GEOMETRIC IMPLICATIONS OF CLASSIFICATION ON REDUCING OPEN SPACE RISK

Matthew Lau, Leyan Pan, Stefan Davidov, Athanasios P Meliopoulos & Wenke Lee
College of Computing and College of Engineering, Georgia Institute of Technology

ABSTRACT

To reduce open space risk of hypotheses, we reexamine the 'simplest' hypothesis class, binary linear classifiers, geometrically. Generalizing linear classification, we establish a surprising fact: linear classifiers can have arbitrarily high VC dimension, stemming from increasing the number of partitions in input space. Hence, linear classifiers with multiple margins are more expressive than single-margin classifiers. Despite a higher VC dimension, such classifiers have less open space risk than halfspace separators. These geometric insights are useful to detect unseen classes, while probabilistic modeling of risk minimization helps with seen classes. In supervised anomaly detection, we show that a classifier that combines a probabilistic and geometric lens can detect both seen and unseen anomalies well.

1 INTRODUCTION

For a machine learning model to be safe, it must be able to reject anomalous data. Conventional halfspace separation, modeled in sigmoid and rectified linear units (ReLU), is limited in open-world settings. Scheirer et al. (2013) uses a margin instead to reduce open space risk (OSR). Lau et al. (2023) provides a linear formulation of this margin from a geometric perspective and closing numbers metric for OSR in neural networks (NNs). We generalize their linear formulation to the full extent, which includes allowing for multiple margins. We gain two geometric insights. First, having more margins is more expressive than one margin. In fact, we challenge the simplicity of linear classifiers, showing an arbitrarily high VC dimension (stemming from the number of margins/partitions in input space). Second, such classifiers with margins have a smaller closing number (lower OSR) than halfspace separators. These two insights guide our experiments on supervised anomaly detection (AD), where we use a small number of margins. In supervised AD, we need to optimize over two objectives. The first is detecting seen anomalies, which can be done through a standard probabilistic approach of empirical risk minimization (ERM). The second is detecting unseen anomalies, usually done by reducing OSR. However, probabilistically modeling unknown data requires assumptions. For instance, Sipple (2020) generates anomalies uniformly in raw data space while Tao et al. (2023) samples in feature space but requires a multi-class classification task. Instead of a probabilistic approach, we construct the hypothesis class of NNs to implicitly control the OSR through geometry. Our combined probabilistic and geometric approach detects both seen and unseen anomalies well.

2 GEOMETRIC THINKING TO COMPLEMENT PROBABILISTIC THINKING

Classification On a joint distribution D_{XY} of data and binary labels, we aim to minimize the risk $\mathbb{P}_{\mathbf{x},y\sim D_{XY}}(h(\mathbf{x})\neq y)$ with respect to hypothesis h in a given hypothesis class $\mathcal{H}$. The simplest hypothesis class is commonly thought to be linear classifiers (Shalev-Shwartz & Ben-David, 2014)

$$\mathcal{H}:=\phi\circ A_n=\{\mathbf{x}\mapsto\phi(h_{\mathbf{w},b}(\mathbf{x})):h_{\mathbf{w},b}\in A_n\},\ A_n=\{\mathbf{x}\mapsto\mathbf{w}^T\mathbf{x}+b:\mathbf{w}\in\mathbb{R}^n,b\in\mathbb{R}\}\quad(1)$$

where A_n is the set of affine functions in $\mathbb{R}^n$ and $\phi:\mathbb{R}\to\{0,1\}$ is the labeling function. To formalize the labeling function, we denote it as the indicator function $\phi:=\mathbb{I}_S$, where set $S\subseteq\mathbb{R}$ denotes the region where one class lives in while the other class lives in the complement $\mathbb{R}\backslash S$. The linearity of the classifier relies on one weight vector (and one bias term) to determine if a datum belongs to S or $\mathbb{R}\backslash S$. Halfspace separation is common, where $S=\mathbb{R}_+$. This approach conveniently models the region where it is more likely for a datum to be the positive class. Halfspaces are

Table 1: AUPR for NSL-KDD. DoS attacks (anomalies) are seen while other types of attacks are unseen. Overall AUPR against all attacks is also reported. We compare halfspace separators (HS) and our geometric separator (GS) with baselines. The suffix (-s) denotes the number of margins used.

