Low-Dimensional Lipschitz Embeddings Invariant to Permutations

Radu Balan

Department of Mathematics and Norbert Wiener Center for Harmonic Analysis and Applications
University of Maryland, College Park, MD

March 17, 2022
One World Mathematics of INformation, Data, and Signals (1-W MINDS) Seminar

Acknowledgments

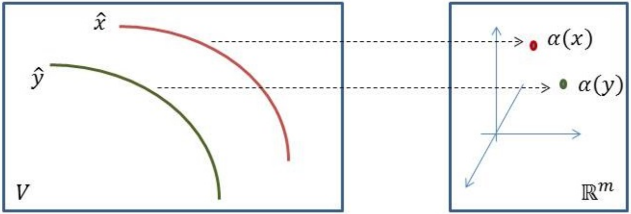
This material is based upon work partially supported by the National Science Foundation under grant no. DMS-2108900 and Simons Foundation. "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

Joint work with:
Naveed Haghani (UMD,APL-JHU)
Maneesh Singh (Verisk)
arXiv preprint: 2203.07546 [math.FA] , [cs.LG]

High-Level View

In this talk, we discuss Euclidean embeddings of metric spaces induced by representations of permutation (sub)groups $\mathcal{S}_n$ on linear spaces V.

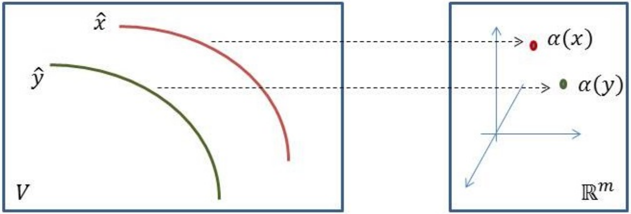
Problem: Construct bi-Lipschitz embeddings of the metric space $\hat{V} = V/\sim$ of orbits, $\alpha : \hat{V} \to \mathbb{R}^m$, where $d(\hat{x}, \hat{y}) = \min_{u\in\hat{x}, v\in\hat{y}} \|u - v\|_V$.



High-Level View

In this talk, we discuss Euclidean embeddings of metric spaces induced by representations of permutation (sub)groups $\mathcal{S}_n$ on linear spaces V.

Problem: Construct bi-Lipschitz embeddings of the metric space $\hat{V} = V/\sim$ of orbits, $\alpha : \hat{V} \to \mathbb{R}^m$, where $d(\hat{x},\hat{y}) = \min_{u\in\hat{x},v\in\hat{y}} \|u-v\|_V$.



Today we focus on the case $V = \mathbb{R}^{n\times d}$, $X \sim Y \Leftrightarrow Y = PX$ for some $P \in \mathcal{S}_n$.

Table of Contents:

Table of Contents

Graph Learning Problems

Given a data graph (e.g., social network, transportation network, citation network, chemical network, protein network, biological networks):

- Graph adjacency or weight matrix, $A \in \mathbb{R}^{n\times n}$;
- Data matrix, $X \in \mathbb{R}^{n\times r}$, where each row corresponds to a feature vector per node.

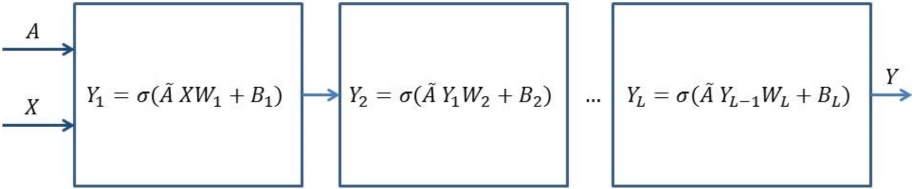
Contruct a map $f : (A, X) \to f(A, X)$ that performs:

1. classification: $f(A, X) \in \{1, 2, \cdots, c\}$
2. regression/prediction: $f(A, X) \in \mathbb{R}$.

Key observation: The outcome should be invariant to vertex permutation: $f(PAP^T, PX) = f(A, X)$, for every $P \in \mathcal{S}_n$.

Graph Convolution Networks (GCN), Graph Neural Networks (GNN)

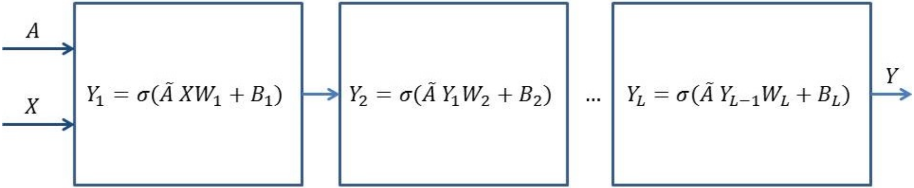
General architecture of a GCN/GNN



GCN (Kipf and Welling ('16)) choses $\tilde{A} = I + A$; GNN (Scarselli et.al. ('08), Bronstein et.al. ('16)) choses $\tilde{A} = p_l(A)$, a polynomial in adjacency matrix. L-layer GNN has parameters $(p_1, W_1, B_1, \cdots, p_L, W_L, B_L)$.

