A Metric Notion of Dimension and Its Applications to Learning

Robert Krauthgamer (Weizmann Institute)

Based on joint works with Lee-Ad Gottlieb, James Lee, and Aryeh Kontorovich
Finite metric spaces

\((X,d)\) is a metric space if

- \(X\) = set of points
- \(d\) = distance function
  - Nonnegative
  - Symmetric
  - Triangle inequality

- Many “ad-hoc” metric data sets (not a vector space)
  - But are “close” to low-dimensional Euclidean metrics.

- Many Euclidean data sets with high embedding dimension.
  - But have “intrinsic” low-dimensional structure, e.g. a low-dimensional manifold.

- Goal: Capture their “intrinsic dimension”.
Intrinsic Dimension

- Can we measure the “complexity” of a metric using a notion of “dimension”?
  - Abstract—should depend only on the distances
  - Analogy—should generalize vector spaces ($\mathbb{R}^m$)

- We borrow a notion from Analysis
  - Show its advantages over previous suggestions
  - And that it controls the complexity of various algorithms!
A Metric Notion of Dimension

- Definition: $B(x,r) = \text{all points of } X \text{ within distance } r \text{ from } x$.

- The doubling dimension of $(X,d)$, denoted $\text{dim}(X)$, is the minimum $k > 0$ s.t. every ball can be covered by $2^k$ balls of half the radius.

  - Defined by [Gupta-K.-Lee’03], inspired by [Assouad’83, Clarkson’97].
  - Call a metric doubling if $\text{dim}(X) = O(1)$.
  - Captures every norm on $\mathbb{U}^k$.

- Robust to:
  - taking subsets,
  - union of sets,
  - small distortion in distances, etc.

- Unlike earlier suggestions based on $|B(x,r)|$ [Plaxton-Richa-Rajaraman’97, Karger-Ruhl’02, Faloutsos-Kamel’94, K.-Lee’03, …]

Here $2^k \approx 7$. A metric notion of dimension and its applications to learning
Example: Earthmover Metric

- The earthmover distance between $S, T \subseteq [0,1]^2$ with $|S| = |T|$, is:

$$\text{EMD}(S, T) = \min_{\pi : S \rightarrow T} \left\{ \frac{1}{k} \sum_{s \in S} \|s - \pi(s)\|_2 \right\}$$

where the minimum is over all one-to-one mappings $\pi : S \rightarrow T$.

- Has several uses e.g. in computer vision applications

- Lemma: This earthmover metric for sets of size $k$ has doubling dimension $\lesssim O(k \log k)$.

- Proof sketch:
  - Fix an $r/2$-grid in the plane, and “approximate” a set $S$ by snapping it to grid point
  - The sets $S$ near a fixed $T$, i.e. $\text{EMD}(S, T) \lesssim r$, can be “approximated” by one of only $k^{O(k)}$ fixed sets
Applications of Doubling Metrics

- **Approximate Nearest Neighbor Search (NNS)** [Clarkson’97, K.-Lee’04,…, Cole-Gottlieb’06, Indyk-Naor’06,…]

- **Dimension reduction** [Bartal-Recht-Sculman’07, Gottlieb-K.’09]

- **Embeddings** [Gupta-K.-Lee’03, K.-Lee-Mendel-Naor’04, Abraham-Bartal-Neiman’08]

- **Networking and distributed systems:**
  - Spanners [Talwar’04,…, Gottlieb-Roditty’08]
  - Compact Routing [Chan-Gupta-Maggs-Zhou’05,…]
  - Network Triangulation [Kleinberg-Slivkins-Wexler’04,…]
  - Distance oracles [HarPeled-Mendel’06]

- **Classification** [Bshouty-Li-Long’09, Gottlieb-Kontorovich-K.’10]
Near Neighbor Search (NNS)

- **Problem statement:**
  - Preprocess $n$ data points in a metric space $X$, so that
  - Given a query point $q$, can quickly find the closest point to $q$ among $X$, i.e. compute $a \in X$ such that $d(q,a) = d(q,X)$.