Model\Attack	DoS	Probe	Privilege	Access	Overall
Random	0.435	0.200	0.007	0.220	0.569
SVM	0.959±0.000	0.787±0.000	0.037±0.000	0.524±0.000	0.948±0.000
OCSVM	0.779±0.000	0.827±0.000	0.405±0.000	0.760±0.000	0.897±0.000
HS MLP	**0.960±0.022**	0.842±0.023	0.024±0.009	0.240±0.039	0.909±0.019
GS-1 (ours)	0.918±0.010	0.798±0.034	0.399±0.006	0.709±0.010	0.929±0.010
GS-2 (ours)	0.936±0.010	**0.853±0.026**	**0.407±0.013**	**0.778±0.014**	**0.953±0.005**

symmetrical, so the 0/1 labels are arbitrary. Since this symmetry does not hold for general S (e.g. $S = \{0\}$ (Lau et al., 2023)), we propose generalizing linear classification to allow label flipping, which we denote as S-membership separation (e.g. Figure 1):

$$\mathcal{H}_S = \{\mathbb{I}_S(h_{\mathbf{w},b}(\cdot)) : h_{\mathbf{w},b} \in A_n\} \cup \{\mathbb{I}_{\mathbb{R}\setminus S}(h_{\mathbf{w},b}(\cdot)) : h_{\mathbf{w},b} \in A_n\}. \tag{2}$$

Expressivity The VC dimension (Vapnik & Chervonenkis, 1971) of classical halfspace separators is $n+1$ in $\mathbb{R}^n$. We claim that S-membership separators can have arbitrarily high VC dimension. We prove this in the specific case where S is a countable set (i.e. membership to multiple hyperplanes):

Theorem 2.1. *Let $x \in \mathbb{R}^n$ for $n \in \mathbb{N}$, $S \subseteq \mathbb{R}$ be a countable set with size $s := |S| \in \{1, 2, ..., \infty\}$. Denote this $\mathcal{H}_S$ as hyperplane separators. Then, $2n + s \leq VCdim(\mathcal{H}_S) \leq 2sn + 2s - 1$.*

To generalize these zero-measure hyperplanes, we allow a margin of error between each hyperplane:

Corollary 2.2. *Let $x \in \mathbb{R}^n$ for $n \in \mathbb{N}$, $S \subseteq \mathbb{R}$ and s be the minimum number of finite-length closed intervals needed to partition S. Denote this $\mathcal{H}_S$ as margin separators. Then, $2n + 2 + s \leq VCdim(\mathcal{H}_S) \leq 2sn + 4s + 1$.*

Infinite sized S can be induced by familiar operators. An example of $s = \infty$ hyperplane separators is identifying if adding n numbers is zero modulo k for $k \in \mathbb{N}$. Here, $S = k\mathbb{Z}$. An example of $s = \infty$ margin separators is sinusodial classifiers with infinite VC dimension (Shalev-Shwartz & Ben-David, 2014). We show that the expressivity of the sine function emerges from its ability to partition the input space into infinite parts. This classifier is, in fact, linear – the sine function merely acts as a labeling function. Conversely, margin separators with finite s have a closing number of n. This is unlike halfspace separators which have a higher closing number of $n + 1$ with stricter constraints required to attain this (Lau et al., 2023). In other words, by increasing the number of margins s, we can increase expressivity while maintaining a lower OSR than halfspace separators.

Experiments We calculated area under the precision-recall curve (AUPR) for a supervised AD task on NSL-KDD cyber-attack dataset (Tavallaee et al., 2009) to measure the separation between normal and anomalous (attack) data. To geometrically control OSR, we aim to enclose normal data within a region (Figure 2). To form closed decision regions, we use a multi-layer perceptron (MLP) with Gaussian bump activations (Eq. 3) in the penultimate layer for $s = 1, 2$ and an RBF activation centered at $\mathbf{1}$ in the output layer. To ensure that the decision region encloses normal data, we disallow label flipping, restricting normal data within the margin(s). Coupling geometric modeling with a standard probabilistic ERM approach, we do not explicitly minimize OSR in the loss function. Table 1 presents key comparisons of our method, denoted as geometric separator (GS), with supervised (SVM and halfspace MLP) and unsupervised baselines (OCSVM). More details can be found in Appendix D.2 and E. GS's detection of unseen and seen anomalies is competitive with unsupervised and supervised methods respectively.

