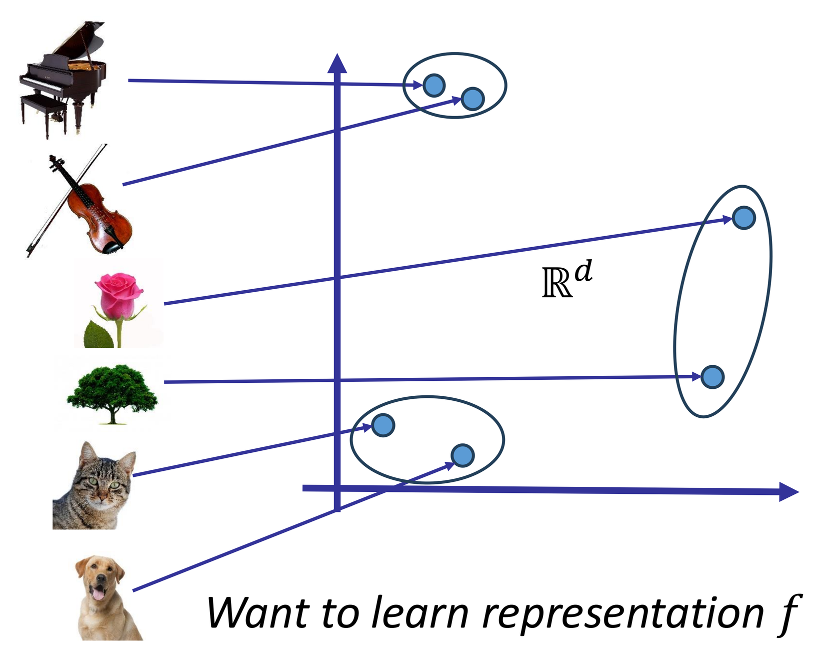
Embedding Dimension of Contrastive Learning and k-Nearest Neighbors

Dmitrii Avdiukhin[1], Vaggos Chatziafratis, Orr Fischer[3], Grigory Yaroslavtsev[4]

[1]Northwestern University, [2]UC Santa Cruz, [3]Weizmann Institute of Science, [4]George Mason University

REPRESENTATION LEARNING



Input type
- Image
- Text
- Sound
- Video
- Graph
- etc.

Neural Networks
- Transformers
- CNNs
- LSTM
- Graph NN
- etc.

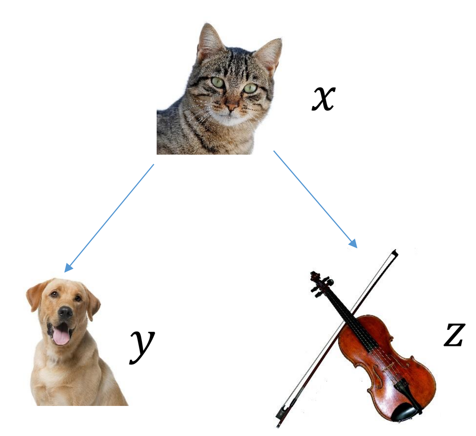
CONTRASTIVE LEARNING

Labeled Input

Triplet (x, y^+, z^-)
- "Anchor" x
- "Positive" element y
- "Negative" element z

Interpretation

"x is more similar to y than to z"
$\|f(x) - f(y)\| < \|f(x) - f(z)\|$



EMBEDDING DIMENSION

Question

Assume we have m contrastive constraints

What is the minimum dimension d required to satisfy all constraints?

E.g. a design choice when designing a neural network

k-NEAREST NEIGHBORS

For each point $x \in V$, know k points $y_1, \dots, y_k$ closest to x
1) $\|f(x) - f(y_1)\| < \|f(x) - f(y_2)\| < \dots < \|f(x) - f(y_k)\|$
2) $\|f(x) - f(y_k)\| < \|f(x) - f(z)\|$ for all other z

What is the minimum dimension d
required to preserve k-NN for all points?

Previous work

[Alon et al.`24]: for a fixed dimension d, the VC-dimension of contrastive learning is $\tilde{O}(nd)$
[Chatziafratis, Indyk`23]:
- $d = \Theta(n)$ is necessary and sufficient to preserve all constraints
- $d = \Theta(k)$ to preserve orderings between k-NN, but not necessarily k-NN

OUR RESULT

CONTRASTIVE LEARNING

Consider a set of m non-contradictory constraints. There exists embedding into ℓ_p space satisfying all constraints if d is chosen as follows.
- $\boldsymbol{d \in \Theta(\sqrt{m})}$ **for** $\boldsymbol{p = 2}$
  - $O(\sqrt{m})$ always suffices, and some set of constraints require $\Omega(\sqrt{m})$
- $d \in O(m^{2/3}) \cap \Omega(\sqrt{m})$ for $p = \infty$
- $d \in O(m) \cap \Omega(\sqrt{m})$ for positive integer p

k-NN

For a positive integer p, there exists embedding into ℓ_p space of dimension $d = \mathrm{poly}(k) \cdot \mathrm{polylog}(n)$ preserving the k-nearest neighbors of each point.
- No polynomial dependence on n
  - Very surprising since k-NN information encodes $\Theta(n^2)$ constraints
- No dependence on p

MAIN TECHNIQUES

Intuition

For $m = \Theta(n^2)$, have $d = \Theta(n)$ [Chatziafratis, Indyk`23]
- Gives the $\Omega(\sqrt{m})$ lower bound

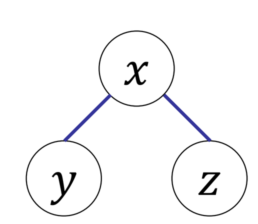
If there are $O(n)$ constraints:
- If the constraints are spread over all n points, $d = \Theta(1)$
- If these constraints are concentrated on $\sqrt{n}$ points, $d = \Theta(\sqrt{n})$.

