Magnitude Distance: A Geometric Measure of Dataset Similarity

Sahel Torkamani, Henry Gouk, Rik Sarkar
University of Edinburgh

Background and Motivation

Distance between Finite Datasets

The problem of measuring the similarity or distance between two finite datasets plays an important role in generative modelling:

- Evaluating the generative models performance by the similarity of generated samples with the reference dataset.
  - Such as Inception Score or the Maximum Mean Discrepancy (MMD).
- Providing a learning signal during the optimization of model parameters.
  - Such as Wasserstein Generative Adversarial Networks (WGANs).

Goal

Introduce a distance, measuring the dissimilarity between finite sets $X, Y \subset \mathbb{R}^D$, which is **outlier robust** and **captures geometric properties of the data**.

Magnitude of Metric Space

For a finite metric space, (X, d), we define the **similarity matrix** as $\zeta_X(x_i, x_j) := \exp(-d(x_i, x_j))$, for every $x_i, x_j \in X$.

A weighting of (X, d) is a function $w_X : X \to \mathbb{R}$ satisfying

$$\sum_{j \in X} \zeta_X(x_i, x_j)\, w_X(x_j) = 1$$

for every $x_i \in X$, where $w_X(x_i)$ is called the **magnitude weight**.

The **magnitude** of (X, d) is defined as $Mag(X, d) = \sum_{x_i \in X} w_X(x_i)$.

When X is a finite subset of $\mathbb{R}^D$, then ζ_X is invertible and magnitude is the sum of all the entries of the similarity matrix's inverse.

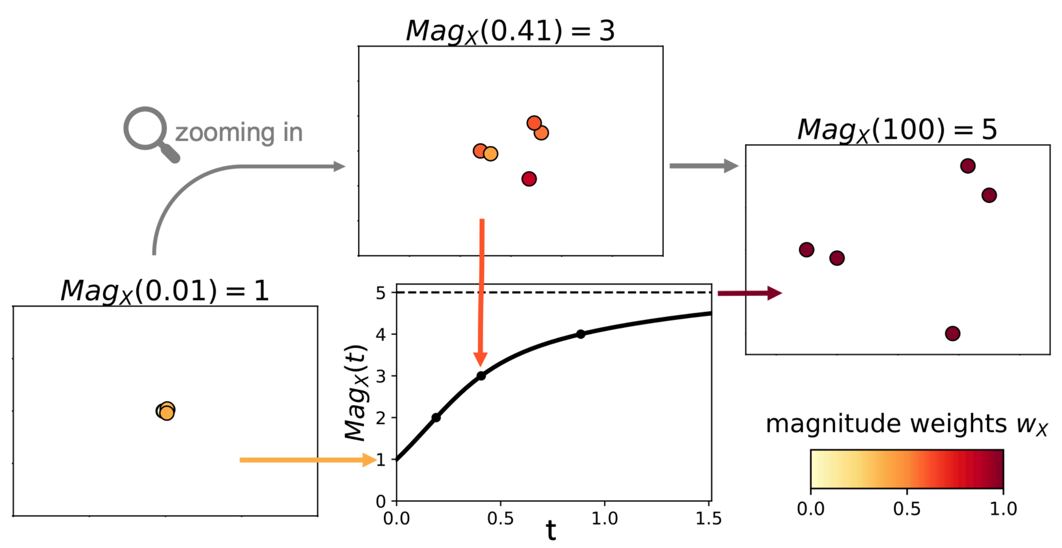


Image Credit: Limbeck, Katharina, et al.(2023).

Magnitude Function

For scaling parameter $t \in \mathbb{R}_+$, the **scaled metric space** (tX, d_t) is the metric with the same points as X and metric $d_t(x, y) = t \cdot d(x, y)$.

The **magnitude function** assigns each finite metric space X to a family of scaled metric spaces $\{tX\}_{t>0}$ by $Mag_X(t) = Mag(tX)$.

Magnitude Distance

Magnitude Distance

In the literature, a metric space is understood to be a **set** of distinct points, i.e., without duplicates. By extending the notion of magnitude to finite **collections** of points that may contain duplicates, we define the magnitude distance for every two finite two collection of point in $\mathbb{R}^D$.