3 Conclusion

In this paper, we viewed linear classification from a geometric lens to generalize it to S-membership separators. We showed an arbitrary high VC dimension coming from the number of margins s the hypothesis class induces. Margin separators are one such type of linear classifiers that control open space risk better than halfspace separators, which is useful for rejecting unseen data. Combining this geometric lens with standard probabilistic modeling (via risk minimization) for supervised anomaly detection, our approach detects both seen and unseen anomalies well.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. *LOF: Identifying Density-Based Local Outliers*, pp. 93–104. Association for Computing Machinery, New York, NY, USA, 2000. ISBN 1581132174. URL https://doi.org/10.1145/342009.335388.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995. ISSN 1573-0565. doi: 10.1007/BF00994018. URL https://doi.org/10.1007/BF00994018.

Gary William Flake. *Nonmonotonic Activation Functions in Multilayer Perceptrons*. PhD thesis, The University of Maryland, 1993.

Matthew Lau, Ismaila Seck, Athanasios P Meliopoulos, Wenke Lee, and Eugene Ndiaye. Revisiting non-separable binary classification and its applications in anomaly detection, 2023.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgH0TEYwH.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=aRtjVZvbpK.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, Cambridge, England, July 2014.

John Sipple. Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9016–9025. PMLR, 2020. URL http://proceedings.mlr.press/v119/sipple20a.html.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JHklpEZqduQ.

Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009. doi: 10.1109/CISDA.2009.5356528.

David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, January 2004. doi: 10.1023/b:mach.0000008084.60811.49. URL `https://doi.org/10.1023/b:mach.0000008084.60811.49`.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL `https://doi.org/10.1137/1116025`.

A Background on open space risk

We refer to Scheirer et al. (2013) for their introduction on open space risk for open set recognition, summarizing key ideas here. We note that, in the one-class case, open set recognition reduces to supervised anomaly detection. In general, the security-related fields of machine learning (novelty, anomaly, out-of distribution detection and open set recognition) have many common ideas (Salehi et al., 2022). Hence, our ideas in supervised anomaly detection are directly applicable to these other fields.

Scheirer et al. (2013) recognizes that in classification, the closed-world assumption may be violated. In other words, inputs to a classifier (e.g. machine learning model) may be unexpected and follow a different distribution that is not captured in the training data, such as inputs that do not belong to any of the classes. In these cases, a classifier that outputs the closest class for a given input is insufficient – the classifier must also indicate if the input belongs to that class. To avoid such cases, a classifier should not label unknown regions of the input space, which Scheirer et al. (2013) refers to as *open space*. We would like to minimize the volume of open space that is labeled by the classifier as a particular class (denoted as $\mathcal{O}$) across all classes, which is referred to as open space risk minimization. Thus, Scheirer et al. (2013) defines the open space risk for classifier $f : \mathcal{X} \to \{0, 1\}$ that (classifies the input as 1 for the class of interest and 0 otherwise) as

$$R_{\mathcal{O}}(f) := \frac{Vol(\text{Positively labeled open space})}{Vol(\text{Large ball } S_O)} = \frac{\int_{\mathcal{O}} f(x)dx}{\int_{S_O} f(x)dx}$$

where $S_O \subseteq \mathcal{X}$ covers all possible inputs (training examples from the class of interest and the positively labeled open space i.e. inputs that are misclassified as the class of interest). Note that S_O is usually a (bounded) ball rather than the whole of $\mathcal{X}$ to prevent both the numerator and denominator from being infinite. For instance, if $S_O = \mathcal{X} = \mathbb{R}^n$, the denominator is infinity and the numerator is potentially infinity too (because $\mathcal{O}$ may be unbounded).

