density gradients can decrease the intrinsic adversarial risk $\widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ to first order.

Assumption 3 (Local class-conditional dominance)**.** *In a neighborhood of* $\mathbf{x}$*, the posterior-dominant class* $k^\star$ *is unique and locally fixed. We consider local configurations where the competition is dominated by* $k^\star$ *and each influential competitor* j*, while contributions from the remaining classes are comparatively small in the density-gradient expansion.*

Proof of Theorem 2. We analyze the first-order alignment of the two density-guided directions $\mathbf{g} \in \{\nabla p(\mathbf{x}), \nabla M(\mathbf{x})\}$ with the intrinsic adversarial risk gradient $\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x})$. A guidance direction locally decreases the adversarial risk to first order when $\left\langle \mathbf{g}, \nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \right\rangle < 0$.

1. Preliminaries and Common Gradient Forms.

First, let's recall the definitions for a Gaussian Mixture Model. The component density is a multivariate Gaussian:

$$q_k(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right). \tag{109}$$

The local score, which is the gradient of the log-density, points towards the component mean:

$$s_k(\mathbf{x}) \triangleq \nabla \log q_k(\mathbf{x}) = -\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k). \tag{110}$$

Using the chain rule, the gradient of the component density is:

$$\nabla q_k(\mathbf{x}) = q_k(\mathbf{x}) \nabla \log q_k(\mathbf{x}) = q_k(\mathbf{x}) s_k(\mathbf{x}). \tag{111}$$

The gradient of the marginal density $p(\mathbf{x}) = \sum_{i=1}^K \pi_i q_i(\mathbf{x})$ is therefore:

$$\nabla p(\mathbf{x}) = \sum_{i=1}^K \pi_i \nabla q_i(\mathbf{x}) = \sum_{i=1}^K \pi_i q_i(\mathbf{x}) s_i(\mathbf{x}). \tag{112}$$

Let $m := K-1$ denote the number of competing classes. Given the dominance of the T_1 term, the intrinsic adversarial gradient $\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x})$ can be approximated as follows:

$$\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \approx -C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\nabla(\pi_{k^\star} q_{k^\star}(\mathbf{x})) - \nabla(\pi_j q_j(\mathbf{x}))\right), \tag{113}$$

where C is a positive constant proportional to $\sqrt{n}$. Formally, each weight $\partial \Phi_m / \partial a_j$ is the marginal probability density of the multivariate normal distribution with respect to the j-th class margin, calculated as:

$$\frac{\partial \Phi_m}{\partial a_j} = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_{j-1}} \int_{-\infty}^{a_{j+1}} \cdots \int_{-\infty}^{a_m} \phi_m(a_j, \mathbf{v}_{-j}; \mathbf{\Sigma}_V)\, d\mathbf{v}_{-j}. \tag{114}$$

This term quantifies how sensitive the overall adversarial probability is to a change in the margin of class j. We define the normalized adversarial sensitivity weight $\tilde{w}_j = \frac{\partial \Phi_m / \partial a_j}{\sum_{\ell \neq k^\star} \partial \Phi_m / \partial a_\ell}, \qquad j \neq k^\star$. Since the Gaussian density is positive, $\partial \Phi_m / \partial a_j > 0$, and therefore $\tilde{w}_j > 0, \sum_{j \neq k^\star} \tilde{w}_j = 1$.

2. Condition for Marginal Density Purification ($\mathbf{g} = \nabla p(\mathbf{x})$).

We want to find the condition for $\langle \nabla p(\mathbf{x}), -\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \rangle > 0$. Substituting Eqs. 112 and 113, and moving constants and the outer summation outside the inner product, we get:

$$
\begin{aligned}
&\langle \nabla p(\mathbf{x}), \nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \rangle \\
&\approx C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left\langle \sum_{i=1}^{K} \pi_i q_i(\mathbf{x}) s_i(\mathbf{x}), \pi_j q_j(\mathbf{x}) s_j(\mathbf{x}) - \pi_{k^\star} q_{k^\star}(\mathbf{x}) s_{k^\star}(\mathbf{x}) \right\rangle \\
&= C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left\langle \pi_{k^\star} q_{k^\star} s_{k^\star} + \sum_{i \neq k^\star} \pi_i q_i s_i, \pi_j q_j s_j - \pi_{k^\star} q_{k^\star} s_{k^\star} \right\rangle \\
&= C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\langle \pi_{k^\star} q_{k^\star} s_{k^\star} + \pi_j q_j s_j + \sum_{i \neq \{k^\star, j\}} \pi_i q_i s_i, \pi_j q_j s_j - \pi_{k^\star} q_{k^\star} s_{k^\star} \rangle \right) \\
&= C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\|\pi_j q_j s_j\|^2 - \|\pi_{k^\star} q_{k^\star} s_{k^\star}\|^2 + \sum_{i \neq \{k^\star, j\}} \langle \pi_i q_i s_i, \pi_j q_j s_j - \pi_{k^\star} q_{k^\star} s_{k^\star} \rangle \right). \quad (115)
\end{aligned}
$$

