Emergent Linear Separability of Unseen Data Points in High-dimensional Last-Layer Feature Space

Taehun Cha CTH127@KOREA.AC.KR
Donghun Lee* HOLY@KOREA.AC.KR
Korea University, Seoul, Republic of Korea

Abstract

In this work, we investigate the emergence of linear separability for unseen data points in the high-dimensional last-layer feature space of deep neural networks. Through empirical analysis, we observe that, after training, in-distribution and out-of-distribution samples become linearly separable in the last-layer feature space when the hidden dimension is sufficiently high—even in regimes where the input data itself is not. We leverage these observations for the task of uncertainty quantification. By connecting our findings to classical support vector machine margin theory, we theoretically show that the separating hyperplane exhibits a higher weight norm when facing in-distribution data points. This work highlights linear separability as a fundamental and analyzable property of trained deep neural networks' representations, offering a geometric perspective on the practical uncertainty quantification task in neural networks.

1. Introduction

Retraining the last layer has been proposed as a solution to various problems. For instance, Kirichenko et al. [11] empirically showed that retraining only the last layer with a small amount of group-annotated data can effectively correct a model's reliance on spurious correlations. Likewise, Kang et al. [10] demonstrated that last-layer retraining can effectively mitigate the effects of label imbalance in training data. Since the last layer is typically a linear transformation, these findings suggest that the last-layer encodings are linearly separable enough to distinguish each target data class. This naturally raises the question: *Does the geometry of linear separability in this space also emerge in other settings, such as separating seen and unseen data points?*

The significance of this question lies in its direct applications, such as uncertainty quantification. Uncertainty quantification aims to address this problem by distinguishing between data the model is familiar with and data it has rarely or never seen. While traditional uncertainty quantification methods often rely on probabilistic modeling or Bayesian inference, the linear separability in the last-layer feature can suggest that a fundamentally geometric approach may also be effective.

In this study, we explore how linear separability arises for unseen data points within the high-dimensional last-layer feature space of deep neural networks. Surprisingly, we observe that the linear separability of unseen data points is a pervasive phenomenon in neural networks trained by standard procedures, occurring in both classification and regression settings. We also find that this phenomenon also emerges in high-dimensional real image datasets.

Building on this observation, we propose a novel uncertainty quantification method inspired by the linear separability of last-layer encodings. Specifically, we measure how easily a test point can

* Corresponding author

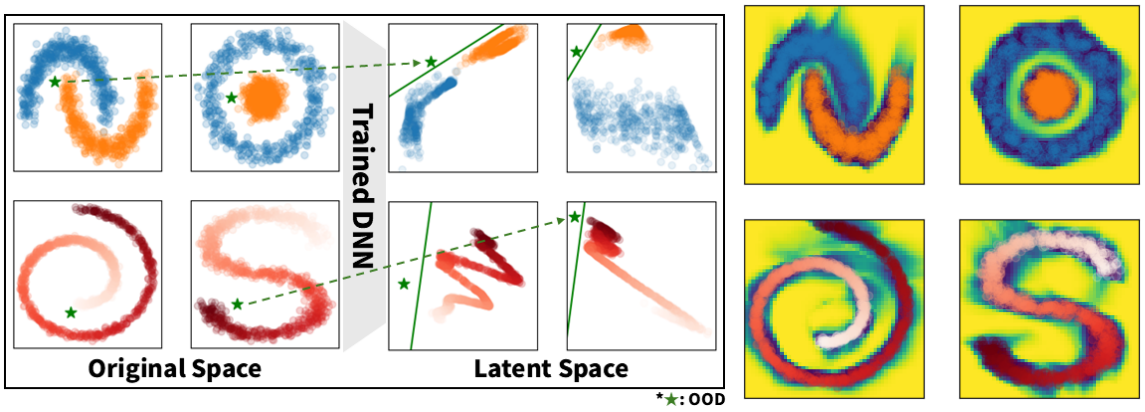

Figure 1: Left: Input data (leftmost) are encoded by a deep neural network (DNN) into a linearly separable latent space (middle), where in-distribution (ID) and out-of-distribution (OOD) test samples (green stars) can be separated by a hyperplane (green lines). Right: Visualization of the linear logistic regressor's predictions, where yellow indicates higher OOD probability. These results imply that OOD data points are linearly separable in the last-layer feature space, across both classification (top) and regression (bottom) tasks.

be linearly separated from the training set in the last-layer encoding space. Leveraging classical support vector machine (SVM) theory, we theoretically demonstrate that the inverse of the norm of the SVM decision boundary vector—which separates the unseen encoding from the seen training encodings—can serve as a meaningful measure of uncertainty.

2. Last-layer Linear Separability of Unseen Data Points

To examine the last-layer linear separability of unseen data points, we first conduct toy experiments on synthetic datasets. As shown on the left side of Figure 1, we generate four toy datasets with various shapes: the top two are binary classification tasks, and the bottom two are regression tasks. We train two-layer DNNs for each task and compute the last-layer encodings over a grid covering the input domain. Grid points that are sufficiently close to any training sample are labeled as in-distribution, while the rest are considered out-of-distribution. We then train a linear logistic regressor on these encodings and visualize the predicted probability of being out-of-distribution on the right side of Figure 1.

Surprisingly, we observe that the last-layer encodings of in-distribution and out-of-distribution data points are linearly separable. Even more interestingly, this linear separability arises not only in models trained on classification tasks but also in those trained on regression tasks. This phenomenon is not limited to toy examples or a particular dataset: encodings from ResNet [7], pre-trained on CIFAR-10 and CIFAR-100 [12], are nearly perfectly linearly separable from those of the SVHN dataset [19], achieving AUROC scores of 0.9996–0.9998.