The density seems to matter

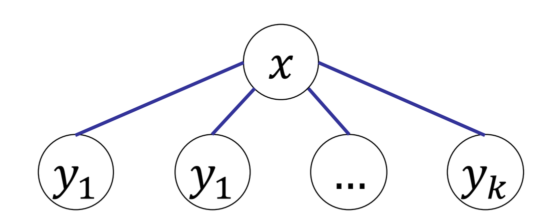
Constraint graph

Create edges for all constraints (contrastive learning) or points (k-NN)

(x, y^+, z^-)

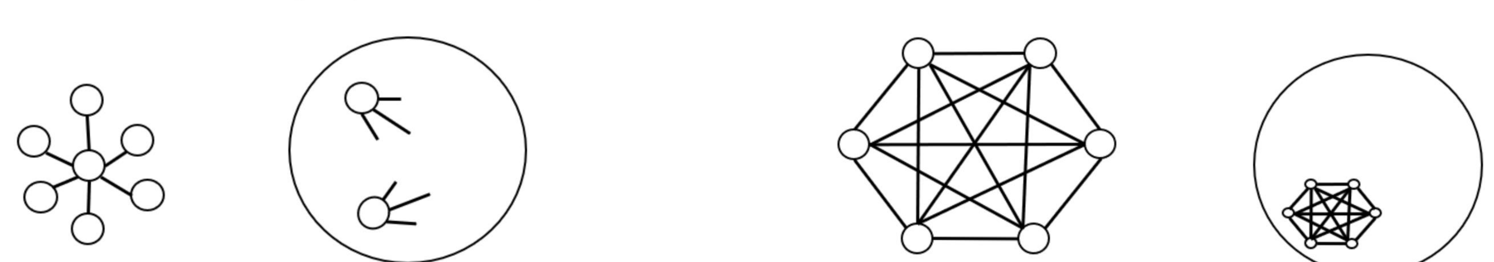


x with its nearest neighbors $y_1, \dots, y_k$



Arboricity

- **Definition**: the minimum number of forests to cover the graph
- Measure of graph density



Bounds on arboricity

- Clique has arboricity $\lceil \sqrt{m/2} \rceil$
- The $\lceil \sqrt{m/2} \rceil$ bound is tight

Why is arboricity useful?

Assume that the graph has arboricity r
- ($2r$-degeneracy) There is an order of vertices so that for any $x \in V$, the number of neighbors of x preceding x is at most $2r$
- The graph is $2r$-vertex colorable

UPPER BOUND FOR CONTRASTIVE LEARNING IN ℓ_2 ($d = \Theta(\sqrt{m})$)

Outline

- Since constraints are non-contradictory, there exists a metric satisfying these constraints
- Let $w(x, y) \in \{1, \dots, 2m\}$ be the rank of edge (x, y) in the metric, starting from the heaviest edge
- We'll construct embedding f:

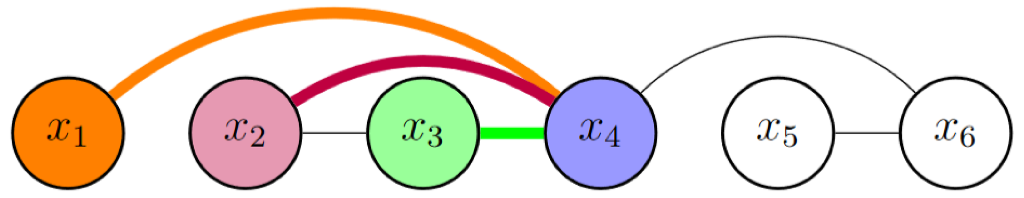
$$\|f(x) - f(y)\| - \|f(x) - f(z)\| = w(x,z) - w(x,y)$$

$$\|f(x) - f(y)\|^2 = \underbrace{\|f(x)\|^2}_{\text{same}} + \underbrace{\|f(y)\|^2}_{\text{same}} - 2\underbrace{\langle f(x), f(y)\rangle}_{= w(x,y)}$$

$\hat{x}$	$\mathring{x}$
$\hat{x}$ sets inner products according to edge ranks	$\mathring{x}$ makes norms of all vectors equal

CONSTRUCTION OF $\hat{x}$

$\langle \hat{x}_1, \hat{x}_4 \rangle = w(\hat{x}_1, \hat{x}_4)$
$\langle \hat{x}_2, \hat{x}_4 \rangle = w(\hat{x}_2, \hat{x}_4)$
$\langle \hat{x}_3, \hat{x}_4 \rangle = w(\hat{x}_3, \hat{x}_4)$



Dimension?

- System of at most $2r$ linear equations over $\hat{x}_i$
- Add small noise to every vector
- If dimension is $2r$ – there is always a solution

CONSTRUCTION OF $\mathring{x}$

Equalize vector norms

- Every vertex chooses one coordinate
- For each vertex, set its coordinate so that $\|f(x_1)\| = \|f(x_2)\| = \cdots \|f(x_n)\|$

How to preserve inner products?

- A vertex chooses a coordinate not used by any of preceding neighbors
  - Then, $\langle f(x), f(y)\rangle = \langle \hat{x}, \hat{y}\rangle + \underbrace{\langle \mathring{x}, \mathring{y}\rangle}_{=0}$
- There are at most $2r$ preceding neighbors