On the Lipschitz Continuity of Set Aggregation Functions and Neural Networks for Sets

Giannis Nikolentzos[1] Konstantinos Skianis[2]

[1]University of Peloponnese, Greece

[2]University of Ioannina, Greece

Introduction

- Complex datasets often decomposed into sets (or multisets) of simpler objects.
  - Computer Vision: objects as sets of 3d points (i. e., point clouds)
  - NLP: documents as sets of word embeddings
  - Graph Mining: graphs as sets of node embeddings
- Typically, a permutation invariant function is applied to aggregate (multi)set elements into a vector (e. g., sum, mean, max)
  ↪Common architectures follow this approach (e. g., DeepSets [Zaheer et al., NIPS'17])
- **Question**: Are these functions and architectures that employ these functions stable (from a Lipschitz perspective)?

Distance Functions for Unordered Multisets

Three functions considered (metrics for sets and pseudometrics for multisets):

- **Earth Mover's Distance (EMD)**: A measure of dissimilarity between two distributions:

$$d_{\text{EMD}}(X,Y) = \min_{\mathbf{F}\geq 0} \sum_{i=1}^{m}\sum_{j=1}^{n}[\mathbf{F}]_{ij}\,\|\mathbf{v}_i-\mathbf{u}_j\|_2, \quad \text{s.t.} \sum_{j=1}^{n}[\mathbf{F}]_{ij}=\frac{1}{m}, \sum_{i=1}^{m}[\mathbf{F}]_{ij}=\frac{1}{n}$$

- **Hausdorff Distance**: Defined as follows:

$$d_H(X,Y) = \max\big(h(X,Y),h(Y,X)\big) \quad \text{where } h(X,Y)=\max_{i\in[m]}\min_{j\in[n]}||\mathbf{v}_i-\mathbf{u}_j||_2$$

  $h(X,Y)$ denotes the maximum distance from a point in X to the closest point in Y.
- **Matching Distance**: Assigns elements of one multiset to elements of the other. The assignments are determined by a permutation of the elements of the larger multiset:

$$d_M(X,Y)=\begin{cases}M(X,Y) & \text{if } m\geq n\\ M(Y,X) & \text{otherwise.}\end{cases} \quad \text{where } M(X,Y)=\min_{\pi\in\mathfrak{S}_m}\left[\sum_{i=1}^{n}||\mathbf{v}_{\pi(i)}-\mathbf{u}_i||_2+\sum_{i=n+1}^{m}||\mathbf{v}_{\pi(i)}||_2\right]$$

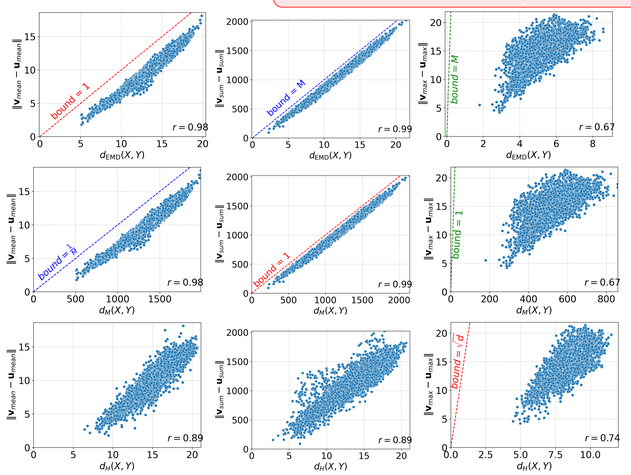
$\mathfrak{S}_m$ denotes the set of all permutations of a multiset with m elements.

Lipschitz Continuity of Aggregation Functions

	Sum	Mean	Max
EMD	$^{\dagger}L=M$	$L=1$	$^{\dagger}L=M$
Hausdorff	–	–	$L=\sqrt{d}$
Matching	$L=1$	$^{\dagger}L=1/M$	$^{\dagger}L=1$

Main result #1:

- Aggregation functions correspond to metrics:
  - EMD ⇝ Mean, Matching ⇝ Sum, Hausdorff ⇝ Max
- For fixed size (†), aggregations are Lipschitz continuous also with respect to other metrics.



Upper Bounds of Lipschitz Constants of Neural Networks for Sets

Neural network that operates on multisets:

$$\text{NN}_g(X) = f_2\Big(g\big(\{\!\{f_1(\mathbf{v}_1),\ldots,f_1(\mathbf{v}_m)\}\!\}\big)\Big)$$

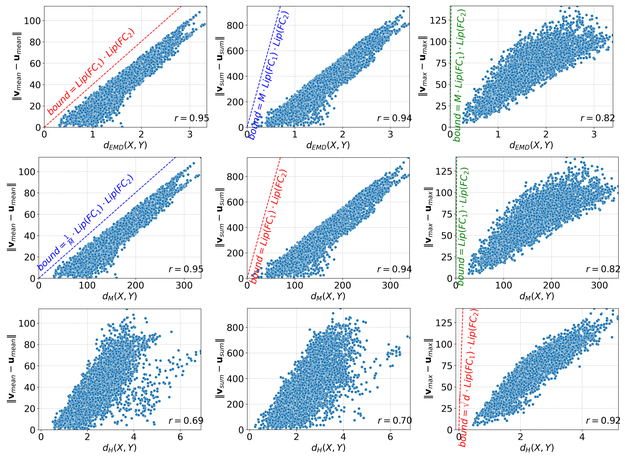
where $\{\!\{\mathbf{v}_1,\ldots,\mathbf{v}_m\}\!\}$ is the input multiset, g the employed aggregation function and f_1, f_2 neural network modules (e. g., MLPs)