To further analyze this phenomenon, we vary the latent feature dimension (16, 32, 64, 128, and 256) and visualize the OOD probability in Figure 2. Interestingly, we observe that last-layer linear

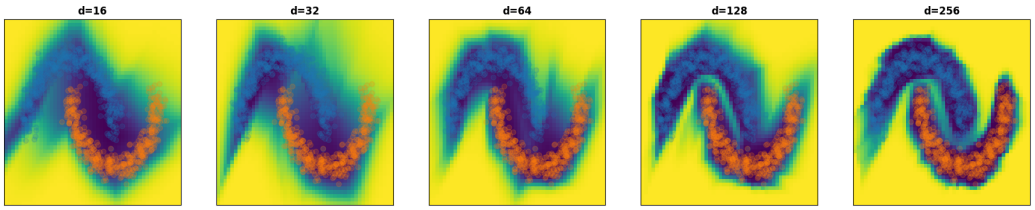

Figure 2: Visualization of OOD probability as the latent feature dimension increases (from left to right: $d = 16, 32, 64, 128, 256$.). As the latent dimension grows, the last-layer feature space exhibits increasingly clear linear separability between in-distribution (blue) and out-of-distribution regions (yellow).

separability emerges when the latent dimension is sufficiently high. This trend is particularly evident when examining the points between the two moon-shaped clusters. The result is intuitive: as the latent feature dimension increases, the last-layer features become more separable, since there are more degrees of freedom to distinguish between ID and OOD data points.

3. Linear Separability for Uncertainty Quantification

To verify the extensibility of this phenomenon to real-world tasks, we propose a novel uncertainty quantification method based on our findings. We begin by formally defining the problem of uncertainty quantification in deep neural networks. Let $\mathbb{P}_{\text{id}}(x, y)$ and $\mathbb{P}_{\text{id}}(x)$ denote the joint and the marginal *in-distribution* probability distributions, respectively, where $x \in \mathcal{X}_{\text{id}}$ and $\mathcal{X}_{\text{id}}$ is the support of the in-distribution distribution. Let $\mathbb{P}_{\text{ood}}(x)$ denote the marginal *out-of-distribution* distribution with domain $x \in \mathcal{X}_{\text{ood}}$. Our goal is to construct an uncertainty measure $u(x)$ that assigns higher values to out-of-distribution inputs, i.e., $u(x) \leq u(x')$ for $x \sim \mathbb{P}_{\text{id}}(x)$ and $x' \sim \mathbb{P}_{\text{ood}}(x)$.

We usually say that two d-dimensional real vectors $e, e' \in \mathbb{R}^d$ are *linearly separable* if there exists a hyperplane that separates them. However, mere separability between the two sets does not necessarily convey uncertainty. To measure the degree of separability, we propose a simple algorithm based on SVM theory [3]. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with binary labels $y_i \in \{-1, 1\}$, the (hard-margin) linear SVM solves a convex optimization problem that minimizes $\|w\|_2$ subject to the constraint $y_i(\langle w, x_i\rangle - b) \geq 1$ for all $i = 1, \ldots, N$. Intuitively, the SVM finds a hyperplane $\langle w, x\rangle - b = 0$ that linearly separates the two classes while maximizing the margin between them. Since the distance from a point x to the hyperplane is given by $\frac{\langle w,x\rangle - b}{\|w\|_2}$, and points on the decision boundary satisfy $\langle w, x\rangle - b = 1$, minimizing $\|w\|_2$ is equivalent to maximizing the margin, which equals $\|w\|_2^{-1}$.

We propose a measure of *linear separability* between two points, defined as $l(x', x) = \|w\|_2^{-1}$, where w is the solution to the hard-margin linear SVM problem with hypothetical labels $y' = +1$ and $y = -1$. This notion can be extended to measure the linear separability between a point and a set: $l(x', \mathcal{X}) = \|w\|_2^{-1}$, where the solution satisfies $\langle w, x'\rangle - b \geq 1$ and $\langle w, x\rangle - b \leq -1$ for all $x \in \mathcal{X}$. If no such w exists, we define $l(x', \mathcal{X}) = 0$.

Building on this definition, we define an uncertainty measure. Given a training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N \sim \mathbb{P}_{\text{id}}(x, y)$, we compute an encoding set $\mathcal{E} = \{e_i\}_{i=1}^N$, where $e_i = \phi(x_i)$ for all $i = 1, \ldots, N$ and ϕ is a pre-trained encoder. For a test point x^*, we measure the uncertainty

$l_{\mathcal{E}}(x^*) := l(\phi(x^*), \mathcal{E})$ — that is, the margin obtained by separating $\phi(x^*)$ from the training encodings $\mathcal{E}$ using a hard-margin linear SVM. Intuitively, this quantity reflects the distance from x to the hyperplane that separates it from the convex hull of the training encodings. Importantly, this notion differs from Euclidean distance: a point near the training set in Euclidean space may still be linearly separable, while a point farther away may not be.

We investigate whether the proposed uncertainty measure $l_{\mathcal{E}}$ satisfies a key property for uncertainty quantification: that $l_{\mathcal{E}}(x) \leq l_{\mathcal{E}}(x')$ for $x \sim \mathbb{P}_{\text{id}}(\text{x})$ and $x' \sim \mathbb{P}_{\text{ood}}(x)$. Importantly, we do not require the in-distribution and out-of-distribution sets to be completely linearly separable. Instead, we show that this property holds even when the two sets are only *partially linearly separable*, which is a much weaker and more realistic condition in high-dimensional settings.

