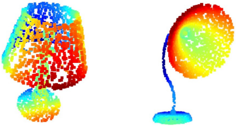
# On the Lipschitz Continuity of Set Aggregation Functions and Neural Networks for Sets

**Giannis Nikolentzos** and **Konstantinos Skianis**

University of Peloponnese
University of Ioannina

# Introduction

Complex datasets often decomposed into sets (or multisets) of simpler objects

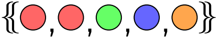


**Point clouds**

$\{\!\{$the, quick, brown, fox, jumps, over, lazy, dog $\}\!\}$

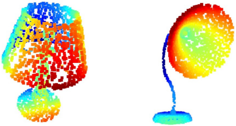
**Textual documents**

WL colors



**Graphs**

Complex datasets often decomposed into sets (or multisets) of simpler objects

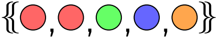


**Point clouds**

$\{\!\{$the, quick, brown, fox, jumps, over, lazy, dog $\}\!\}$

**Textual documents**

WL colors



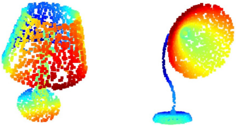
**Graphs**

Typically, a permutation invariant function is applied to aggregate (multi)set elements into a vector (e. g., sum, mean, max)

Most architectures, such as DeepSets [Zaheer et al., NIPS'17], use such functions

## Introduction

Complex datasets often decomposed into sets (or multisets) of simpler objects

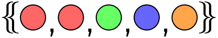


**Point clouds**

$\{\!\{$the, quick, brown, fox, jumps, over, lazy, dog $\}\!\}$

**Textual documents**

WL colors



**Graphs**

Typically, a permutation invariant function is applied to aggregate (multi)set elements into a vector (e. g., sum, mean, max)

Most architectures, such as DeepSets [Zaheer et al., NIPS'17], use such functions

**Question**: Are these functions and architectures that employ these functions stable (from a Lipschitz perspective)?

# Distance Functions for Unordered Multisets

# Distance Functions for Unordered Multisets

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

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

# Distance Functions for Unordered 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 $Y$.

# Distance Functions for Unordered Multisets

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

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

- **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 ($m \geq n$):

$$d_M(X, Y) = \min_{\pi \in \mathfrak{S}_m} \left[ \sum_{i=1}^{n} ||\mathbf{v}_{\pi(i)} - \mathbf{u}_i||_2 + \sum_{i=n+1}^{m} ||\mathbf{v}_{\pi(i)}||_2 \right]$$

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

Functions considered are metrics for sets and pseudometrics for multisets

# Lipschitz Continuity of Aggregation Functions

**Main result #1**:

- There is some correspondence between the three aggregation functions and the three metrics for unordered multisets.
  - Mean $\leftrightsquigarrow$ EMD
  - Sum $\leftrightsquigarrow$ Matching Dist.
  - Max $\leftrightsquigarrow$ Hausdorff Dist.
- If the multisets have fixed size (denoted by $^\dagger$ in the Table), the aggregation functions are Lipschitz continuous also with respect to other metrics.

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

# Upper Bounds of Lipschitz Constants of Neural Networks for Sets

Neural network that operates on multisets:

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

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)

**Main result #2**:

- The results for the aggregation functions 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 $^\dagger$ in the Table).

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

# Conclusion

**Experiments reported in the paper**:

- Validation of theoretical results
- Stability under perturbations of input multisets
- Generalization under distribution shifts

**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.