Final Report: Unrestricted Adversarial Training

Marcin Podhajski
marcin.podhajski@akces-ncbr.pl

Tony Wang
twang6@mit.edu

Supervised Program for Alignment Research — Spring 2026

Abstract

The Unrestricted Adversarial Examples Challenge (UAEC) is a longstanding open problem in adversarial robustness that goes beyond ϵ-bounded threat models. It requires a classifier that (i) achieves high accuracy on natural, unambiguous images (e.g., birds vs. bikes), (ii) does not misclassify unambiguous images of one class as another, and (iii) is allowed to abstain when inputs are ambiguous, where "unambiguous" is defined by human agreement. While conceptually appealing, UAEC remains largely unsolved.

A key limitation of ϵ-bounded robustness is that increasing ϵ to cover broader variations eventually forces the model to treat genuinely valid examples as adversarial, effectively collapsing the distinction between correct and incorrect inputs rather than capturing semantic robustness. As a result, standard adversarial training methods do not extend naturally to the unrestricted setting.

In this work, we study UAEC through a simplified high-dimensional toy setting and obtain positive results, providing a first step toward understanding unrestricted robustness.

1 Introduction

In vision, adversarial robustness has traditionally been studied under the ϵ-bounded perturbation framework, where an adversary is restricted to small norm-bounded changes to a natural input. This formulation underlies most existing work on adversarial training, including the approach of Madry et al. [5]. While effective within its domain, this framework relies on a fundamental assumption: that robustness can be characterized by a fixed local neighborhood around each data point.

However, extending ϵ-bounded robustness by simply increasing ϵ does not yield a notion of unrestricted robustness. As ϵ grows, the threat model eventually includes transformations that move between distinct semantic classes, meaning that enforcing invariance to all such perturbations would force the model to misclassify valid examples in order to remain "robust". In this regime, robustness ceases to correspond to preserving label-consistent semantics and instead collapses the distinction between classes. This highlights a fundamental mismatch between bounded perturbation robustness and the notion of semantic robustness we ultimately care about.

The Unrestricted Adversarial Examples Challenge (UAEC) [1], introduced in 2018, addresses this mismatch directly. Formally, the task is to train a classifier that takes as input an image and outputs one of three labels: "bird", "bike", or "abstain". The classifier must satisfy the following conditions:

- Accuracy on natural data: On a held-out set of natural (i.e., non-adversarially designed) unambiguous images of birds and bikes, the classifier must achieve at least 80% accuracy. The training algorithm has access to a training set drawn from the same distribution.

- No targeted misclassification (bird → bike): It should be infeasible to construct an unambiguous image of a bird (as judged by a panel of humans) that the classifier labels as "bike". The model is allowed to abstain.

- No targeted misclassification (bike → bird): Symmetrically, it should be infeasible to construct an unambiguous image of a bike that the classifier labels as "bird". The model is allowed to abstain.

A central challenge in UAEC is that adversaries are not restricted by any geometric notion of distance. This removes the structure that standard adversarial training relies on and makes ϵ-bounded techniques fundamentally misaligned with the problem formulation. Consequently, most progress in adversarial robustness has remained within bounded threat models, and these methods have not translated to the unrestricted setting.

We view UAEC as an operationalization of unrestricted robustness: robustness to any input that preserves human-recognizable semantics, together with the ability to abstain when semantics are unclear. This better reflects real-world deployment requirements, where systems should be reliable on clear cases and defer otherwise. It is also closely related to latent adversarial training, where robustness is defined in learned representation spaces rather than pixel space.

In this work, we study UAEC through a simplified high-dimensional toy problem involving two classes of points drawn from spheres of different radii (one centered and one offset). We show that this setting captures key aspects of unrestricted robustness and can be solved, providing a first step toward more general unrestricted adversarial learning problems.