Theorem 1 *Let $\mathbb{P}_{id}$ and $\mathbb{P}_{ood}$ be two probability distributions over $\mathbb{R}^d$, and suppose there exists a hyperplane defined by $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ such that $\langle w, x \rangle < b$, $\mathbb{P}_{id}$ almost surely and $\mathbb{P}_{ood}[\langle w, x' \rangle \geq b] > \epsilon$ for some $\epsilon > 0$. Then the $l_{\mathcal{E}}(x)$ has a lower upper bound than $l_{\mathcal{E}}(x')$ for $x \sim \mathbb{P}_{id}(x)$ and $x' \sim \mathbb{P}_{ood}(x')$.*

As SVM optimization fundamentally minimizes $\|w\|_2$ (equivalently maximizing $l_{\mathcal{E}}(x) = \|w\|_2^{-1}$), the theorem suggests that we can achieve higher $l_{\mathcal{E}}(x')$ than $l_{\mathcal{E}}(x)$, even when the in-distribution encoded point $\phi(x)$ is linearly separable from the training set encoding $\mathcal{E}$. This insight is supported by the statistical learning theory-style proposition, which is provided in Appendix A. Briefly speaking, we upper bound $l_{\mathcal{E}}$ with generalization error, which is higher for OOD data.

4. OOD detection on real-world datasets

We evaluate the performance of our uncertainty quantification method against several standard OOD detection approaches. For a fair comparison, we only include methods that do not require additional training or fine-tuning. In other words, all methods are implemented using models pre-trained on the in-distribution dataset with the standard cross-entropy loss.

First, as a baseline, we implement 3 methods, including maximum softmax probability [8] (**MSP**), maximum logit [9] (**MaxLogit**), and energy score [17] (**Energy**). These baseline methods only utilize output logit values, without access to the last layer encodings. Second, we compare 2 distance-based methods, **MDS** [15] and **KNN** [20]. For a more comprehensive comparison, we also include 4 additional methods: an input perturbation-based approach [16] (**ODIN**), a dropout-based Bayesian approximation [6] (**Dropout**), virtual-logit based method [23] (**VIM**), and the most recent pruning-based activation scaling method [24] (**Scale**). In total, we compare 9 methods.

For our method (**Geo**), as it requires solving an SVM optimization problem over the entire training set, we introduce a simple yet effective sampling-based approximation. Rather than using the full dataset, we randomly sample $M \ll N$ training embeddings, specifically from the class predicted by the model in classification problem. We set $M = 300$ throughout all experiments. We also extend our method to consider model's confidence, by computing $l_{\mathcal{E}}^{\text{logit}}(x) = [\|w\|_2 \cdot \max_i [f(x)]_i]^{-1}$, where $[f(x)]_i$ denotes the i-th logit output of the classifier (**GeoLogit**). We summarize our base and modified algorithms in Appendix B.

We utilize the OpenOOD v1.5 framework [26] to implement each method. OpenOOD categorizes OOD datasets into two groups: Near OOD and Far OOD. As ImageNet-1K [4] is used as the in-distribution dataset, SSB-hard [22] and NINCO [1] are used as Near OOD datasets, while iNaturalist

	DenseNet				WideResNet				ViT			
	Near OOD		Far OOD		Near OOD		Far OOD		Near OOD		Far OOD	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	65.56	76.54	51.55	85.53	75.01	75.80	53.93	85.85	78.72	71.77	54.23	84.10
MaxLogit	68.38	76.69	44.14	88.65	83.81	72.76	77.28	82.42	85.69	69.54	65.51	83.48
Energy	68.91	76.01	44.78	88.31	85.66	68.27	82.42	76.18	86.53	66.54	70.26	80.86
ODIN	74.33	74.31	50.55	87.91	87.72	63.71	87.75	64.09	88.08	61.22	80.22	75.97
Dropout	65.56	76.54	51.55	85.53	75.01	75.80	53.93	85.85	78.72	71.77	54.23	84.10
MDS	87.13	54.64	56.88	78.20	69.92	74.16	**26.43**	93.06	71.79	**75.30**	**33.77**	**91.01**
KNN	69.33	70.19	43.78	84.52	66.25	75.68	27.35	**93.22**	71.37	70.67	38.68	85.10
VIM	71.63	70.94	31.91	88.65	78.72	69.96	28.06	92.48	81.31	71.88	35.82	90.52
Scale	65.41	**78.56**	**31.55**	**92.83**	93.06	45.60	85.74	62.90	84.48	67.12	66.69	81.11
Geo	67.54	75.11	43.04	88.49	66.24	77.02	30.22	92.93	70.58	72.48	42.19	85.99
GeoLogit	**63.56**	78.25	34.20	91.19	**63.70**	**78.47**	30.85	91.54	**69.79**	73.70	38.74	87.35

Table 1: OOD detection performance on the ImageNet dataset (as an in-distribution) using DenseNet-201, WideResNet-50, and ViT-B/32 as backbone models. We report the mean and standard deviation of FPR@95 (%, ↓ lower is better) and AUROC (%, ↑ higher is better). Bold, real, and dashed underlined values denote the best, second-best, and third-best results.

[21], Textures [2], and OpenImage-O [23] are used as Far OOD datasets. We present the results on Table 1.

The result highlights the strong generalizability of our methods, Geo and GeoLogit, to various model architectures. Our algorithms consistently rank among the top performers across all OOD settings. For example, with WideResNet, GeoLogit and Geo achieve the best and second-best FPR@95 and AUROC scores on Near OOD setting, outperforming all baselines. The result shows that our method adapts well across diverse architectures. These results validate the applicability of the last-layer linear separability for uncertainty quantification in large-scale scenarios.

We further investigate the performance of our method on CIFAR datasets [12] in Appendix C. We also present the performance of our method on the regression task in Appendix D.