However, it is not clear how to properly define $\mathcal{O}$ (for a particular class) and S_O because it may not be straightforward to define what open space we are concerned with – if we have apriori knowledge of the unknowns, such a problem would not appear to violate the closed-world assumption in the first place (and we can just perform regular classification). Hence, our work considers that S_O is potentially unbounded (e.g. $S_0 = \mathcal{X} = \mathbb{R}^n$) and ensures that $\mathcal{O}$ is bounded (through mandating that the whole decision region, i.e. the positively labeled region, is bounded). In our paper, we show that reducing the open space risk by reducing the volume of $\mathcal{O}$ from infinite to finite is empirically effective. This geometric approach of controlling the open space risk differs from other works that probabilistically control the open space risk such as by sampling outliers to perform empirical risk minimization (e.g. Sipple (2020); Tao et al. (2023)).

B Proof of VC Dimension of $\mathcal{H}_S$

We provide proofs for the VC dimension claims in Section 2.

B.1 Countable S

We prove Theorem 2.1 by induction, referring to $\mathcal{H}_{S_s}$ as S-membership separators with $|S| = s$. The base case of $s = 1$ where VC dimension is $2n + 1$ is proved in Lau et al. (2023). For induction from s to $s + 1$, we provide upper and lower bounds:

$$1 \leq VCdim(\mathcal{H}_{S_{s+1}}) - VCdim(\mathcal{H}_{S_s}) \leq 2n + 2.$$

The upper bound holds because of the union property (Shalev-Shwartz & Ben-David, 2014): all hyperplanes of the S-membership separator are parallel, so having another hyperplane will increase the VC dimension by at most $(2n + 1) + 1 = 2n + 1$. Hence, the upper bound is $2n + 1 + (s - 1)(2n + 2) = 2sn + 2s - 1$.

To prove the lower bound, we show that the VC dimension of $\mathcal{H}_S$ is monotonically increasing with the number of hyperplanes s. We consider a set of points that can be shattered with $s - 1$ hyperplanes. Adding an additional point to this set that is not coplanar with any other subset of n points from the original set, we shatter this new set with an additional hyperplane by either letting the added hyperplane pass through the point or not (depending on the added point's label). We can shatter the new set because, by design, this added hyperplane does not have to pass through any points from the original set.

B.2 S IS A COLLECTION OF INTERVALS

To prove Corollary 2.2, we use the same technique to show that for s intervals,

$$1 \leq VCdim(\mathcal{H}_{S_{s+1}}) - VCdim(\mathcal{H}_{S_s}) \leq 2n + 4,$$

with the base case of $s = 1$ of VC dimension $2n + 3$ proved in Lau et al. (2023). By increasing the VC dimension by at most $(2n + 3) + 1 = 2n + 4$ from s to $s + 1$, the upper bound is $2n + 3 + (s - 1)(2n + 4) = 2sn + 4s + 1$. The lower bound is from Theorem 2.1, since we can make the margin width arbitrarily small.

A sample visualization of margins separators is shown in Figure 1, where points in the shaded region are classified as one class (e.g. positive) while other points are classified as the other class (e.g. negative). Figure 1 visually demonstrates the linear property – the weight vector is constant, inducing parallel lines or regions for classification.

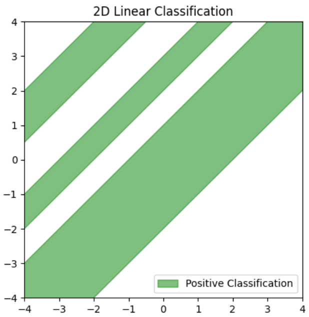

Figure 1: Example of margin separation in 2 dimensions, where $\mathbf{w} = [1, -1]^T$, $b = 0$ and $S = [-6, -4.5] \cup [-3, -2] \cup [-1, 2]$.

C CLOSING NUMBER OF S-MEMBERSHIP SEPARATORS

As proposed in Lau et al. (2023), the closing number of a hypothesis class is the minumum number of hypotheses from the class such that the volume (Lebesgue measure) of their intersection is finite and non-zero. This is especially useful in describing how neurons work together in a neural network to form closed decision regions. Forming closed decisions is beneficial for anomaly detection and controlling the open space risk.