2 Solving an easier problem first

Instead of tackling the UAEC directly, we have instead been prototyping algorithms in a simpler toy setting inspired by Gilmer et al. [2]'s Adversarial Spheres paper:

We consider two nested "spheres" of radius $r = 1$ and $r = 2$ in R^d ($d = 100$). The spheres are distributions that one can sample from (e.g., first sample an isotropically random direction, then sample a radius from a Gaussian centered at r=1 or r=2). The variance of this Gaussian controls the "thickness" of the sphere. The outer sphere is centered at zero, while the inner sphere is offset. The task is then to robustly classify samples drawn from these two spheres.

We define robust classification to meet the following three conditions (analogous to the conditions of the UAEC):

- On samples sampled directly from the inner or outer sphere distributions, the model classifies these points correctly with at least 95% probability.

- There does not exist any point which lies sufficiently close (as defined by a Minkowski interpolation distance) to the inner sphere that the model classifies as "outer".

- There does not exist any point which lies sufficiently close to the outer sphere (as defined by a Minkowski interpolation distance) to the outer sphere that the model classifies as "inner".

This setup has the following nice properties:

- Firstly, the nested high-dimensional spheres setup is borrowed directly from Gilmer et al. [2], and yields a setup in which natural training does not yield adversarial robustness. (This we have confirmed empirically).

- Secondly, standard bounded adversarial training is also insufficient for meeting the three conditions of the toy problem. (This we have confirmed empirically).
- Due to the toy nature of the problem, we can define what "unambiguous" means mathematically, and do not need to have a human in the loop to do the evaluation. This greatly speeds up our iteration speed.
- The offset of the inner spheres breaks symmetry. This means that certain parts of the inner distribution are closer to the outer distribution than other parts of the inner distribution. This is a property that natural image distributions also have.

3 Evaluation

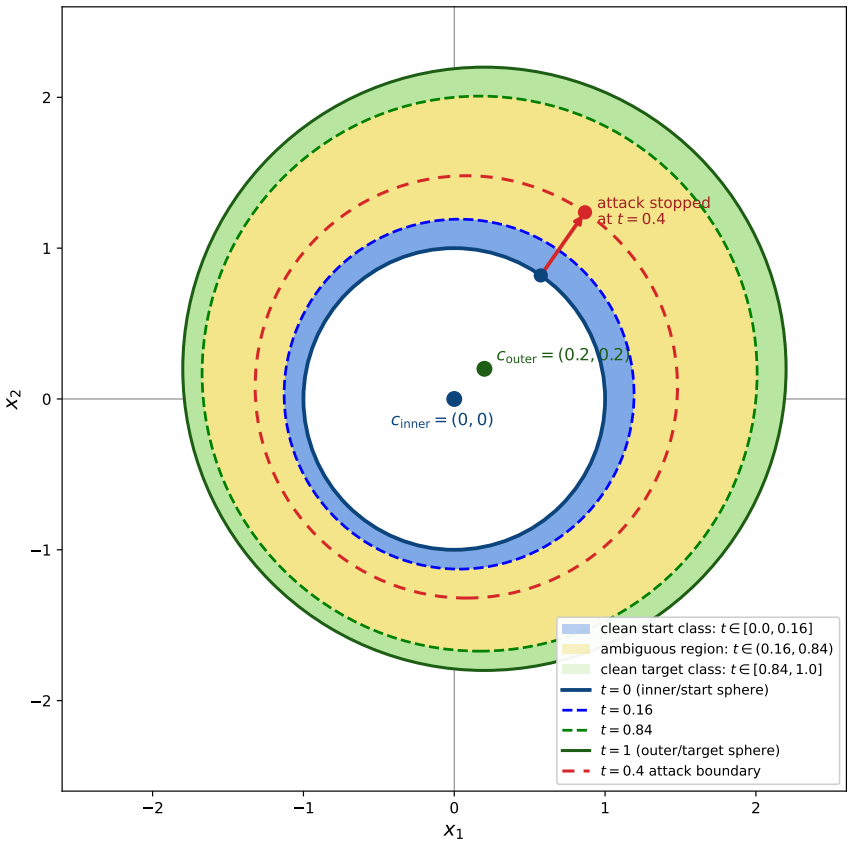

Figure 1: Attack (inner->outer) boundaries using Minkowski interpolation for a 2D dataset.