5. Discussion and Limitation

In the context of calibration, a model is expected to produce well-calibrated probability estimates—meaning its predicted confidence should align with actual accuracy. Since our uncertainty scores do not represent probabilities in the $[0, 1]$ range, we assess calibration by dividing the test set into quantile-based bins based on uncertainty and computing the accuracy within each bin. We find that our uncertainty measure shows a consistent trend: accuracy generally decreases as uncertainty increases, closely aligning with the behavior of the Ensemble method. We present these calibration results in Appendix E.

Core limitations of our method are memory and time inefficiency: our method requires a user to save the last layer encoding of the training dataset, and it requires computing the last-layer encoding vectors for the entire training set in advance. This is a common limitation of feature-based methods, including MDS [18] and KNN [20]. To alleviate this, one can consider memory-efficient approximations such as encoding compression, prototype selection, or clustering-based summarization of the training encodings. On the other hand, to further reduce computation time, one possible extension is to implement batch-wise uncertainty quantification. While these skills

may slightly reduce accuracy, they offer a practical trade-off for deploying uncertainty estimation in resource-constrained environments.

Acknowledgment

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1102828).

References

[1] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 2471–2506. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bitterwolf23a.html.

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. Machine learning, 20:273–297, 1995.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[5] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

[9] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for

real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 8759–8773. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html.

[10] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

[11] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30, 2017.

[14] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS 231N, 7(7):3, 2015.

[15] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

[16] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

[17] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf.

[18] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-), 80:S1–S7, 2018.

[19] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[20] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 20827–20840. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

[21] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[22] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=5hLP5JY9S2d.

[23] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[24] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=RDSTjtnqCg.

[25] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. CoRR, abs/1605.07146, 2016. URL http://arxiv.org/abs/1605.07146.

[26] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. Journal of Data-centric Machine Learning Research, 2024. URL https://openreview.net/forum?id=cnnTnJQigs. Dataset Certification.

[27] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.

Appendix A. Proof of Theorem 1

First, we restate our target theorem.

Theorem 2 *Let $\mathbb{P}_{id}$ and $\mathbb{P}_{ood}$ be two probability distributions over $\mathbb{R}^d$, and suppose there exists a hyperplane defined by $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ such that $\langle w, x\rangle < b$, $\mathbb{P}_{id}$ almost surely and $\mathbb{P}_{ood}(\langle w, x'\rangle \geq b) > \epsilon$ for some $\epsilon > 0$. Then the uncertainty measure $l_{\mathcal{E}}(x)$ has a lower upper bound than $l_{\mathcal{E}}(x')$ for $x \sim \mathbb{P}_{id}(x)$ and $x' \sim \mathbb{P}_{ood}(x')$.*

To prove the theorem, we need one proposition and one lemma.

Proposition 3 *Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be a dataset with a deterministic labelling scheme, $y_i = -1, \forall i = 1, \ldots, N-1$ and $y_N = +1$. Assume each $x_i \overset{iid}{\sim} \mathbb{P}[x|y=-1], \forall i = 1, \ldots, N-1$ and $x_N \sim \mathbb{P}[x|y=+1]$ with $x_i \in \mathbb{R}^d$ and $\|x_i\|_2 \leq B$. Assume there exists a linear classifier $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ s.t. $\langle x_i, w\rangle - b < -1, \forall i = 1, \ldots, N-1$ and $\langle x_N, w\rangle - b \geq +1$. For the zero-one loss function $L(x, y; w, b) = \mathbf{1}_{\{sign[\langle x,w\rangle - b] \neq y\}}$, with probability $1-\delta$, the following holds,*

$$\mathbb{E}[L(x, y; w, b)] < C_1 \|w\|_2 + C_0,$$

where the constant $C_1 = BC_0, C_0 = \frac{1+\sqrt{N-1}}{N}\sqrt{2^{-1}ln(2\delta^{-1})}$.

In statistical learning theory on SVM, the weight norm $\|w\|_2$ is typically assumed to be upper bounded, and this bound is used to derive an upper bound on the generalization error. In contrast, we use a similar reasoning in reverse: we show that the norm of the weight vector is lower bounded by the generalization error.

Proof Without loss of generality, we ignore the b term, as we can simply concatenate it as an element of w with concatenating 1 to x. Define a hinge loss function, $h(x, y; w) = \max(0, 1 - y\langle x, w\rangle)$. We define two versions, $h^+(x) = \max(0, 1 - \langle x, w\rangle)$, $h^-(x) = \max(0, 1 + \langle x, w\rangle)$. Define a sample version of the hinge loss:

$$\hat{h}^+(\mathcal{X}_n) = \frac{1}{n}\sum_{i=1}^{n} h^+(x_i), \quad \text{where } x_i \in \mathcal{X}_n,$$

where $\mathcal{X}_n$ is an n-sized subsample set and x_is are from $\mathcal{D}$. Likewise, define $\hat{h}^-(\mathcal{X}_n)$. Set $\mathcal{X}_1^+ = \{x_N\}$ and $\mathcal{X}_{N-1}^- = \{x_1, \ldots, x_{N-1}\}$, so that $\hat{h}^+(\mathcal{X}_1^+) = 0$ and $\hat{h}^-(\mathcal{X}_{N-1}^-) = 0$. For a $\mathcal{X}_n$ s.t. $\hat{h}^+(\mathcal{X}_n) = 0$, define $\mathcal{X}_n' = \{x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n\}$, i.e. one element of $\mathcal{X}_n$ is altered. Then,

$$\begin{aligned}
\sup_{x'} |\hat{h}^+(\mathcal{X}_n) - \hat{h}^+(\mathcal{X}_n')| &= \frac{1}{n}\sup_{x_i'} |h^+(x_i')| \\
&\leq \frac{1}{n}\sup_{x_i'} |1 - \langle x_i', w\rangle| \\
&\leq \frac{1}{n}\sup_{x_i'} 1 + |\langle x_i', w\rangle| \quad \text{by Triangle inequality} \\
&\leq \frac{1}{n}\sup_{x_i'} 1 + \|x_i'\|_2 \cdot \|w\|_2 \quad \text{by Cauchy-Schwarz inequality} \\
&\leq \frac{1 + B\|w\|_2}{n} \quad \text{by assumption}
\end{aligned}$$