Motivated by the separation condition in Appendix C.4, Assumption 3 treats the dominant class and its influential competitors as the leading contributors to the local density gradient. Accordingly, the marginal alignment is approximated by

$$
\langle \nabla p(\mathbf{x}), \nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \rangle \approx C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\|\pi_j q_j(\mathbf{x}) s_j(\mathbf{x})\|^2 - \|\pi_{k^\star} q_{k^\star}(\mathbf{x}) s_{k^\star}(\mathbf{x})\|^2 \right). \quad (116)
$$

For purification to be successful, this expression must be negative. Since $C > 0$ and $\partial \Phi_m / \partial a_j > 0$:

$$
\sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \left(\|\pi_j q_j s_j\|^2 - \|\pi_{k^\star} q_{k^\star} s_{k^\star}\|^2 \right) < 0. \quad (117)
$$

Rearranging the terms, we get:

$$
\sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \|\pi_{k^\star} q_{k^\star} s_{k^\star}\|^2 > \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \|\pi_j q_j s_j\|^2. \quad (118)
$$

Using the normalized weights $\tilde{w}_j$, this becomes

$$
\left(\sum_{\ell \neq k^\star} \frac{\partial \Phi_m}{\partial a_l} \right) \|\pi_{k^\star} q_{k^\star} s_{k^\star}\|^2 > \sum_{j \neq k^\star} \left(\sum_{\ell \neq k^\star} \frac{\partial \Phi_m}{\partial a_l} \right) \tilde{w}_j \|\pi_j q_j s_j\|^2. \quad (119)
$$

Canceling the positive constant term $\sum_{l \neq k^\star} \partial \Phi_m / \partial a_l$ from both sides yields:

$$
\|\pi_{k^\star} q_{k^\star}(\mathbf{x}) s_{k^\star}(\mathbf{x})\|^2 > \sum_{j \neq k^\star} \tilde{w}_j \|\pi_j q_j(\mathbf{x}) s_j(\mathbf{x})\|^2. \quad (120)
$$

Since $q_i(\mathbf{x})$ and π_i are scalars, we can expand the squared norm $\|\pi q s\|^2 = \pi^2 q(\mathbf{x})^2 \|s(\mathbf{x})\|^2$. This gives us the final condition for marginal density purification:

$$
\pi_{k^\star}^2 q_{k^\star}^2(\mathbf{x}) \|s_{k^\star}(\mathbf{x})\|^2 > \sum_{j \neq k^\star} \tilde{w}_j \pi_j^2 q_j^2(\mathbf{x}) \|s_j(\mathbf{x})\|^2. \quad (121)
$$

This is precisely the condition stated in Eq. 8. The proof for maximum conditional density purification follows a similar derivation.

3. Condition for Maximum-Conditional Density Purification ($\mathbf{g} = \nabla M(\mathbf{x})$).

Maximum-conditional purification selects

$$
k_M(\mathbf{x}) = \arg\max_i q_i(\mathbf{x}), \; M(\mathbf{x}) = q_{k_M(\mathbf{x})}(\mathbf{x}). \quad (122)
$$

At a point where $k_M(\mathbf{x})$ is unique and locally fixed,

$$
\nabla M(\mathbf{x}) = \nabla q_{k_M}(\mathbf{x}) = q_{k_M}(\mathbf{x}) s_{k_M}(\mathbf{x}). \quad (123)
$$