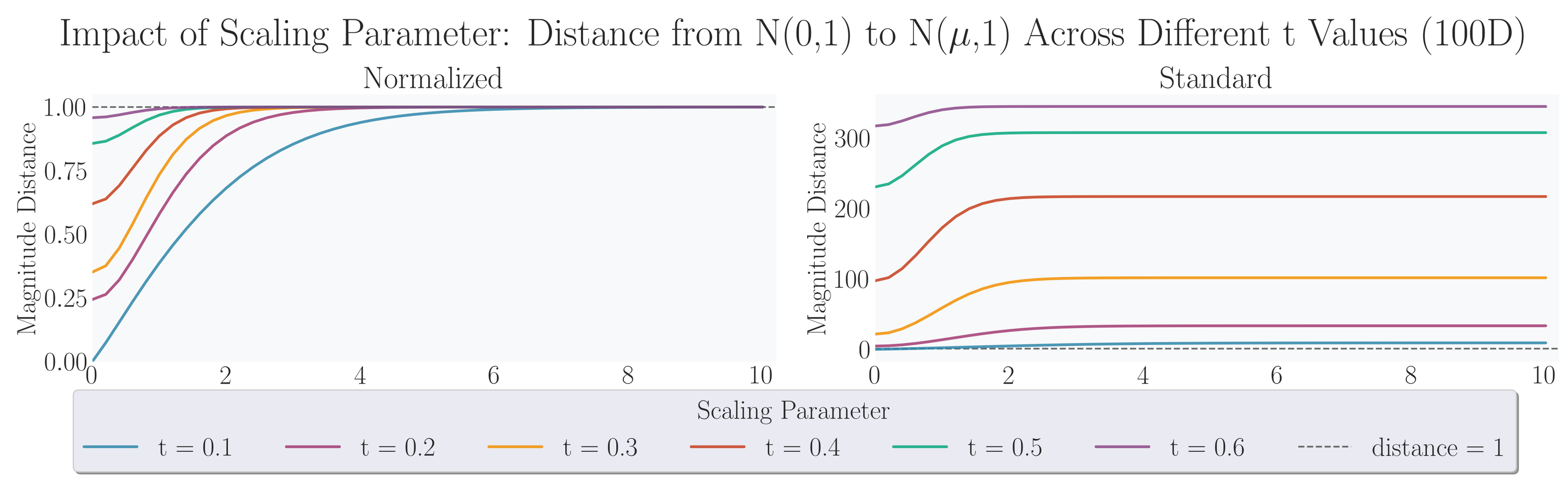
Definition

For two finite sets $X, Y \subset \mathbb{R}^D$, the magnitude distance with scale parameter $t \in \mathbb{R}_+$ is defined as

$$d^t_{Mag}(X, Y) = 2Mag_{X \cup Y}(t) - Mag_X(t) - Mag_Y(t),$$

and the normalized magnitude distance is defined as

$$\tilde{d}^t_{Mag}(X, Y) = \frac{d^t_{Mag}(X, Y)}{Mag_{X \cup Y}(t)}.$$



Scaling Parameter t

We show that the magnitude distance inherits similar properties of the magnitude function stated in [Proposition 2.2.6, Leinster *et al.*, 2013].

Theorem

For every finite metric sets X and Y, the magnitude distance $d^t_{Mag}(X)$:

- Converges to 0 as $t \to 0$.
- Converges to the cardinality of $X \Delta Y$ as $t \to \infty$.
- For $t \gg 0$, the magnitude distance $d^t_{Mag}(X)$ is increasing with respect to t.

The lower semicontinuity with respect to the Gromov-Hausdorff distance of the magnitude function on finite subsets of Euclidean space ensures that the magnitude distance is also lower semicontinuous.

➡ For every two finite sets $X, Y \in \mathbb{R}^D$, there exists a sufficiently small value of t for which the magnitude distance is meaningful.

Result

$d^t_{Mag}(X)$ remains discriminative even in high-dimensional settings.

In contrast, classical distances are known to suffer from the **curse of dimensionality.**

Properties

Metric Axioms

Theorem

Magnitude distance satisfies the following properties for $X, Y \subset \mathbb{R}^D$ and $t > 0$:

- **Symmetry:** $d^t_{\text{Mag}}(X, Y) = d^t_{\text{Mag}}(Y, X)$ by definition.
- **Non-negativity:** For any $t > 0$, we have $d^t_{\text{Mag}}(X, Y) \geq 0$.
- **Identity of indiscernibles:** $d^t_{Mag}(X, Y) = 0 \iff X \underset{Mag(t)}{=} Y$.
- **Triangle inequality:** $d^t_{\text{Mag}}(X, Y)$ does not satisfy the triangle inequality in $\mathbb{R}^D$ for $D > 1$.