Graph Convolution Networks (GCN), Graph Neural Networks (GNN)

General architecture of a GCN/GNN

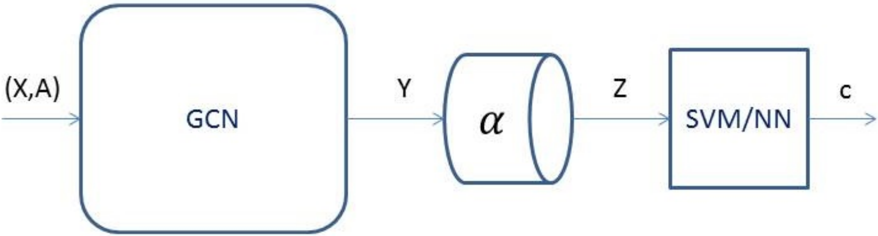


GCN (Kipf and Welling ('16)) choses $\tilde{A} = I + A$; GNN (Scarselli et.al. ('08), Bronstein et.al. ('16)) choses $\tilde{A} = p_l(A)$, a polynomial in adjacency matrix. L-layer GNN has parameters $(p_1, W_1, B_1, \cdots, p_L, W_L, B_L)$.

Note the *covariance (or, equivariance) property*: for any $P \in O(n)$ (including $\mathcal{S}_n$), if $(A, X) \mapsto (PAP^T, PX)$ and $B_i \mapsto PB_i$ then $Y \mapsto PY$.

Deep Learning with GCN/GNN

The approach for the two learning tasks (classification or regression) is based on the following scheme (see also Maron et.al. ('19)):



where α is a permutation invariant map (embedding), and SVM/NN is a single-layer or a deep neural network (Support Vector Machine or a Fully Connected Neural Network) trained on invariant representations.

The purpose of this talk is to analyze the α component.

Table of Contents

The metric space $\hat{V}$ when $V = \mathbb{R}^{n \times d}$

Recall the equivalence relation $\sim$ on $V = \mathbb{R}^{n \times d}$ induced by the group of permutation matrices $\mathcal{S}_n$ acting on V by left multiplication: for any $X, X' \in \mathbb{R}^{n \times d}$,

$$X \sim X' \quad \Leftrightarrow \quad X' = PX \ , \ \text{for some } P \in \mathcal{S}_n$$

Let $\widehat{\mathbb{R}^{n \times d}} = \mathbb{R}^{n \times d} / \sim$ be the quotient space endowed with the natural distance induced by Frobenius norm $\| \cdot \|_F$

$$d(\hat{X}_1, \hat{X}_2) = \min_{P \in S_n} \|X_1 - PX_2\|_F \quad , \quad \hat{X}_1, \hat{X}_2 \in \widehat{\mathbb{R}^{n \times d}}.$$

The metric space $\hat{V}$ when $V = \mathbb{R}^{n \times d}$

Recall the equivalence relation $\sim$ on $V = \mathbb{R}^{n \times d}$ induced by the group of permutation matrices $\mathcal{S}_n$ acting on V by left multiplication: for any $X, X' \in \mathbb{R}^{n \times d}$,

$$X \sim X' \quad \Leftrightarrow \quad X' = PX \;, \text{ for some } P \in \mathcal{S}_n$$

Let $\widehat{\mathbb{R}^{n \times d}} = \mathbb{R}^{n \times d} / \sim$ be the quotient space endowed with the natural distance induced by Frobenius norm $\| \cdot \|_F$

$$d(\hat{X}_1, \hat{X}_2) = \min_{P \in \mathcal{S}_n} \|X_1 - PX_2\|_F \quad , \quad \hat{X}_1, \hat{X}_2 \in \widehat{\mathbb{R}^{n \times d}}.$$

The computation of the minimum distance is performed by solving the Linear Assignment Problem (LAP) whose convex relaxation is exact:

$$\max_{P \in \mathcal{S}_n} trace(PX_2X_1^T) = \max_{W \in DS(n)} trace(WX_2X_1^T)$$

where $DS(n) = \{W \in [0,1]^{n \times n} : W1 = 1, W^T 1 = 1\}$ is the convex set of doubly stochastic matrices.

The embedding problem

Problem: Construct a bi-Lipschitz embedding $\hat{\alpha} : \widehat{\mathbb{R}^{n \times d}} \to \mathbb{R}^m$, i.e., an integer $m = m(n, d)$, a map $\alpha : \mathbb{R}^{n \times d} \to \mathbb{R}^m$ with constants $0 < a \leq b < \infty$ so that for any $X, X' \in \mathbb{R}^{n \times d}$,

1. If $X \sim X'$ then $\alpha(X) = \alpha(X')$.
2. If $\alpha(X) = \alpha(X')$ then $X \sim X'$.
3. $a \cdot d(\hat{X}, \hat{X}') \leq \|\alpha(X) - \alpha(X')\|_2 \leq b \cdot d(\hat{X}, \hat{X}')$.

where $d(\hat{X}, \hat{X}') = \min_{P \in \mathcal{S}_n} \|X - PX'\|_F$.