Using Eq. 113, we obtain

$$\left\langle \nabla M(\mathbf{x}), -\nabla \widetilde{P}_{\mathrm{adv}}(\mathbf{x}) \right\rangle$$
$$\approx C \sum_{j \neq k^\star} \frac{\partial \Phi_m}{\partial a_j} \Big[\pi_{k^\star} q_{k_M}(\mathbf{x}) q_{k^\star}(\mathbf{x}) \left\langle s_{k_M}(\mathbf{x}), s_{k^\star}(\mathbf{x}) \right\rangle - \pi_j q_{k_M}(\mathbf{x}) q_j(\mathbf{x}) \left\langle s_{k_M}(\mathbf{x}), s_j(\mathbf{x}) \right\rangle \Big].$$

Because $C > 0$ and $\partial \Phi_m / \partial a_j > 0$, this inner product is negative when the sensitivity-weighted expression inside the brackets is positive. Using

$$\tilde{w}_j = \frac{\partial \Phi_m / \partial a_j}{\sum_{\ell \neq k^\star} \partial \Phi_m / \partial a_\ell}, \qquad \sum_{j \neq k^\star} \tilde{w}_j = 1,$$

we obtain the general alignment condition

$$\pi_{k^\star} q_{k_M}(\mathbf{x}) q_{k^\star}(\mathbf{x}) \left\langle s_{k_M}(\mathbf{x}), s_{k^\star}(\mathbf{x}) \right\rangle > \sum_{j \neq k^\star} \tilde{w}_j \pi_j q_{k_M}(\mathbf{x}) q_j(\mathbf{x}) \left\langle s_{k_M}(\mathbf{x}), s_j(\mathbf{x}) \right\rangle . \tag{124}$$

This is precisely the general condition stated in Eq. 9.

With equal priors $k_M = k^\star$, the preceding condition reduces to

$$\pi_{k^\star} q_{k^\star}^2(\mathbf{x}) \| s_{k^\star}(\mathbf{x}) \|^2 > \sum_{j \neq k^\star} \tilde{w}_j \pi_j q_{k^\star}(\mathbf{x}) q_j(\mathbf{x}) \left\langle s_{k^\star}(\mathbf{x}), s_j(\mathbf{x}) \right\rangle , \tag{125}$$

which recovers Eq. 10. □

C.6 KDE-BASED EMPIRICAL VISUALIZATION OF GUIDANCE ALIGNMENT

Setup. We use the same 3-class two-dimensional GMM as in Fig. 3. The component means are $\mu_1 = (0, 0)$, $\mu_2 = (4, 1.5)$, and $\mu_3 = (4, -1.5)$. For the prior–variance heterogeneous setting, we use priors $(0.90, 0.05, 0.05)$ and variances $(1, 4, 4)$, and draw $n = 2000$ training samples from this distribution.

KDE estimation and empirical EAR. For each class, we fit a Gaussian KDE to estimate the class-conditional density $\hat{q}_k(\mathbf{x})$, with class-wise bandwidths selected by leave-one-out log-likelihood over a fixed multiplicative grid around Silverman's rule-of-thumb bandwidth. Following Definition 3.1, we approximate the local class masses by

$$\hat{p}_k(\mathbf{x}) = \hat{p}(\mathbf{x}, y = k) V_d(r),$$

with radius $r = 0.4$. For $K = 3$, we compute $\widehat{P}_{\mathrm{adv}}(\mathbf{x})$ using the bivariate Gaussian approximation of Theorem 1. The gradients $\nabla \hat{p}(\mathbf{x})$, $\nabla \hat{M}(\mathbf{x})$, and $\nabla \widehat{P}_{\mathrm{adv}}(\mathbf{x})$ are obtained by automatic differentiation on a grid.

Alignment statistics. We measure the angle between each density-guided direction $g(\mathbf{x}) \in \{\nabla \hat{p}(\mathbf{x}), \nabla \hat{M}(\mathbf{x})\}$ and the empirical EAR descent direction $-\nabla \widehat{P}_{\mathrm{adv}}(\mathbf{x})$:

$$\theta_g(\mathbf{x}) = \cos^{-1} \left(\frac{\langle g(\mathbf{x}), -\nabla \widehat{P}_{\mathrm{adv}}(\mathbf{x}) \rangle}{\|g(\mathbf{x})\| \, \|\nabla \widehat{P}_{\mathrm{adv}}(\mathbf{x})\|} \right).$$

Statistics are computed over grid points with nonzero gradient norms. The high-EAR region denotes the top 15% grid points ranked by $\widehat{P}_{\mathrm{adv}}(\mathbf{x})$.

The global acute ratios show that density-guided directions are mostly aligned with empirical EAR descent, while the larger angles in high-EAR regions indicate that misalignment is localized and most pronounced in high-risk regions.