We evaluate our method using two metrics: *accuracy* and *robustness*.

We calculate the accuracy using clean data, where we treat abstention as a wrong prediction.

We define an attack boundary sphere with radius and center that evolve via a Minkowski interpolation between a starting sphere and a target sphere. The radius and center at interpolation parameter $t \in [0, 1]$ are given by:

$$radius_t = t \cdot radius_{start} + (1 - t) \cdot radius_{target}$$

$$center_t = t \cdot center_{start} + (1 - t) \cdot center_{target}$$

where t controls the interpolation between the two spheres. We then interpret points relative to this interpolation path as follows:

- For $t \in [0, 0.16]$, any point within the corresponding interpolated sphere is considered part of the clean starting data class.
- For $t \in [0.84, 1.0]$, any point within the corresponding interpolated sphere is considered part of the clean target data class.
- For $t \in (0.16, 0.84)$, points are considered ambiguous, as they lie in the transition region between the two classes.

The exact t values were chosen so that $\sim 5\%$ of clean data (outliers) fall in the ambiguous section. A visual example of this definition for a 2D dataset is illustrated in Figure 1. We measure robustness by evaluating robust accuracy (where abstentions are treated as correct predictions) under adversarial attacks bounded by the t-spheres, for multiple values of $t \in [0, 1]$.

4 Baselines

We evaluate multiple baselines: a standard (natural) classifier, a bounded adversarially trained classifier [5], a diffusion classifier [4], and an Equilibrium Matching (EqM) energy model [6]. For all methods, we additionally introduce an abstention mechanism via a calibrated confidence threshold: for each model, we select the threshold such that predictions are made on the top 95% of clean samples ranked by confidence, while the remaining 5% are abstained.

"Perfect" model We also define a "perfect" model, which predicts the starting class for any point within the corresponding interpolated sphere for $t \in [0, 0.16]$, abstain for $t \in (0.16, 0.84)$, and the target class for $t \in [0.84, 1.0]$. This model abstains on $\sim 5\%$ of clean data (as some of the outliers naturally fall outside of "$t = 0.16$"-sphere).

We plot the robustness curves for the baselines in Figure 2. We observe that none of the methods achieves a robust solution for the clean data sphere ($t \leq 0.16$).

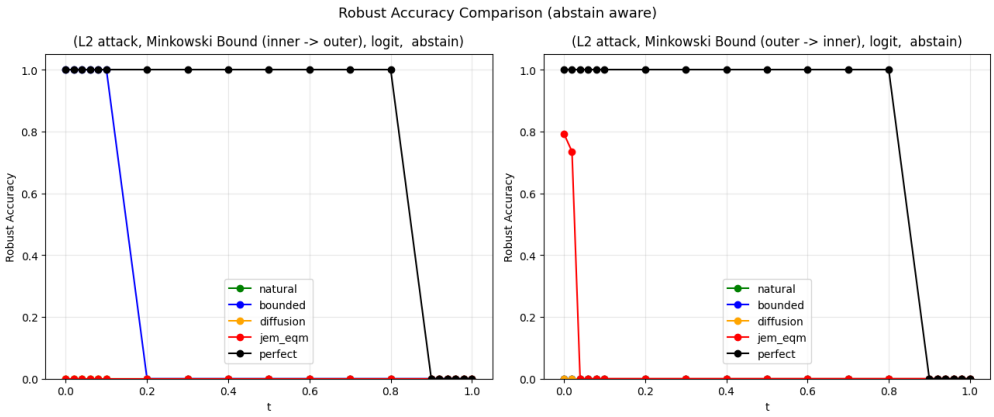

Figure 2: Robustness evaluation for the baselines using a D=100 dataset.