- **Naive solution:**
  - No preprocessing, query time $O(n)$

- **Ultimate goal (holy grail):**
  - Preprocessing time: about $O(n)$
  - Query time: $O(\log n)$
  - Achievable on the real line
NNS in Doubling Metrics

- **A simple \((1+\varepsilon)\)-NNS scheme** \([K.-Lee’04a]\):
  - Query time: \((1/\varepsilon)^{O(\text{dim}(X))} \cdot \log \Phi\). \([\Phi = d_{\text{max}}/d_{\text{min}} \text{ is spread}]\)
  - Preprocessing: \(n \cdot 2^{O(\text{dim}(X))}\).
  - Insertion / deletion time: \(2^{O(\text{dim}(X))} \cdot \log \Phi \cdot \log \log \Phi\).

- **Outperforms previous schemes** \([Plaxton-Richa-Rajaraman’98, Clarkson’99, Karger-Ruhl’02]\):
  - Simpler, wider applicability, deterministic, bound not needed
  - Nearly matches the Euclidean case \([Arya et al.’94]\)
  - Explains empirical successes—it’s just easy…

- **Subsequent enhancements**:
  - Optimal storage \(O(n)\) \([Beygelzimer-Kakade-Langford]\)
    - Also implemented and obtained very good empirical results
  - No dependence on \(\Phi\) \([K.-Lee’04b, Mendel-HarPeled’05, Cole-Gottlieb’06]\)
  - Faster scheme for doubling Euclidean metrics \([Indyk-Naor]\)
Nets

- **Motivation:** Approximate the metric at one scale \( r > 0 \).
  - Provide a spatial “sample” of the metric space
  - E.g., grids in \( \mathbb{U}^2 \).

- **Definition:** \( Y \) \( \rightarrow \) \( X \) is called an \( r \)-net if it is an \( r \)-separated subset that is maximal, i.e.,
  1. For all \( y_1, y_2 \in Y \), \( d(y_1, y_2) \geq r \) [packing]
  2. For all \( x \in X \setminus Y \), \( d(x, Y) < r \) [covering]
Navigating nets

NNS scheme (**simplest variant**):

- **Preprocessing**
  - Compute a $2^i$-net for all $i$.  
  - Add “local links”.

- **Query**
  - Iteratively go to finer nets  
  - Navigating towards query point.

From a $2^i$-net point to **nearby** $2^{i-1}$-net points,  
# local links $\lesssim 2^\Theta(dim(X))$. 

A metric notion of dimension and its applications to learning
The JL Lemma

**Theorem [Johnson-Lindenstrauss, 1984]:**
For every n-point set $X \subseteq \mathbb{R}^m$ and $0 < \varepsilon < 1$, there is a map $\Psi: X \rightarrow \mathbb{R}^k$, for $k = O(\varepsilon^{-2} \log n)$, that preserves all distances within $1 + \varepsilon$:

$$ ||x-y||_2 < ||\Psi(x)-\Psi(y)||_2 < (1+\varepsilon) ||x-y||_2, \quad ; x, y \in X. $$

- Can be realized by a simple linear transformation
  - A random $k \times d$ matrix works – entries from $\{-1,0,1\}$ [Achlioptas’01] or Gaussian [Gupta-Dasgupta’98, Indyk-Motwani’98]

- Many applications (e.g. computational geometry)

- Can we do better?
A Stronger Version of JL?

Recall [Johnson-Lindenstrauss, 1984]:

Every \( n \)-point set \( X \rightarrow l_2 \) and \( 0 < \varepsilon < 1 \), has a linear embedding \( \Psi: X \rightarrow l_2^k \), for \( k = O(\varepsilon^{-2} \log n) \), such that for all \( x, y \in X \),

\[
\|x - y\|_2 < \|\Psi(x) - \Psi(y)\|_2 < (1 + \varepsilon) \|x - y\|_2.
\]

- A matching lower bound of [Alon’03]:
  - \( X = \) uniform metric, then \( \dim(X) = \log n \), \( k = \Omega(\varepsilon^{-2} \log n) \)

- Open: JL-like embedding into dimension \( k = k(\varepsilon, \dim X) \)?
  - Even constant distortion would be interesting [Lang-Plaut’01, Gupta-K.-Lee’03]:
  - Cannot be attained by linear transformations [Indyk-Naor’06]

We present two partial resolutions, using \( \tilde{O}(\dim X^2) \) dimensions:

1. Distortion 1+\( \varepsilon \) for a single scale, i.e. pairs where \( \|x - y\| \geq \delta r, r \).
2. Global embedding of the snowflake metric, \( \|x - y\|^{\frac{1}{2}} \).
2’. Conjecture correct whenever \( \|x - y\|^2 \) is Euclidean (e.g. for every ultrametric).
I. Embedding for a Single Scale

- **Theorem 1 [Gottlieb-K.]:** For every finite subset $X \subseteq l_2$, and all $0 < \delta < 1$, $r > 0$, there is embedding $f: X \rightarrow l_2^k$ for $k = \tilde{O}(\log(1/\delta)(\text{dim } X)^2)$, satisfying
  1. Lipschitz: $||f(x) - f(y)|| \leq ||x - y||$ for all $x, y \in X$
  2. Bi-Lipschitz at scale $r$: $||f(x) - f(y)|| \geq \Omega(||x - y||)$ whenever $||x - y|| \in [\delta r, r]$
  3. Boundedness: $||f(x)|| \leq r$ for all $x \in X$

- **Compared to open question:**
  - Bi-Lipschitz only at one scale (weaker)
  - But achieves distortion = absolute constant (stronger)

- **Improved version:** $1 + \epsilon$ distortion whenever $||x - y|| \in [\delta \epsilon r, \epsilon r]$

- **Divide and conquer approach:**
  - Net extraction
  - Padded Decomposition
  - Gaussian Transform (kernel)
  - JL (locally)
  - Glue partitions
  - Extension theorem
II. Snowflake Embedding

- **Theorem 2 [Gottlieb-K.]:** For every $0 < \varepsilon < 1$ and finite subset $X \subseteq \mathbb{R}^2$ there is an embedding $F : X \rightarrow \ell_2^k$ of the snowflake metric $||x-y||^{1/2}$ achieving dimension $k = \tilde{O}(\varepsilon^{-4}(\dim X)^2)$ and distortion $1 + \varepsilon$, i.e.

  $$1 \leq \frac{||F(x) - F(y)||}{||x - y||^{1/2}} \leq 1 + \varepsilon, \quad \forall x, y \in X.$$ 

- **Compared to open question:**
  - We embed the snowflake metric (weaker)
  - But achieve distortion $1 + \varepsilon$ (stronger)

- **We generalize [Kahane’81, Talagrand’92] who embed Wilson’s helix (real line w/distances $|x-y|^{1/2}$)**
Distance-Based Classification

- **How to classify data presented as points in a metric space?**
  - No inner-product structure…

- **Framework for large-margin classification [von Luxburg-Bousquet’04]:**
  - Hypotheses (classifiers) are Lipschitz functions \( f: X \rightarrow \mathbb{U} \).
  - Classification reduces to finding such \( f \) consistent with labeled data (which is a classic problem in Analysis, known as Lipschitz extension)
  - They establish generalization bounds
  - Evaluating the classifier \( f(.) \) is reduced to 1-NNS
    - Exact NNS
    - Assuming zero training error

- **Left open:**
  - What about approximate NNS? Exact is hard…
  - What about training error?
Efficient Classification of Metric Data

- **[Gottlieb-Kontorovich-K.’10]:** Data of low doubling dimensions admits accurate and computationally-efficient classification
  - First relation between classification efficiency and data's dimension
    - Contrast to [Bshouty-Li-Long’09]

I. Choose a classifier quickly (deal w/training error)
- Need to find the errors and optimize bias-variance tradeoff
- Key step: Given target # Lipschitz constant, minimize # of outliers (and find them)

II. Evaluate the classifier using approximate NNS
- For which efficient algorithms are known
- Key idea: approx. NNS is “indetermined” but we only need to evaluate sign(f)
The Bigger Picture

Summary

- Algorithms for general (low-dimensional) metric spaces

Future

- Other contexts:
  - Different/weaker assumptions, e.g. similarity
  - Deal with graphs (social networks? Chemicals?)
- Rigorous analysis of heuristics
  - Explain or predict empirical behavior of algorithms.
- Connections to other fields
  - Databases, networking, machine learning, Fourier analysis, communication complexity, information theory