D EXPERIMENTAL DETAILS

D.1 SYNTHETIC GMM.

We construct a 2D paired-overlap GMM with $K = 5$ classes and $L = 4$ components per class. The class anchors and their neighbor relationships are fixed across seeds; the component perturbations, variances, and mixture weights are sampled independently for each seed.

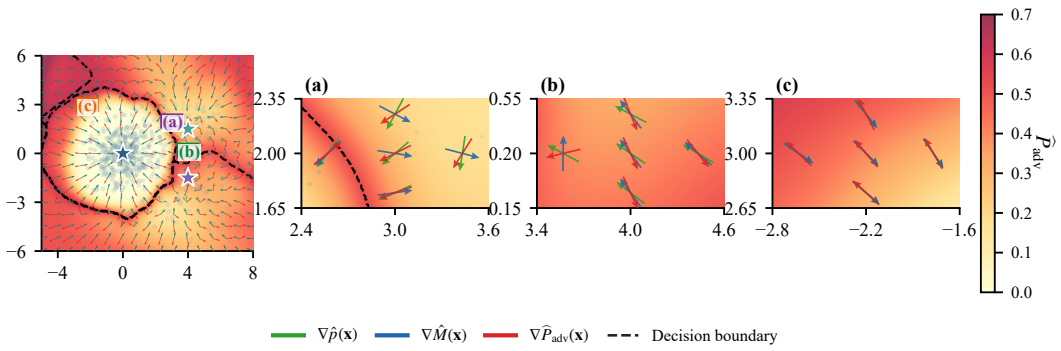


Figure 7: **KDE-estimated alignment of density guidance and EAR descent.** Panel (a) shows the full KDE-estimated field. Panels (b) and (c) enlarge boundaries between high- and low-prior components and between two low-prior components, respectively. Panel (d) shows a region where both density directions are aligned with EAR descent. Green, blue, and red arrows denote $\nabla\hat{p}$, $\nabla\hat{M}$, and $\nabla\widehat{P}_{\text{adv}}$, respectively; EAR descent points opposite to the red arrows. All arrows are normalized to emphasize direction rather than magnitude.

Table 3: **Empirical alignment statistics.** Angles are measured between each KDE-guided direction and empirical EAR descent $-\nabla\widehat{P}_{\text{adv}}$. The acute ratio is the fraction of grid points with angle smaller than 90°.

Region	Direction	Mean angle	Median angle	Acute ratio
Global	$\nabla\hat{p}$	43.1°	21.2°	0.820
Global	$\nabla\hat{M}$	32.4°	16.1°	0.907
High-EAR	$\nabla\hat{p}$	78.7°	67.0°	0.616
High-EAR	$\nabla\hat{M}$	71.3°	61.1°	0.659

1. **Define the class anchors and neighbors.** The five class anchors are fixed at equally spaced positions on a circle of radius 1.2:

$$\mathbf{c}_k = 1.2\begin{bmatrix}\cos(2\pi(k-1)/5)\\ \sin(2\pi(k-1)/5)\end{bmatrix}, \qquad k \in \{1,\dots,5\}.$$

Neighbor relationships follow this fixed circular order. For $k = 1,\dots,5$, define

$$\text{next}(k) = \begin{cases} k+1, & k<5,\\ 1, & k=5,\end{cases} \qquad \text{prev}(k) = \begin{cases} k-1, & k>1,\\ 5, & k=1.\end{cases}$$

Thus, for example, classes 5 and 2 are the two neighbors of class 1.

2. **Sample four component means per class.** We first define four deterministic reference locations:

$$\begin{aligned}\overline{\boldsymbol{\mu}}_{k,1} &= \mathbf{c}_k,\\ \overline{\boldsymbol{\mu}}_{k,2} &= 0.58\mathbf{c}_k + 0.42\mathbf{c}_{\text{next}(k)},\\ \overline{\boldsymbol{\mu}}_{k,3} &= 0.58\mathbf{c}_k + 0.42\mathbf{c}_{\text{prev}(k)},\\ \overline{\boldsymbol{\mu}}_{k,4} &= 0.72\mathbf{c}_k + 0.28\mathbf{c}_{\text{next}(k)}.\end{aligned}$$