5 Our method

In our algorithm, which we call Unrestricted Adversarial Training (UAT), we train the network to minimize:

$$L(x, y) = CE(x, y) + L_{adv}(x_{adv}, y).$$

where CE denotes standard Cross Entropy. x_{adv} is constructed by an adversarial PGD-like attack maximizing the difference between the target class $(1 - y)$ logit and the maximum of the other two logits. The attack bound is specified in the following sections. L_{adv} is defined by:

$$L_{adv}(x, y) = -\log\left(1 - \text{softmax}(x)[1 - y]\right)$$

5.1 Using ground truth as the attack bound

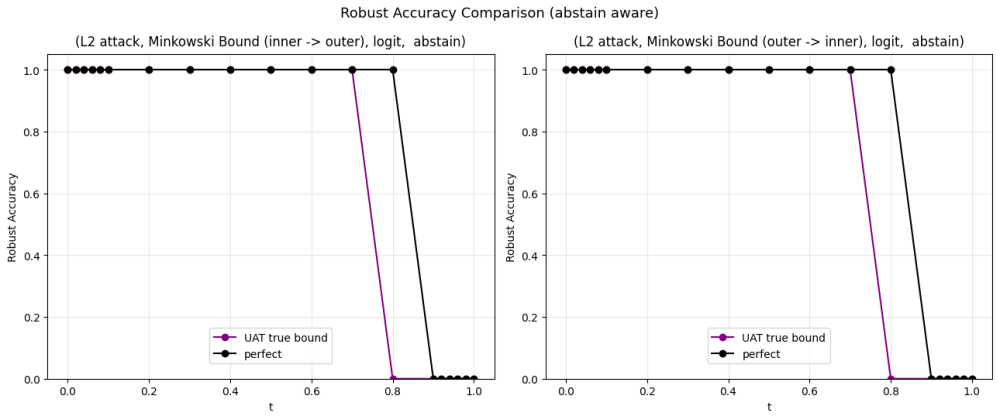

Figure 3: Robustness evaluation for the UAT with true boundary method using a D=100 dataset.

The first experiment we ran was a sanity check to verify our algorithm. Specifically, we evaluate on a sphere with the predefined stopping boundary $t = 0.84$ for the attack. We evaluate the robustness of this method in Figure 3.

Figure 4 shows the training dynamics of a representative successful run. We identify four distinct phases: (1) a warm-up period in which the model outputs only hard predictions; (2) a sharp transition steps where the model begins to abstain heavily, reaching up to 50–100% abstention on clean samples depending on dimensionality; (3) an exploration phase where the model learns to push adversarial inputs toward the stopping boundary; and (4) a recovery phase where clean accuracy is restored as abstention on clean samples gradually decreases, while adversarial inputs remain at the boundary.

5.2 Using an auxiliary classifier to predict the attack bound

We consider a method that uses an auxiliary classifier to determine when an attack should be terminated. We first adapt the auxiliary classifier to include an abstention mechanism by thresholding the predicted probability of the target class. The threshold is chosen such that the classifier predicts the target label on the top 95% of clean samples (ranked by target-class probability) and abstains on the remaining 5%.

During adversarial attack generation against the UAT model, we then use this auxiliary classifier as a stopping criterion: the attack is terminated once the auxiliary model predicts the target class, while it continues to run if the auxiliary model abstains.