In Section 2, we claimed that margin separators $\mathcal{H}_S$ with finite number of intervals s have a closing number of n. Similar to the proof in Lau et al. (2023), one can prove that a positive-(non-zero) and finite-volumed decision region can be induced by the intersection of n hypotheses from $\mathcal{H}_S$. As with Lau et al. (2023), the only constraint is that the weight vectors from these n hypotheses are a linearly independent set of vectors.

D GEOMETRIC INTERPRETATION

D.1 RELATED WORKS

Geometric modeling is probably the default approach for unsupervised anomaly detection to detect unseen anomalies. Some works aim to embed the normal data into a hypersphere, such as support vector data description (Tax & Duin, 2004) / one-class support vector machines (OCSVM) (Schölkopf et al., 1999) with radial basis function (RBF) kernel and deep semi-supervised anomaly detection (Ruff et al., 2020). Others measure distance, such as LOF (Breunig et al., 2000).

Probabilistic modeling is convenient because of the ERM paradigm. A preference for probabilistic modeling is even suggested in the terminology "open space *risk*" to quantify incorrectly labeled unknown data (Scheirer et al., 2013). Some works that already have a probabilistic formulation of seen data (via ERM) opt to cast open space risk minimization into an ERM problem by sampling data that are assumed to be (mostly) outliers, such as in out-of-distribution (OOD) detection (Tao et al., 2023). Note that these probabilistic formulations of open space risk complement and can be used with our geometric approach.

D.2 OUR PROPOSED METHOD

Our work aims to combine both probabilistic approach for seen anomalies and geometric approach for unseen anomalies. We refer to Figure 2 to visualize the geometry in our proposed neural network architecture. First, we feed the raw data into a feature extractor $f : \mathcal{X} \to \mathbb{R}^d$. In simple cases such

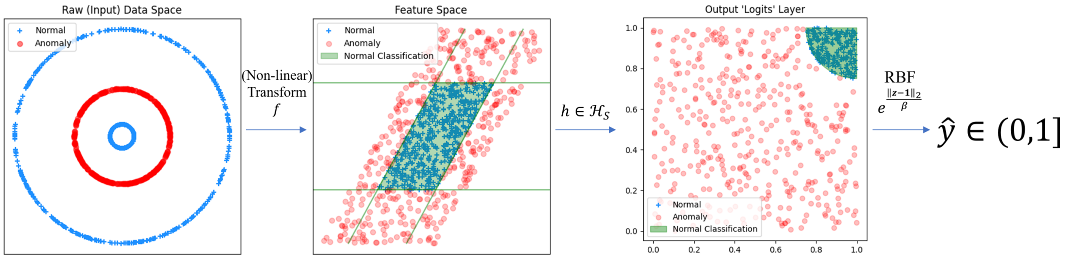

Figure 2: Geometric approach to anomaly detection. Data shown is for visualization purposes.

as when the raw data are sampled from a Gaussian, f may be just the identity. The goal of f is to embed the raw data into a feature space such that we can enclose the normal data. Ideally, to ensure that the eventual network is activated only on a finite volume in input space (to induce closed decision regions in input space), we want to design f such that any set of finite volume in feature space has a finite-volumed pre-image under f as well.

In the following hidden layer, we follow principles from closing numbers by taking the intersection of d margin separators to induce the closed decision region. This is modeled by choosing d neurons in this hidden layer with bump activations, followed by an RBF unit centered at $\mathbf{1}$. We use the Gaussian bump function

$$B(z; \mu, \sigma) = \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right] \tag{3}$$

for fixed μ, σ. We also tried another bump function $\tanh(\sigma^2/z^2)$ (Flake, 1993) and had similar results in our experiments. Note that RBF activations can also be used here, but they can only model norm balls, while margin separators are flexible and do not depend on the features having identity covariance matrix. To have the effect of having multiple closed decision regions (similar to having multiple RBF units), the margin separator can have more than 1 margin (i.e. $s > 1$), such as adding multiple bump functions with a learnable μ for each (and then normalizing it so the maximum of the function is 1).