The first reference location is the class anchor. The other three lie between the current anchor and one of its two fixed neighbors, creating cross-class overlap. We then add independent Gaussian perturbations:

$$\boldsymbol{\mu}_{k,\ell} = \overline{\boldsymbol{\mu}}_{k,\ell} + \boldsymbol{\epsilon}_{k,\ell},$$

where

$$\boldsymbol{\epsilon}_{k,1}, \boldsymbol{\epsilon}_{k,4} \sim \mathcal{N}(\mathbf{0}, 0.08^2\mathbf{I}), \qquad \boldsymbol{\epsilon}_{k,2}, \boldsymbol{\epsilon}_{k,3} \sim \mathcal{N}(\mathbf{0}, 0.05^2\mathbf{I}).$$

3. **Generate the component variances.** Each Gaussian has a diagonal covariance matrix. Its coordinate-wise variances are sampled independently from $[0.1160, 0.1570]$ for component 1, $[0.2754, 0.4050]$ for components 2 and 3, and $[0.2430, 0.3888]$ for component 4.
4. **Generate the component weights.** Each class has independently sampled mixture weights
$$\mathbf{w}_k \sim \mathrm{Dirichlet}(\mathbf{1}_4).$$
Class priors are uniform.
5. **Sample the data.** To sample from the class-conditional GMM $p(\mathbf{x} \mid y)$, we first draw a component index
$$z \mid y \sim \mathrm{Categorical}(\mathbf{w}_y),$$
and then sample
$$\mathbf{x} \mid y, z \sim \mathcal{N}(\boldsymbol{\mu}_{y,z}, \mathrm{diag}(\mathbf{v}_{y,z}))\,.$$
Marginalizing over z gives the desired mixture distribution:
$$p(\mathbf{x} \mid y) = \sum_{z=1}^{L} w_{y,z}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{y,z}, \mathrm{diag}(\mathbf{v}_{y,z}))\,.$$
For the complete dataset, the class label y is sampled from the uniform class prior.

Classifier and KDE. The downstream classifier is an MLP with hidden widths $(256, 256, 128)$ and SiLU activations. The KDE is constructed using 9,000 samples, with bandwidth $h = 0.2946344794$ and neighborhood radius $r = 0.12$. For the simplified EAR objective, we use continuity correction $c = 0.5$.

Neighborhood radius. We select the radius r on held-out validation data to maximize the accuracy of the fixed-radius nearest-neighbor reference rule among candidates satisfying $\mathrm{median}(n_{\mathrm{eff}}) \geq 8,\ Q_{0.1}(n_{\mathrm{eff}}) \geq 1.5$. The selected radius is shared across all objectives and attacks. Under the small-radius approximation, $p_{k,r}(\mathbf{x}) \approx V_d(r)p_k(\mathbf{x})$. Hence, replacing local class-mass gaps by joint-density gaps changes their gradients only by the common positive factor $V_d(r)$, which cancels under gradient normalization. Thus, r affects the reported EAR estimates but not the resulting joint-density-gap direction under this approximation.

Attacks. We first map each input coordinate to $[0, 1]$ using an attack box fitted to the training data. Its lower and upper bounds are the 0.5% and 99.5% coordinate-wise quantiles, respectively, expanded by 5% of their range. In this normalized space, both attacks use 10-step PGD with a random start, an ℓ_∞ perturbation budget of $\epsilon = 0.014$, and step size $\alpha = 0.00175$.

Purification and distillation. Each objective defines a 10-step normalized-gradient purifier g with step size $\eta = 0.01$. We distill it as
$$g_\theta^{\mathrm{dist}}(\mathbf{x}) = \mathbf{x} + R_\theta(\mathbf{x}),$$
where R_θ is a feed-forward MLP trained to predict $g(\mathbf{x}) - \mathbf{x}$. It has two hidden layers of width 256 with ReLU activations. Distillation samples are generated by perturbing clean samples within a 2ϵ box. We train R_θ using AdamW with learning rate 10^{-3}, weight decay 10^{-5}, batch size 128, and validation-MSE-based early stopping.

Conditional evaluation. We retain examples satisfying
$$f(\mathbf{x}_i) = y_i,\ f(\mathbf{x}_{i,\mathrm{adv}}) \neq y_i,\ k^\star(\mathbf{x}_{i,\mathrm{adv}}) = y_i.$$
Adaptive examples are generated against $f \circ g_\theta^{\mathrm{dist}}$, while the filtering condition is evaluated before purification. From each retained set, we uniformly sample 2,000 examples; conditional ASR is the fraction that remain misclassified after purification.