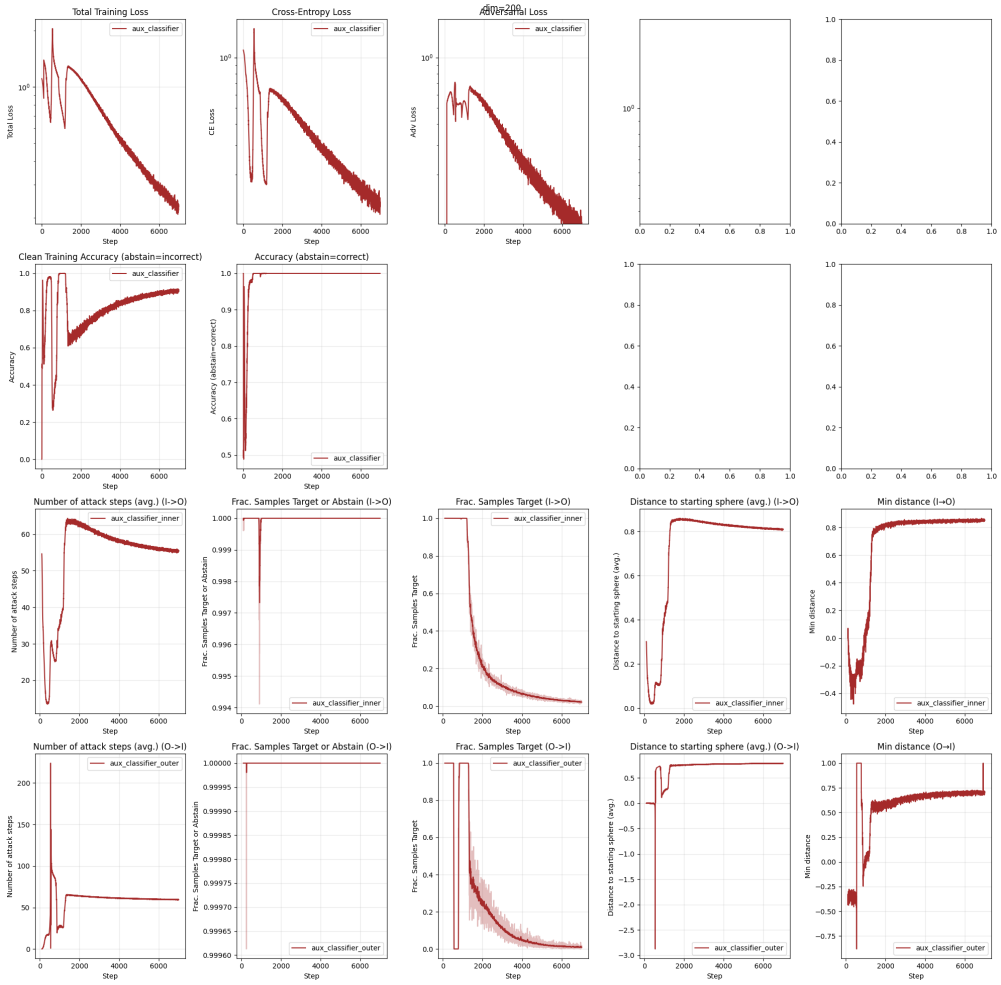

Figure 4: Training curves using a D=200 dataset.

We achieved the best results using a bounded adversarial training model as the auxiliary classifier. We evaluate the robustness of this method in Figure 5.

Unlike the oracle stopping boundary t, the auxiliary classifier induces a *soft* boundary: different attack trajectories are terminated at varying distances from the original sample, ranging approximately from $t = 0.6$ to $t = 0.9$ in our experiments. As a result, the robustness achieved by this method is lower than that of the oracle, since some attacks are terminated too early.

Nevertheless, this level of robustness is sufficient to satisfy our problem requirements, which technically only demand robustness up to $t = 0.16$. Furthermore, being robust up to $t = 0.5$ carries a natural geometric interpretation: any misclassified example is closer to the target class than to the original class in the input space, which we consider a strong and meaningful robustness guarantee in its own right.

5.3 Using the model itself to predict the attack bound

We next consider a fully adaptive variant of the attack-stopping mechanism, in which the model itself is used to estimate when an attack should terminate. Instead of relying on a fixed auxiliary

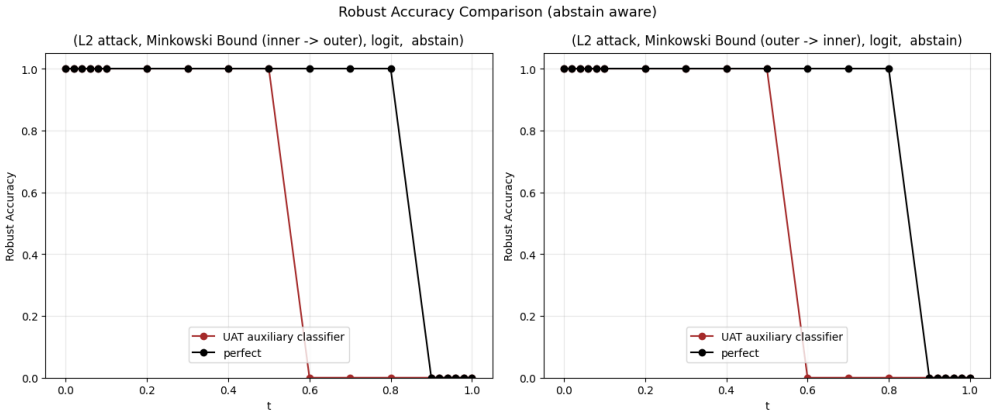

Figure 5: Robustness evaluation for the UAT with auxiliary classifier method using a D=100 dataset.