This holds regardless of the altered position i. Define $\mathbb{P}^+ = \mathbb{P}[\cdot|y=+1]$ and $\mathbb{E}^+[\cdot] = \mathbb{E}[\cdot|y=+1]$. Similarly define $\mathbb{P}^-$ and $\mathbb{E}^-$. As each x_i's are sampled independently, we can apply McDiarmid's inequality. For any $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}^+\left[\hat{h}^+(\mathcal{X}_n) - \mathbb{E}^+\left[\hat{h}^+(\mathcal{X}_n)\right] < -\epsilon\right] &\leq \exp\left[-\frac{2\epsilon^2}{n(n^{-1}(1+B\|w\|_2))^2}\right] \\ &= \exp\left[-\frac{2\epsilon^2}{n^{-1}(1+B\|w\|_2)^2}\right]\end{aligned}$$

i.e. with a probability $1 - \frac{\delta}{2}$,

$$\mathbb{E}^+\left[\hat{h}^+(\mathcal{X}_n)\right] < (1+B\|w\|_2)\sqrt{(2n)^{-1}\ln(2\delta^{-1})},$$

as we assumed $\hat{h}^+(\mathcal{X}_n) = 0$. As each data points in $\mathcal{X}_1^+$ and $\mathcal{X}_{N-1}^-$ are sampled i.i.d. respectively, $\mathbb{E}^+\left[\hat{h}^+(\mathcal{X}_1^+)\right] = \mathbb{E}^+[h^+(x_N)] = \mathbb{E}^+[h^+(x)]$. (Recall $\mathbb{E}^+ := \mathbb{E}[\cdot|y=+1]$.) So,

$$\mathbb{E}^+\left[h^+(x)\right] < (1+B\|w\|_2)\sqrt{2^{-1}\ln(2\delta^{-1})}.$$

Likewise, with a probability $1 - \frac{\delta}{2}$,

$$\mathbb{E}^-\left[h^-(x)\right] < (1+B\|w\|_2)\sqrt{(2(N-1))^{-1}\ln(2\delta^{-1})},$$

and both event occurs at least with a probability $1-\delta$.
So

$$\begin{aligned}\mathbb{E}[h(x,y;w)] &= \mathbb{P}(y=+1)\mathbb{E}[h(x,y;w)|y=+1] + \mathbb{P}(y=-1)\mathbb{E}[h(x,y;w)|y=-1] \\ &= \mathbb{P}(y=+1)\mathbb{E}^+\left[h^+(x)\right] + \mathbb{P}(y=-1)\mathbb{E}^-\left[h^-(x)\right] \\ &< \frac{1}{N}(1+B\|w\|_2)\sqrt{2^{-1}\ln(2\delta^{-1})} + \frac{N-1}{N}(1+B\|w\|_2)\sqrt{(2(N-1))^{-1}\ln(2\delta^{-1})} \\ &= \frac{1+\sqrt{N-1}}{N}(1+B\|w\|_2)\sqrt{2^{-1}\ln(2\delta^{-1})}\end{aligned}$$

and the zero-one loss L is dominated by the hinge loss function h, we obtain the result.

■

Now, we will demonstrate that the generalization error is higher when the test point $x_N = x \sim \mathbb{P}_{\text{id}}(x)$ is drawn from the same distribution as the in-distribution training set $x_i \sim \mathbb{P}_{\text{id}}(x)$, than when $x_N = x' \sim \mathbb{P}_{\text{ood}}(x)$ comes from an out-of-distribution distribution. This result aligns with intuition: when x_N is from a different distribution, it is easier to distinguish it from the in-distribution samples, yielding a lower classification error and thus a smaller lower bound on $\|w\|_2$. Conversely, when x_N is from the same distribution, the classifier struggles more, leading to a higher error and hence a higher lower bound on $\|w\|_2$. As a result, the upper bound of uncertainty measure $l_{\mathcal{E}}(x) = \frac{1}{\|w\|_2}$ is lower for in-distribution points than for out-of-distribution points, if they share the same constants C_0 and C_1.

Lemma 4 *Define two probability distributions $\mathbb{P}$ and $\mathbb{P}'$, such that, $\mathbb{P}'[x|y=+1] = \mathbb{P}'[x|y=-1] = \mathbb{P}'(x)$ (i.e. random label) and $\mathbb{P}[x|y=+1] \neq \mathbb{P}[x|y=-1] = \mathbb{P}'(x)$. Define the corresponding expectations, $\mathbb{E}$ and $\mathbb{E}'$. Assume the marginal distribution of label is equal, i.e. $\mathbb{P}(y) = \mathbb{P}'(y)$ with $\mathbb{P}'(y=+1) \leq \mathbb{P}'(y=-1)$. Assume $\exists (w,b) \in \mathbb{R}^d \times \mathbb{R}$ such that $\langle w, x\rangle < b$, $\mathbb{P}'(x)$ almost surely, while $\mathbb{P}(\langle w, x\rangle \geq b) > \epsilon$ for some $\epsilon > 0$. Then for the zero-one loss function $L(x,y;w,b) = \mathbf{1}_{\{sign[\langle x,w\rangle - b] \neq y\}}$, the following holds,*

$$\min_{w,b} \mathbb{E}\left[L(x,y;w,b)\right] < \min_{w',b'} \mathbb{E}'\left[L(x,y;w',b')\right]$$