D.2 Ablation on EAR-Derived Surrogate Scores

The main experiment uses a joint-density margin as a concise surrogate for EAR-guided purification. We selected this objective by evaluating a family of surrogate scores obtained by simplifying the full Gaussian EAR estimation. This appendix describes these simplifications and reports their empirical behavior on the synthetic dataset.

Let $\widehat{q}_k(\mathbf{x})$ denote the KDE estimate of the class-conditional density $q_k(\mathbf{x})$. We estimate the local class masses as $\widehat{p}_{k,r}(\mathbf{x}) := V_d(r)\alpha_k\widehat{q}_k(\mathbf{x})$. Define the locally dominant class by $\widehat{k}^\star(\mathbf{x}) := \arg\max_k \widehat{p}_{k,r}(\mathbf{x})$. For each rival class $j \neq \widehat{k}^\star$, let $\Delta_j(\mathbf{x}) := \widehat{p}_{\widehat{k}^\star,r}(\mathbf{x}) - \widehat{p}_{j,r}(\mathbf{x})$. The class $\widehat{k}^\star$ is detached and held fixed within each gradient step. Let $c = 0.5$ denote the continuity correction used in the Gaussian-based surrogates, and define

$$a_j^{(c)} := \sqrt{n}\,\Delta_j - \frac{c}{\sqrt{n}}, \qquad \sigma_j^2 := \widehat{p}_{j,r}(\mathbf{x}) + \widehat{p}_{\widehat{k}^\star,r}(\mathbf{x}) - \Delta_j^2,$$

where n is the number of KDE samples in our experiments.

We start from the full multivariate Gaussian approximation. Let $\mathbf{a}^{(c)} = (a_j^{(c)})_{j\neq\widehat{k}^\star}$. The full Gaussian EAR estimate is

$$\widetilde{P}_{\mathrm{adv}}(\mathbf{x}) := 1 - \Phi_{K-1}\left(\mathbf{a}^{(c)}; \mathbf{\Sigma}_V\right),$$

where $\Phi_{K-1}(\cdot; \mathbf{\Sigma}_V)$ denotes the CDF of $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_V)$, and $\mathbf{\Sigma}_V$ is computed from the local masses $\widehat{p}_{k,r}$. We maximize the negative EAR:

$$S_1(\mathbf{x}) := \Phi_{K-1}\left(\mathbf{a}^{(c)}; \mathbf{\Sigma}_V\right) - 1.$$

The implementation estimates this multivariate Gaussian probability using Monte Carlo samples with smooth indicator functions.

The first simplification removes cross-rival covariance while retaining the marginal variance of each rival gap. Define $z_j^{\mathrm{diag}} := \frac{a_j^{(c)}}{\sigma_j}$. The diagonal-covariance surrogate is

$$S_2(\mathbf{x}) := -\left[1 - \prod_{j\neq\widehat{k}^\star} \Phi\left(z_j^{\mathrm{diag}}\right)\right].$$

An alternative simplification removes both cross-rival covariance and marginal variance normalization while retaining the Gaussian-CDF transformation. Define $z_j^{\mathrm{unit}} := a_j^{(c)}$. The unit-covariance surrogate is

$$S_3(\mathbf{x}) := -\left[1 - \prod_{j\neq\widehat{k}^\star} \Phi\left(z_j^{\mathrm{unit}}\right)\right].$$

For the diagonal-covariance approximation, the gradient contribution of rival j is weighted by $w_j := \phi\left(z_j^{\mathrm{diag}}\right)\prod_{\ell\neq j,\widehat{k}^\star}\Phi\left(z_\ell^{\mathrm{diag}}\right)$. We next replace these CDF-induced weights by uniform weights while retaining the marginal variance normalization:

$$S_4(\mathbf{x}) := \frac{1}{K-1}\sum_{j\neq\widehat{k}^\star} z_j^{\mathrm{diag}} = \frac{1}{K-1}\sum_{j\neq\widehat{k}^\star}\frac{a_j^{(c)}}{\sigma_j}.$$

Finally, removing the remaining variance normalization yields our final surrogate, the average local class-mass margin:

$$S_5(\mathbf{x}; r) := \frac{1}{K-1}\sum_{j\neq\widehat{k}^\star}\left[\widehat{p}_{\widehat{k}^\star,r}(\mathbf{x}) - \widehat{p}_{j,r}(\mathbf{x})\right].$$

