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

Taehun Cha and **Donghun Lee**
Department of Mathematics, Korea University

AIML@K

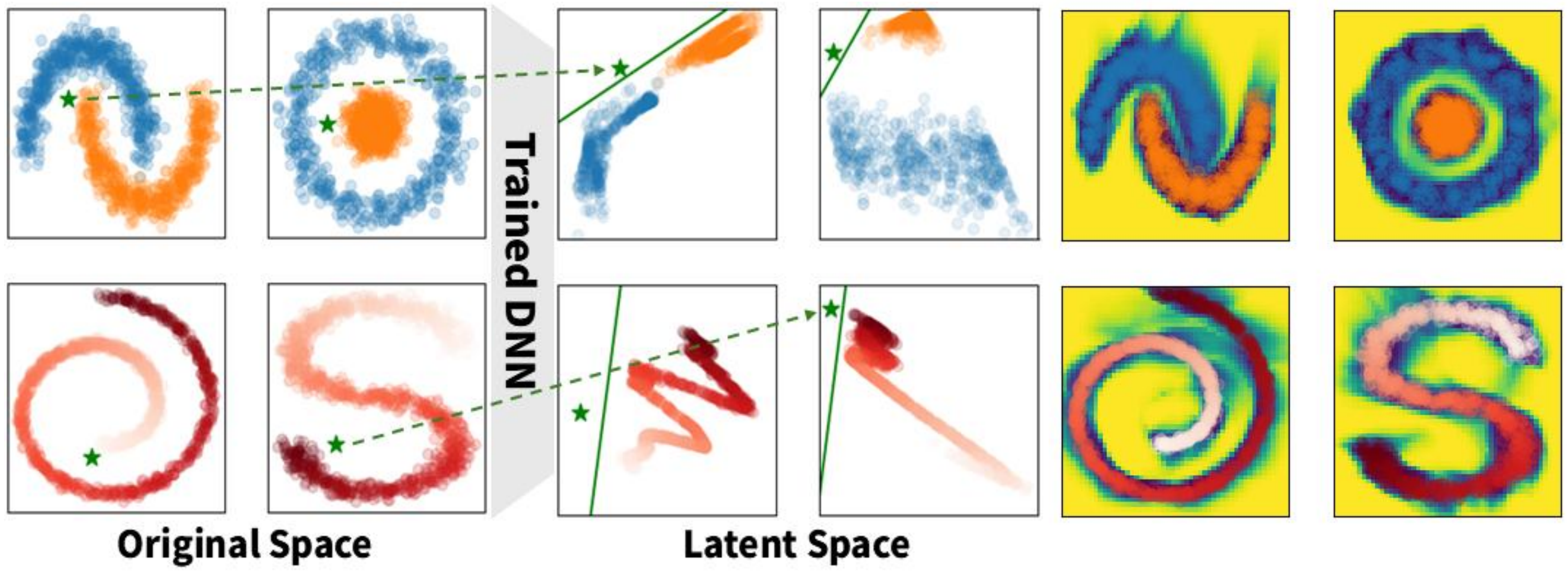
ICML International Conference On Machine Learning

Contributions

- Empirically verify the **emergent linear separability** between seen and unseen data points
- Theoretically show the relationship between "unseen-ness" and the separating hyperplane
- Based on these findings, we propose and evaluate the novel **geometric uncertainty**

Observation: Linear Separability

- Last-layer features of seen and unseen data points from a normally trained neural network are linearly separable
- This tendency strengthens as NN gets wider



Theory: Geometric Uncertainty

- **Notations**
  - P_{id}, P_{ood}: Probability distributions over R^d
  - $E = \{e_1, \dots, e_N\}$: A set of the last-layer encodings of training dataset, $e_i \in R^d$
  - $l_E(x) = \frac{1}{\|w\|}$, where $\langle \mathrm{w}, \phi(x) \rangle - b \geq +1$ and $\langle \mathrm{w}, e \rangle - b < -1, \forall e \in E$ if such w exists

    (In other words, w and b represent a separating hyperplane between x and E)

- **Theorem**
  - If $\langle \mathrm{w}, \phi(x) \rangle < b, P_{id}$ almost surely, and $P_{\mathrm{ood}}[\langle \mathrm{w}, \phi(x') \rangle > \geq b] > \epsilon$ for some $\epsilon > 0$,