Proof For the random label case, minimum loss can be achieved by the majority vote. If $\mathbb{P}'(y=+1) = p, \mathbb{P}'(y=-1) = 1-p$, then

$$\min_{w',b'} \mathbb{E}'\left[L(x,y;w',b')\right] = \min(p, 1-p) = p,$$

as we assumed $p < 1-p$. Set w and b which satisfy our assumption. Then we obtain

$$\begin{aligned}
\mathbb{E}\left[L(x,y;w,b)|y=-1\right] &= \int \mathbf{1}_{\{\langle x,w\rangle > b\}} d\mathbb{P}\left[x|y=-1\right] \\
&= \int \mathbf{1}_{\{\langle x,w\rangle > b\}} d\mathbb{P}'(x) \\
&= 0 \quad \text{by assumption.}
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}\left[L(x,y;w,b)|y=+1\right] &= \int \mathbf{1}_{\{\langle x,w\rangle < b\}} d\mathbb{P}\left[x|y=+1\right] \\
&< 1-\epsilon \quad \text{by assumption.}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[L(x,y;w,b)\right] &= \mathbb{P}(y=+1)\mathbb{E}\left[L(x,y;w,b)|y=+1\right] + \mathbb{P}(y=-1)\mathbb{E}\left[L(x,y;w,b)|y=-1\right] \\
&< p(1-\epsilon) \\
&< p = \min_{w',b'} \mathbb{E}'\left[L(x,y;w',b')\right]
\end{aligned}$$

■

From the Theorem 4, we can regard $\mathbb{P}'(x)$ as an in-distribution data distribution with random label y. Conversely, $\mathbb{P}[x|y=+1]$ represents an out-of-distribution data distribution with a proper OOD label. Now we can prove the main theorem.

Proof [*Proof of Theorem 1*] By setting $\mathbb{P}' = \mathbb{P}_{\text{id}}$ and $\mathbb{P} = \mathbb{P}_{\text{ood}}$, we assume the existence of (w,b) which linearly separate two distributions, at least with ϵ probability. So $\min_{w,b} \mathbb{E}_{\text{ood}}\left[L(x,y;w,b)\right] < \min_{w',b'} \mathbb{E}_{\text{id}}\left[L(x,y;w',b')\right]$ by Theorem 4. We have already showed $\mathbb{E}_{\text{id}}\left[L(x,y;w',b')\right] < C_1 \left\|w'\right\|_2 + C_0$ and $\mathbb{E}_{\text{ood}}\left[L(x,y;w,b)\right] < C_1 \left\|w\right\|_2 + C_0$ on Theorem 3 and trivially $\min_{w',b'} \mathbb{E}_{\text{id}}\left[L(x,y;w',b')\right] \leq \mathbb{E}_{\text{id}}\left[L(x,y;w',b')\right]$ and $\min_{w,b} \mathbb{E}_{\text{ood}}\left[L(x,y;w,b)\right] \leq \mathbb{E}_{\text{ood}}\left[L(x,y;w,b)\right]$. As the two cases share the same constants C_0 and C_1 by setting, the lower bound of $\left\|w'\right\|_2$ trained on in-distribution dataset is higher than the lower bound of $\left\|w\right\|_2$ trained on out-of-distribution dataset. Conversely, $l_{\mathcal{E}}(x')$ for in-distribution x' has a lower upper bound than $l_{\mathcal{E}}(x)$ for out-of-distribution x. ■

Appendix B. Our Algorithms

Here we present our algorithm computing uncertainty.

Algorithm 1 Uncertainty Quantification with Linear Separability

Input: pre-trained model and encoder f and ϕ, training input set $\mathcal{X}$, training data encoding set $\mathcal{E}$, test data point x^*, SVM optimizer $\mathcal{O}$, number of samples M
Output: uncertainty measure $l_{\mathcal{E}}(x^*)$ or $l_{\mathcal{E}}^{\text{logit}}(x^*)$

Sample M training data encoding $\mathcal{E}_M = \{e_i\}_{i=1}^M \subset \mathcal{E}$
(If classification task, choose e_is such that $f(x_i) \approx f(x^*)$)
Set a dataset $\mathcal{D} = \{(e_i, -1)\}_{i=1}^M \cup \{(\phi(x^*), +1)\}$
$\mathcal{D} \leftarrow$ Normalize $\mathcal{D}$
Compute SVM $w, b = \mathcal{O}(\mathcal{D})$
if $\nexists w, b$ **then**
    return 0
else if GeoLogit **then**
    return $l_{\mathcal{E}}^{\text{logit}}(x^*) = \frac{1}{\|w\|_2 \cdot \max_i [f(x^*)]_i}$
else
    return $l_{\mathcal{E}}(x^*) = \frac{1}{\|w\|_2}$
end if

Appendix C. Results on CIFAR Datasets

Here we present the results for CIFAR datasets. We also utilize the OpenOOD v1.5 framework [26] to implement each method. When CIFAR-10 or CIFAR-100 [12] is used as the in-distribution dataset, the other CIFAR dataset and TinyImageNet [14] serve as Near OOD datasets, while MNIST [5], SVHN [19], Textures [2], and Places365 [27] are considered Far OOD datasets.

We employ two types of models: ResNet [7] and WideResNet [25]. For ResNet, we use three pre-trained models provided by the OpenOOD framework [26], which achieve average accuracies of 95% on CIFAR-10 and 77% on CIFAR-100. For WideResNet, we train three models using a standard training scheme, achieving 96% and 80% accuracy on CIFAR-10 and CIFAR-100, respectively. Note that for the CIFAR datasets, we use three independently trained models, allowing us to compute the mean and standard deviation of each metric and to apply the Ensemble method [13]. We present the results on Table 2 and Table 3.