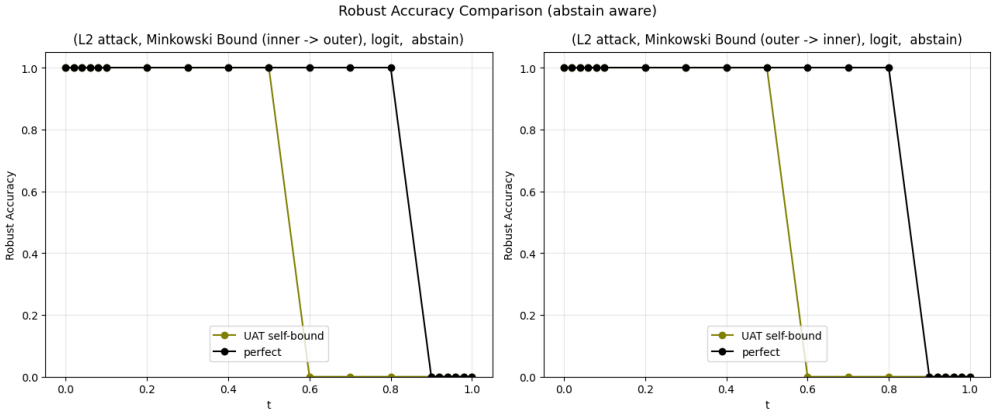

Figure 6: Robustness evaluation for the UAT with auxiliary classifier method using a D=100 dataset.

classifier, we use the current state of the UAT model to dynamically determine an implicit attack bound at each training step.

Concretely, at each iteration, we compute a probability threshold from the model's predictions on clean training data. The threshold is chosen so that the model predicts the target class on the top 95% of clean samples (ranked by target-class probability), while abstaining on the remaining 5%. This induces a moving decision boundary that adapts as the model evolves during training. We then use this threshold as a stopping criterion for adversarial attacks: during attack generation, the optimization is terminated as soon as the model assigns the target class with probability above the current threshold. If the model instead outputs abstain (i.e., falls below the threshold), the attack is allowed to continue.

We evaluate the robustness of this method in Figure 6. While the induced boundary does shift during training, we observe that it consistently converges prematurely at approximately $t = 0.5$ rather than the optimal range of $t = 0.7$–0.8; understanding and correcting this behavior remains an open direction we are currently working on.

5.4 Scaling to higher dimensions

We observe that scaling the input dimensionality requires a corresponding increase in network width and training duration. For example, at $d = 1000$, a width of 10^6 in the first layer is necessary to

achieve strong performance. We verify this by distilling the perfect classifier via bounded adversarial training with a KL divergence loss, and evaluating the resulting robustness directly.

6 Future Work

Our immediate goal is to further stabilize training and evaluation on the current toy problem. While we already obtained a solution that satisfies the problem constraints, we aim to better understand the dynamics of the training algorithm and develop methods that are more robust in the ambiguous region.

Beyond the toy setting, we plan to extend the approach to real-world datasets. A natural next step is CIFAR-10 [3], which provides a standard benchmark for assessing whether the mechanisms developed in the simplified setting transfer to more complex, natural image distributions. This will help evaluate whether the ideas underlying our approach can scale to realistic unrestricted robustness settings.

References

[1] T. B. Brown, N. Carlini, C. Zhang, C. Olsson, P. Christiano, and I. Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

[2] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018. URL https://arxiv.org/abs/1801.02774.

[3] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[4] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, October 2023.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[6] Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.