    (In other words, there exists a (at least partial) domain mismatch between P_{id} and P_{ood})
  - Then, $UB(l_E(x)) < UB(l_E(x'))$ for $x \sim P_{id}, x' \sim P_{ood}$ and UB: upper bound

- As this characteristic is desirable for uncertainty quantification, we define $l_E(x)$ as the **Geometric Uncertainty**
- Intuitively, $\frac{1}{\|w\|}$ represents the distance between and the hyperplane and the closest data points
- This theorem is supported by the classical statistical learning theory-like proposition

- **Proposition**
  - Let $D = \{(x_i, y_i)\}_{i=1}^N$ with $y_i = -1, \forall i = 1, \dots, N-1$ and $y_N = +1$
  - Assume $x_i \sim P[x|y=-1], i.i.d.$ and $x_N \sim P[x|y=+1]$ with $x_i \in R^d$ and $\|x_i\| \leq B$
  - Assume the linear classifier $\langle \mathrm{x_i}, w \rangle - b < -1, \forall i = 1, \dots, N-1$ and $\langle \mathrm{x}_N, w \rangle + b \geq +1$
  - Then for zero-one loss $L(x, y; w, b) = 1_{\{\mathrm{sgn}[\langle x, w \rangle - b] \neq y\}}$, with probability $1 - \delta$,

    $E[L(x, y; w, b)] < C_1 \|w\| + C_0$, where C_0, C_1: constants

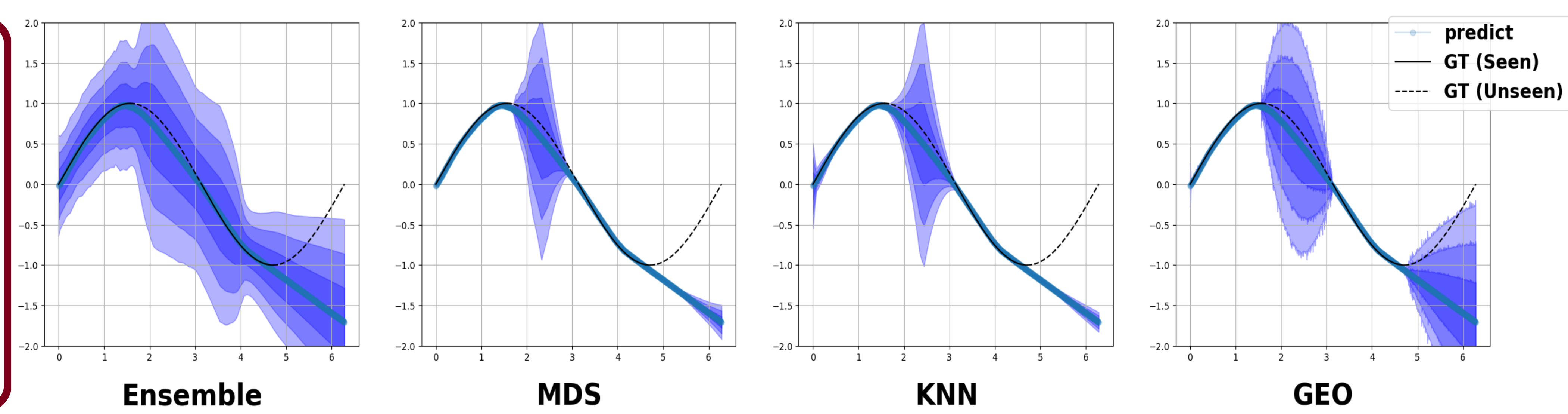
(Usually, this form of theory is used to bound the error term, but we reversely use it to bound $||w||$)

Experiments

- OOD Detection in image classification using (Wide-) ResNet (AUROC)

ID	CIFAR10		CIFAR100		ImageNet	
Target	Near	Far	Near	Far	Near	Far
MDS	89.89	**94.80**	81.63	**83.84**	74.16	93.06
KNN	**92.09**	94.01	82.55	82.36	75.68	**93.22**
Ours	91.77	94.20	**83.01**	82.46	**78.47**	91.54

- Sine function regression with unseen domains



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

I'm currently looking for a postdoctoral position in the mathematical foundations of DNNs and LLMs!

Personal