Our proposed methods, Geo and GeoLogit, consistently demonstrate strong OOD detection performance across both CIFAR-10 and CIFAR-100 in-distribution datasets, for both ResNet and WideResNet structures. In particular, GeoLogit achieves competitive or superior results across all near and far OOD scenarios, especially excelling in the more challenging CIFAR-100 setting. For example, on CIFAR-100 with WideResNet, Geo and GeoLogit record the best or second-best performance in Near OOD setting, outperforming several strong baselines such as KNN, VIM, and Scale. Notably, GeoLogit performs comparably to or better than the Ensemble method in many cases, without requiring training multiple models.

	CIFAR 10				CIFAR 100			
	Near OOD		Far OOD		Near OOD		Far OOD	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	48.17 (3.94)	88.03 (0.25)	31.72 (1.84)	90.73 (0.43)	54.80 (0.33)	80.27 (0.11)	58.70 (1.06)	77.76 (0.44)
MaxLogit	61.32 (4.65)	87.52 (0.47)	41.68 (5.27)	91.10 (0.89)	55.47 (0.66)	81.05 (0.07)	56.72 (1.33)	79.68 (0.57)
Energy	61.33 (4.65)	87.58 (0.46)	41.70 (5.33)	91.21 (0.92)	55.61 (0.61)	80.91 (0.08)	56.58 (1.38)	79.77 (0.61)
ODIN	81.69 (5.07)	79.35 (2.52)	64.79 (5.55)	85.38 (1.01)	58.08 (0.56)	79.85 (0.11)	58.96 (0.86)	79.25 (0.22)
Dropout	48.17 (3.94)	88.03 (0.25)	31.72 (1.84)	90.73 (0.43)	54.80 (0.33)	80.27 (0.11)	58.70 (1.06)	77.76 (0.44)
MDS	49.90 (3.96)	84.20 (2.40)	32.21 (3.39)	89.72 (1.36)	83.52 (0.60)	58.69 (0.09)	72.26 (1.56)	69.39 (1.39)
KNN	**33.99** (0.39)	**90.64** (0.20)	**24.28** (0.41)	92.96 (0.14)	61.23 (0.13)	80.18 (0.15)	53.65 (0.28)	82.40 (0.17)
VIM	44.84 (2.30)	88.68 (0.28)	25.06 (0.51)	**93.48** (0.24)	62.64 (0.27)	74.98 (0.13)	**50.73** (1.00)	81.70 (0.62)
Scale	93.79 (0.57)	53.95 (3.84)	92.23 (0.70)	52.41 (4.53)	74.68 (1.22)	74.30 (0.58)	64.57 (1.22)	79.83 (0.43)
Geo	39.21 (1.20)	89.74 (0.24)	26.77 (0.10)	92.26 (0.11)	56.85 (0.16)	80.73 (0.16)	51.50 (0.68)	**82.71** (0.18)
GeoLogit	39.07 (1.72)	89.96 (0.23)	25.94 (0.51)	92.74 (0.28)	**54.17** (0.35)	**81.48** (0.05)	51.85 (0.83)	81.52 (0.25)
Ensemble	34.30	90.34	22.80	92.82	51.34	82.33	55.39	79.47

Table 2: OOD detection performance on CIFAR-10 and CIFAR-100 (as an in-distribution) using ResNet-18 as the backbone.

	CIFAR 10				CIFAR 100			
	Near OOD		Far OOD		Near OOD		Far OOD	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	44.67 (1.84)	89.42 (0.06)	32.39 (2.05)	91.74 (0.21)	52.80 (0.47)	81.85 (0.11)	55.99 (1.81)	79.76 (0.47)
MaxLogit	56.47 (1.57)	89.30 (0.05)	38.81 (1.60)	92.28 (0.36)	54.27 (0.57)	82.35 (0.26)	54.73 (1.99)	81.38 (0.49)
Energy	56.47 (1.57)	89.38 (0.05)	38.79 (1.60)	92.38 (0.37)	54.33 (0.52)	82.31 (0.29)	54.57 (2.04)	81.67 (0.50)
ODIN	86.39 (0.15)	78.02 (0.43)	70.11 (0.53)	80.05 (2.16)	61.50 (0.25)	80.34 (0.24)	61.09 (2.10)	79.63 (0.89)
Dropout	44.69 (1.84)	89.42 (0.06)	32.40 (2.05)	91.74 (0.21)	52.80 (0.47)	81.85 (0.11)	55.99 (1.81)	79.76 (0.47)
MDS	39.56 (1.63)	89.89 (0.33)	19.25 (0.53)	94.80 (0.14)	55.99 (0.76)	81.63 (0.14)	49.47 (1.92)	83.84 (0.78)
KNN	**29.54** (0.09)	**92.09** (0.14)	21.48 (0.50)	94.01 (0.16)	54.34 (0.77)	82.55 (0.06)	53.45 (1.52)	82.36 (0.45)
VIM	37.05 (1.19)	91.16 (0.26)	**17.19** (0.96)	**95.89** (0.21)	54.27 (0.46)	81.09 (0.08)	**45.87** (1.78)	**84.57** (0.55)
Scale	88.57 (0.36)	79.31 (1.25)	67.29 (4.11)	85.09 (2.36)	75.49 (0.39)	75.83 (0.41)	53.48 (1.08)	84.43 (0.32)
Geo	31.39 (0.47)	91.67 (0.13)	21.96 (0.49)	93.86 (0.14)	52.31 (0.20)	82.93 (0.05)	51.57 (1.67)	82.73 (0.42)
GeoLogit	31.64 (0.46)	91.77 (0.11)	21.73 (0.44)	94.20 (0.16)	**51.92** (0.29)	**83.01** (0.12)	51.22 (1.52)	82.46 (0.41)
Ensemble	32.39	91.15	24.39	93.26	48.34	83.76	51.43	81.69

Table 3: OOD detection performance on CIFAR-10 and CIFAR-100 (as an in-distribution) using WideResNet-50 as the backbone.