As another geometric design, we disallow label flipping, assigning normal data to the positive class within the margin(s). Lau et al. (2023) observes that such a structure can improve the estimation error bound by reducing the VC dimension of the classifier while reducing approximation error with domain knowledge that normal data has more structure than anomalies. Note that both seen and unseen anomalies are assigned to the negative class.

E Experimental details

To simulate supervised anomaly detection, we follow the set-up from Lau et al. (2023) and share code in `https://github.com/mattlaued/Geometric-Implications-Classification-OSR`. We treat denial of service (DoS) attacks as seen anomalies, including them with normal data during training (with labels). Other attacks (probe, privilege escalation and remote access) are removed from the training data to be treated as zero-day attacks (unseen anomalies) during inference. The aim is to detect if a particular datum is normal or anomalous, with the expectation that unseen anomalous data will be present during inference.

We also note that DoS attacks have some similarities with probe attacks, so some generalization in detecting probe attacks is expected. We report detailed results in Table 3, with shallow models (i.e. non-neural network approaches) taken from Lau et al. (2023). These shallow models are (1) the random baseline (calculated in expectation), (2) supervised approach: support vector machine (SVM) (Cortes & Vapnik, 1995) and (3) unsupervised approaches: OCSVM (Schölkopf et al., 1999), isolation forest (IsoF) (Liu et al., 2008) and local outlier factor (LOF) (Breunig et al., 2000).

Table 2: Detailed AUPR results for NSL-KDD. DoS attacks (anomalies) are seen while other types of attacks are unseen. Overall AUPR against all attacks is also reported. We compare shallow baselines (provided in Lau et al. (2023)) with neural networks: halfspace separators (HS), equality separators (Lau et al., 2023) and our geometric separator (GS). The suffix (-s) denotes the number of margins used.

	Model\Attack	DoS	Probe	Privilege	Access	Overall
	Random	0.435	0.200	0.007	0.220	0.569
Shallow	SVM	0.959±0.000	0.787±0.000	0.037±0.000	0.524±0.000	0.948±0.000
	OCSVM-N	0.835±0.000	0.849±0.000	0.382±0.000	0.745±0.000	0.920±0.000
	OCSVM-A	0.779±0.000	0.827±0.000	**0.405±0.000**	**0.760±0.000**	0.897±0.000
	IsoF-N	**0.964±0.006**	**0.960±0.003**	0.039±0.007	0.438±0.015	**0.957±0.002**
	IsoF-A	0.765±0.073	0.850±0.066	0.089±0.044	0.392±0.029	0.865±0.031
	LOF-N	0.759±0.000	0.501±0.000	0.046±0.000	0.451±0.000	0.824±0.000
	LOF-A	0.495±0.000	0.567±0.000	0.039±0.000	0.455±0.000	0.718±0.000
MLPs	HS	**0.960±0.022**	0.842±0.023	0.024±0.009	0.240±0.039	0.909±0.019
	ES	0.953±0.011	0.772±0.029	0.114±0.029	0.625±0.098	0.943±0.015
	GS-1 (ours)	0.918±0.010	0.798±0.034	0.399±0.006	0.709±0.010	0.929±0.010
	GS-2 (ours)	0.936±0.010	**0.853±0.026**	**0.407±0.013**	**0.778±0.014**	**0.953±0.005**

During training, supervised approaches (including our geometric separator) perform binary classification with standard empirical risk minimization (ERM)[1]. We train halfspace separators and equality separators (Lau et al., 2023) as neural network baselines, while SVM is the shallow baseline. We also perform ablations of our geometric separator for $s = 1, 2, 3, 4, 5$. Our geometric separator is a 3-layer multi-layer perceptron (MLP), where activation functions in f (i.e. first hidden layer) are leaky rectified linear units (ReLU) and activation functions in the penultimate layer are bump activations to model the margin separator. To model more than 1 margin, we space the peak of each bump equidistant from each other (i.e. the μ parameters are fixed distance away from each other across consecutive bump functions). To more accurately model arbitrary placement of the margin, one can implement a linear layer with shared weights but separate bias terms or learnable μ parameters. We opt for equidistant bumps for simplicity and to prevent separate bumps from getting too close together and possibly collapsing into themselves.