	Sum	Mean	Max
EMD	$^{\dagger}L\leq \text{Lip}(f_1)\text{Lip}(f_2)M$	$L\leq \text{Lip}(f_1)\text{Lip}(f_2)$	$^{\dagger}L\leq \text{Lip}(f_1)\text{Lip}(f_2)M$
Hausdorff	–	–	$L\leq \text{Lip}(f_1)\text{Lip}(f_2)\sqrt{d}$
Matching	$^{\dagger}L\leq \text{Lip}(f_1)\text{Lip}(f_2)$	$^{\dagger}L\leq \text{Lip}(f_1)\text{Lip}(f_2)(1/M)$	$^{\dagger}L\leq \text{Lip}(f_1)\text{Lip}(f_2)$

Main result #2:

- The results for aggregations transfer to neural networks that operate on multisets:
  - This does **not** hold for the Sum function!
  - **Lipschitz continuity**: only guaranteed if multisets have fixed size (denoted by † in the Table).

Experiment: validate theoretical results



Stability under Perturbations of Input Multisets

Proposition: Given a multiset of vectors $X=\{\!\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}\!\}\in\mathcal{S}_{\leq M}(\mathbb{R}^d)$, let $X'=\{\!\{\mathbf{v}_1,\ldots,\mathbf{v}_n,\mathbf{v}_{n+1}\}\!\}\in\mathcal{S}_{\leq M}(\mathbb{R}^d)$ be the multiset where element $\mathbf{v}_{n+1}$ has been added to X, where $n+1\leq M$. Then,

1. The EMD between X and X' is bounded as $d_{\text{EMD}}(X,X')\leq\frac{1}{n(n+1)}\sum_{i=1}^{n}||\mathbf{v}_i-\mathbf{v}_{n+1}||$
2. The Hausdorff distance between X and X' is equal to $d_{\text{H}}(X,X')=\min_{i\in[n]}||\mathbf{v}_i-\mathbf{v}_{n+1}||$

Experiment: two types of perturbations:

- Pert.#1: add element that has the highest norm to each test sample of ModelNet40.
- Pert.#2: noise sampled from $\mathcal{U}(0,0.2)^d$ is added to each element (i. e., word vector) of each test sample of Polarity.

Measure average drop in accuracy after perturbation is applied to test samples.

Model	ModelNet40 Pert. #1	Polarity Pert. #2
NN_{mean}	2.0 (± 1.3)	13.6 (± 7.1)
NN_{max}	20.1 (± 1.8)	4.8 (± 3.7)

Generalization under Distribution Shifts

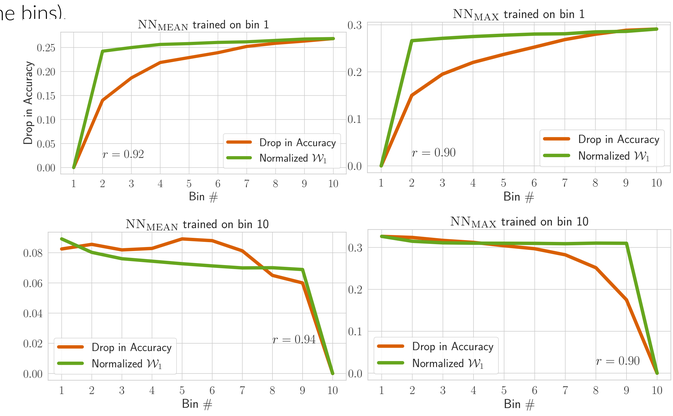
Theorem (Shen et al., AAAI'18): For all hypotheses $h\in\mathcal{H}$, the target error is bounded as:

$$\epsilon_T(h)\leq\epsilon_S(h)+2\,L\,\mathcal{W}_1(\mu_S,\mu_T)+\lambda$$

where L is the Lipschitz constant of h and λ is the combined error of the ideal hypothesis h^* that minimizes the combined error $\epsilon_S(h)+\epsilon_T(h)$.

Experiment:

- Sort documents from Polarity (i. e., multisets of word vectors) based on their cardinality.
- Construct 10 bins, each containing 200 multisets.
- Train models on first or last bin (source distribution) and compute accuracy on the rest of the bins (target distributions).
- Compute the Wasserstein distance with $p=1$ between domain distributions (i. e., between the first bin and the rest of the bins, and also between the last bin and the rest of the bins).



Conclusion

Contributions:

- Analysis of Lipschitz continuity of common aggregation functions under three multiset distance functions.
- Analysis of Lipschitz continuity of neural networks operating on multisets.
- Connections between Lipschitz continuity, robustness to perturbations, and generalization under distribution shift.

Main guideline: choose an aggregation function that is Lipschitz continuous with respect to the distance function that best reflects similarity in the dataset (might require domain knowledge).

Data Characteristic	Ideal Distance Metric	Aggregator
Shape and Boundary Outliers (e. g., 3D scans)	Hausdorff	MAX
Overall Distribution (e. g., documents represented as sets of word vectors)	EMD	MEAN
Direct 1-to-1 Element Correspondence (e. g., images represented as sets of keypoints)	Matching	SUM

Acknowledgements: We thank the anonymous reviewers for their helpful comments and suggestions. This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

nikolentzos@uop.gr