Under the small-radius approximation, $\widehat{p}_{k,r}(\mathbf{x}) \approx V_d(r)\alpha_k\widehat{q}_k(\mathbf{x})$, so the local class-mass margin is proportional to the joint-density margin:

$$S_5(\mathbf{x}; r) \approx V_d(r)\frac{1}{K-1}\sum_{j\neq\widehat{k}^\star}\left[\alpha_{\widehat{k}^\star}\widehat{q}_{\widehat{k}^\star}(\mathbf{x}) - \alpha_j\widehat{q}_j(\mathbf{x})\right].$$

For a fixed radius r, $V_d(r)$ is independent of $\mathbf{x}$. Because every purification step uses a normalized objective gradient, this shared positive factor does not affect the update direction. Consequently,

Table 4: Conditional ASR and mean pointwise EAR before and after purification for the EAR-derived surrogate ablation using seed 20260817. $\mathrm{EAR_B}$ and $\mathrm{EAR_A}$ denote EAR before and after purification, respectively. Lower is better for ASR and $\mathrm{EAR_A}$; bold indicates the best result in each column.

	Transfer (%)			Transfer (Distilled, %)			Adaptive (Distilled, %)		
Surrogate	ASR	$\mathrm{EAR_B}$	$\mathrm{EAR_A}$	ASR	$\mathrm{EAR_B}$	$\mathrm{EAR_A}$	ASR	$\mathrm{EAR_B}$	$\mathrm{EAR_A}$
S_1: full covariance	1.85	47.09	**31.12**	5.90	47.09	**31.79**	6.65	47.05	31.75
S_2: diagonal covariance	**1.70**	47.09	31.48	6.00	47.09	32.47	5.95	46.30	**31.44**
S_3: unit covariance	2.65	47.09	31.30	**4.55**	47.09	31.86	**5.55**	47.22	31.87
S_4: uniform CDF weights	2.70	47.09	31.37	5.60	47.09	32.14	6.60	47.20	32.28
S_5: joint-density margin	2.70	47.09	31.33	7.35	47.09	32.24	7.80	46.69	31.93

under the small-radius approximation, maximizing S_5 with a normalized-gradient update yields the same update direction as maximizing the corresponding joint-density margin.

All scores are maximized using the same normalized-gradient purification update and the same distillation protocol as the main experiment.

All five surrogates reduce pointwise EAR to similar levels: 31.12%–31.48% under transfer, 31.79%–32.47% under distilled transfer, and 31.44%–32.28% under distilled adaptive evaluation. This stability indicates that the EAR-reducing direction derived from the full Gaussian formulation is largely preserved through the successive approximations from S_1 to S_5. The small non-monotonic differences among the surrogates may partly reflect accumulated numerical errors in covariance estimation and Gaussian-probability approximation.

We therefore use S_5 as a practical surrogate in the main method: it provides comparable EAR reduction, avoids multivariate Gaussian computation, and is independent of r. Although S_5 does not attain the lowest ASR among the EAR-derived surrogates in every setting, its ASR remains within 2.8 percentage points of the best variant. More importantly, Table 1 shows that the resulting EAR-guided method consistently achieves the lowest ASR and post-purification EAR among $p(\mathbf{x})$-, $M(\mathbf{x})$-, and EAR-based guidance.

In the transfer settings, all methods start from the same adversarial examples, yielding identical $\mathrm{EAR_B}$, while their different purification updates produce different $\mathrm{EAR_A}$. Under adaptive evaluation, the adversarial examples are method-specific, so $\mathrm{EAR_B}$ may also vary.

D.3 CIFAR-10 PURIFICATION DETAILS

We evaluate the first 512 CIFAR-10 test images. APGD-CE examples are generated against an undefended WideResNet-70-16-dropout using 100 iterations, $\ell_\infty = 8/255$, and random seed 1, and are then transferred to each purification pipeline with the same WRN checkpoint as the final classifier. The attack does not differentiate through purification. The $p(\mathbf{x})$ baseline uses the unconditional RDC objective, whereas $M(\mathbf{x})$ and EAR-guided purification use a conditional diffusion classifier to estimate a fixed $\widehat{k}^\star(\mathbf{x})$. All methods use five Adam steps with learning rate 0.1 and diffusion times sampled from $[0.4, 0.6]$; the diffusion classifier uses 126 integration steps, and EAR-guided purification uses $T = 32$.