The common hyperparameters for all neural networks are stated below:

1. Models are trained with logistic loss with the Adam optimizer under an exponentially decaying learning rate.
2. Models are trained for 500 epochs under an early stopping patience of 10 epochs with validation loss. Validation split is 0.1.
3. All layers have the same width (122 neurons[2]) less the output layer with 1 neuron.
4. Halfspace separators have leaky ReLU in hidden layers and sigmoid in the output layer, while equality separators have bump activations in all layers. Our geometric separator activations are detailed above.
5. Bump activations are initialized with variance parameter $\sigma = 0.5$, while RBF scale parameter β (as in Figure 2) is set to 1. Bump activations with multiple margins are of a distance of σ away from each other.
6. Weight initializer is a seeded Glorot initializer for reproducibility.
7. All other hyperparameters are set to default Tensorflow parameters.

[1] Note that SVMs have a more geometric interpretation of finding the maximum margin, but our main comparisons are with neural networks, so we ignore this exception.

[2] In our geometric separator, the benefit of designing f to be a leaky ReLU under a linear transformation is that, if the linear transform is not a projection, then f maintains the good property that a finite-volumed set in feature space has a finite-volumed pre-image under f.

Table 3: AUPR results for NSL-KDD on our geometric separators (GS). The suffix (-s) denotes the number of margins used, and we perform ablations for $s = 1, 2, 3, 4, 5$.

Model\Attack	DoS	Probe	Privilege	Access	Overall
GS-1 (ours)	0.918±0.010	0.798±0.034	0.399±0.006	0.709±0.010	0.929±0.010
GS-2 (ours)	**0.936±0.010**	**0.853±0.026**	0.407±0.013	**0.778±0.014**	**0.953±0.005**
GS-3 (ours)	0.817±0.024	0.709±0.021	0.500±0.022	0.692±0.049	0.880±0.019
GS-4 (ours)	0.785±0.043	0.668±0.038	**0.507±0.002**	0.666±0.061	0.851±0.037
GS-5 (ours)	0.724±0.010	0.607±0.012	0.504±0.001	0.618±0.012	0.793±0.013

In contrast, unsupervised approaches cannot capitalize on labeled anomalies. To test them, we consider 2 approaches and report them both with a suffix of “-N” and “-A”. The former only uses normal data, while the latter uses all data. The unsupervised baselines are OCSVM, IsoF and LOF.

From our results, we notice that shallow baselines typically trade-off detecting seen anomalies and detecting unseen anomalies – a high AUPR in one column is balanced with a low AUPR in another. For neural networks, we corroborate with Lau et al. (2023) that equality separators perform better than halfspace separation in detecting the unseen privilege escalation and remote access attacks, while halfspace separation performs well on seen attacks. Nevertheless, our geometric separator for $s = 1, 2$ clearly outperforms the equality separation, achieving supervised-level performances on seen anomalies (i.e. DoS attacks and, to some extent, probe attacks) and unsupervised-level performance on unseen anomalies (i.e. privilege escalation and remote access attacks). Here, we see the benefits of combining an ERM approach with geometric intuitions to control open space risk – we minimize empirical risk through standard optimization, while we control open space risk by carefully choosing our hypothesis class. In this case, the hypothesis class is a neural network with edited activation functions in the penultimate and last layer. In addition, choosing the width of a hidden layer has a geometric interpretation with closing numbers (Lau et al., 2023) which we use to our advantage to control the open space risk.

We note that training $s \geq 3$ for our geometric separator incurs a training loss of more than a magnitude higher than $s \leq 2$ and does not converge. With a higher empirical risk and higher VC dimension for $s \geq 3$, we chose $s \leq 2$ for our experiments for a better generalization bound. By observing training loss, we can perform structural risk minimization and reject models with high losses without needing to observe test set performance. High loss without convergence also suggests that the model is stuck in a bad local minimum. An interesting way forward would be to understand how to initialize such networks better and get out of local minima. Nevertheless, our finite s geometric separators still err on the more conservative side, preferring to reject unseen data more than halfspace separators despite a higher loss, as seen in the significantly higher AUPR on unseen attacks.