Appendix D. Application to Regression Task

Unlike methods that rely on logit values from the final layer or decision-boundary-based approaches such as VIM, purely feature-based methods like MDS, KNN, and Geo are directly applicable to regression problems. In this section, we examine how these methods behave in regions of high uncertainty, particularly on unseen input domains.

We construct a regression task using a sine function, sampling 1,000 data points from the intervals $x \in \left[0, \frac{1}{2}\pi\right] \cup \left[\pi, \frac{3}{2}\pi\right]$. A two-layer neural network with 256 hidden units is trained for 30 epochs to fit this function. Consequently, the model has no exposure to the regions $x \in \left(\frac{1}{2}\pi, \pi\right) \cup \left(\frac{3}{2}\pi, 2\pi\right)$.

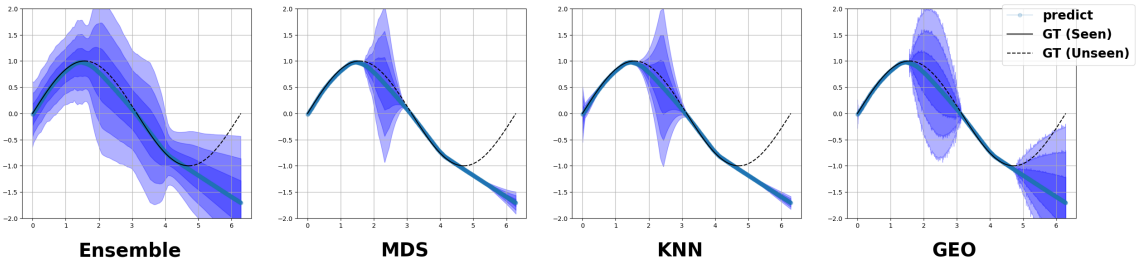

Figure 3: Uncertainty quantification results on a sine regression task using different methods: Ensemble, MDS, KNN, and Geo. The shaded blue regions represent the predictive uncertainty estimated by each method. The solid black line indicates the ground truth (GT) in the seen domain, while the dashed black line shows the GT in the unseen domain.

We expect uncertainty quantification methods to yield higher uncertainty values in these intervals. For comparison, we also implement an Ensemble method by training five additional models and using the prediction variance as a measure of uncertainty. The results are shown in Figure 3.

The Ensemble method shows generally high uncertainty regardless of whether the input lies within the seen or unseen domain. This behavior suggests that the Ensemble method struggles to capture uncertainty effectively in this simple sine function regression task. Meanwhile, both MDS and KNN—being distance-based methods—also exhibit limited ability to detect the rightmost unseen interval. Although they show a visible uncertainty peak in the middle unseen interval, they significantly underestimate uncertainty beyond $\frac{3}{2}\pi$, failing to recognize the out-of-distribution.

In contrast, our proposed Geo method demonstrates strong and localized uncertainty responses in both unseen regions. It produces distinct high-uncertainty bands in the unseen domain while maintaining low uncertainty within the seen domain. This behavior reflects Geo's ability to capture uncertainty with geometric separability from known data. Overall, these results highlight Geo's suitability for regression tasks and its ability to deliver accurate and interpretable uncertainty estimates.

Appendix E. Calibration Analysis

We assess the calibration of uncertainty estimates by binning test samples into 10 quantiles based on each method's confidence score (or inverse of uncertainty score) and computing the classification accuracy within each bin. This setup does not require the uncertainty to be probabilistic but tests whether lower-confidence predictions correspond to lower accuracy, as expected in well-calibrated systems.

As shown in Figure 4, on the CIFAR-10 dataset, all methods—including MDS, Geo, and Ensemble—exhibit relatively smooth, increasing trends, indicating that their confidence scores are moderately aligned with actual accuracy. However, on the more complex CIFAR-100 dataset, only Geo and Ensemble maintain a clear monotonic calibration trend, while MDS fails to distinguish uncertainty levels, particularly in the lowest quantiles. This suggests that Geo is not only effective in detecting out-of-distribution and uncertain samples, but also potentially provides well-ranked confidence estimates that are competitive with established methods like Ensembles.

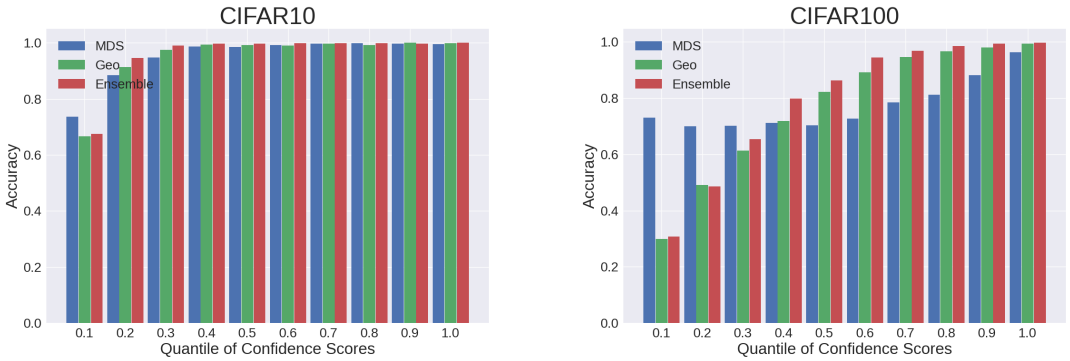

Figure 4: Calibration results on CIFAR-10 (left) and CIFAR-100 (right) using quantile-binned accuracy. Each method's uncertainty score is divided into 10 quantile bins, and the corresponding accuracy is computed per bin.