Outlier Robustness

Theorem

Let $X, Y \subset \mathbb{R}^D$ be finite sets with nonnegative weighting vectors of X, Y, and $X \cup Y$. Then, we have:

$$0 \leq d^t_{Mag}(X, Y) \leq 2(|X \cup Y|).$$

where $|X|$ and $|Y|$ denote the number of points in X and Y respectively.

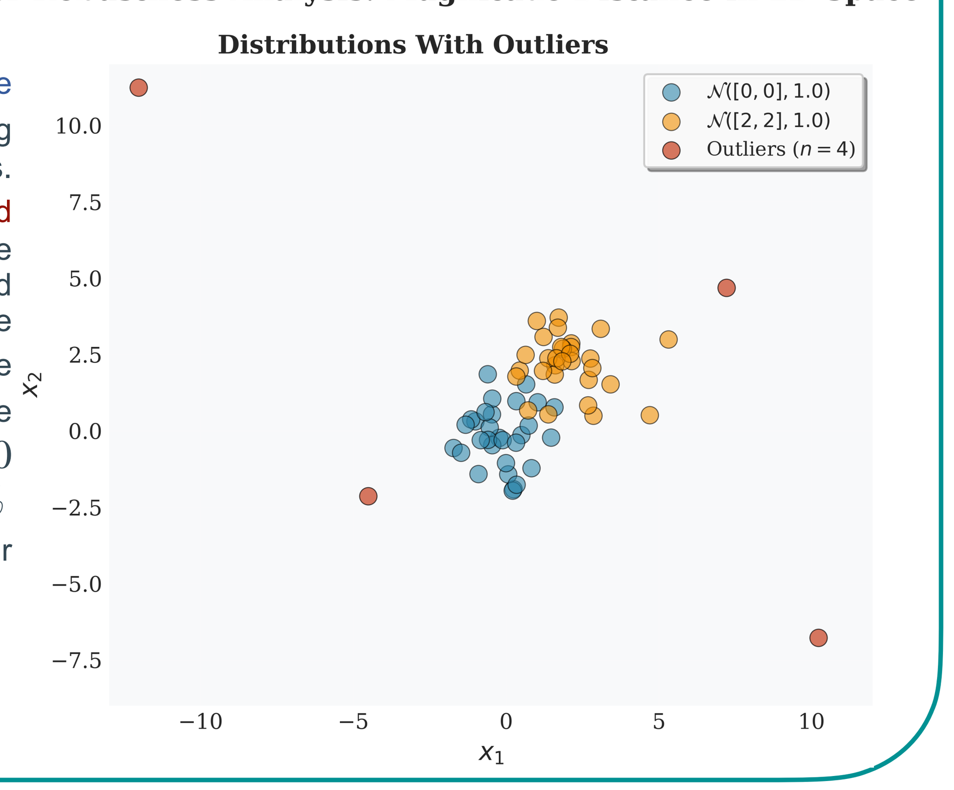
Nonnegative weighting vectors are guaranteed in all subsets of metric spaces when scaled up sufficiently, i.e., $t \gg 0$, and also $\mathbb{R}$ which this bound exists for any scaling parameter.

Result

$d^t_{Mag}(X)$'s sensitivity to adding or adjusting samples is also bounded.

Outlier Robustness Analysis: Magnitude Distance in 2D Space

Caption: We generate two datasets, B (blue points) and Y (yellow points), by sampling from normal distributions with different means. We also generate a third set of points, Y' (red points), with much higher dispersion. We consider the magnitude distance and Wasserstein distance for two cases: the distance between B and Y, and the distance between B and $Y^* = Y \cup Y'$. The relative change in magnitude distance with $t = 20$ and $t = 5$ are $6.85\,\%$ and $10.29\,\%$ respectively, compared to $17.61\,\%$ for Wasserstein distance.



References

Tom Leinster (2013). "The magnitude of metric spaces." In: Documenta Mathematica, 18:857–905, 2013
Tom Leinster (2021). "Entropy and diversity: the axiomatic approach." In: Cambridge University Press
Rayna Andreeva (2025) "Approximating metric magnitude of point sets." In: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 15374–15381, 2025.

PriGM@EurIPS2025

sahel.torkamani@ed.ac.uk