A Universal Embedding

Consider the map

$$\mu : \widehat{\mathbb{R}^{n\times d}} \rightarrow \mathcal{P}(\mathbb{R}^d) \quad , \quad \mu(X)(x) = \frac{1}{n}\sum_{k=1}^{n} \delta(x - x_k)$$

where $\mathcal{P}(\mathbb{R}^d)$ denotes the convex set of probability measures over $\mathbb{R}^d$, and δ denotes the Dirac measure. x_k is the k^{th} row of X.
Clearly $\mu(X') = \mu(X)$ iff $X' = PX$ for some $P \in \mathcal{S}_n$.
The Wasserstein-2 distance is equivalent to the natural metric:

$$W_2(\mu(X), \mu(Y))^2 := \inf_{q \in J(\mu(X),\mu(Y))} \mathbb{E}_q[\|x - y\|_2^2] = \min_{P \in \mathcal{S}_n} \|Y - PX\|^2$$

By Kantorovich-Rubinstein theorem, the Wasserstein-1 distance (the Earth moving distance) extends to a norm on the space of signed Borel measures.
Main drawback: $\mathcal{P}(\mathbb{R}^d)$ is infinite dimensional!

Finite Dimensional Embeddings

Idea: "Project" the measure onto a finite dimensional space. This is accomplished by *kernel methods*:

Fix a family of functions $f_1, \cdots, f_m$ and consider:

$$\mu(X) \mapsto \int_{\mathbb{R}^d} f_j(x) d\mu(X) = \frac{1}{n} \sum_{k=1}^{n} f_j(x_k) \quad , \quad j \in [m]$$

Finite Dimensional Embeddings

Idea: "Project" the measure onto a finite dimensional space. This is accomplished by *kernel methods*:

Fix a family of functions $f_1, \cdots, f_m$ and consider:

$$\mu(X) \mapsto \int_{\mathbb{R}^d} f_j(x) d\mu(X) = \frac{1}{n}\sum_{k=1}^{n} f_j(x_k) \quad , \quad j \in [m]$$

Possible choices:

1. Polynomial embeddings: $\mathbb{R}[X]^{\mathcal{S}_n}$, ring of invariant polynomials; [Lipman&al.],[Peyré&al.],[Sanay&al.],[Kemper book] ...
2. Gaussian kernels: $f_j(x) = exp(-\|x - a_j\|^2/\sigma_j^2)$; [Gilmer&al.],[Zaheer&al.], [Vinyals&al.],...
3. Fourier kernels (cmplx embd): $f_j(x) = exp(2\pi i\langle x, \omega_j\rangle)$; related to Prony method; [Li&Liao] for bi-Lipschitz estimates.

Main drawback: No global bi-Lipschitz embeddings [Cahill&al.]. Ok on (some) compacts.

Table of Contents

The Max Pool approach

The idea is provided by the following observation.
Let $\downarrow: \mathbb{R}^n \to \mathbb{R}^n$ denote the *sorting map* $x \mapsto \downarrow x = \Pi x$, $\Pi \in \mathcal{S}_n$, so that

$$(\Pi x)_1 \geq (\Pi x)_2 \geq \cdots \geq (\Pi x)_n.$$

The Max Pool approach

The idea is provided by the following observation.
Let $\downarrow: \mathbb{R}^n \to \mathbb{R}^n$ denote the *sorting map* $x \mapsto \downarrow x = \Pi x$, $\Pi \in \mathcal{S}_n$, so that

$$(\Pi x)_1 \geq (\Pi x)_2 \geq \cdots \geq (\Pi x)_n.$$

Lemma

$\downarrow: \widehat{\mathbb{R}^n} \to \mathbb{R}^n$ *is an isometry (hence bi-Lipschitz):*

$$\| \downarrow (x) - \downarrow (y) \| = \min_{P\in\mathcal{S}_n} \|x - Py\| \ , \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$

Proof is based on the rearrangement inequality (see Wikipedia, or Hardy-Littlewood-Pólya "Inequalities" §10.2).

Our main goal is to extend this construction from $\mathbb{R}^n$ to $\mathbb{R}^{n\times d}$

The Encoder β_A

Notations

Recall the equivalence relation, for $X, Y \in \mathbb{R}^{n\times d}$,

$$X \sim Y \quad \Leftrightarrow \quad \exists \Pi \in \mathcal{S}_n \ , \ Y = \Pi X$$

that induces a quotient space $\widehat{\mathbb{R}^{n\times d}} = \mathbb{R}^{n\times d} / \sim$ and the natural distance

$$d : \widehat{\mathbb{R}^{n\times d}} \times \widehat{\mathbb{R}^{n\times d}} \to \mathbb{R} \quad , \quad d(X,Y) = \min_{\Pi \in \mathcal{S}_n} \|X - \Pi Y\|_F$$

In the following we look for an Euclidean embedding of the form

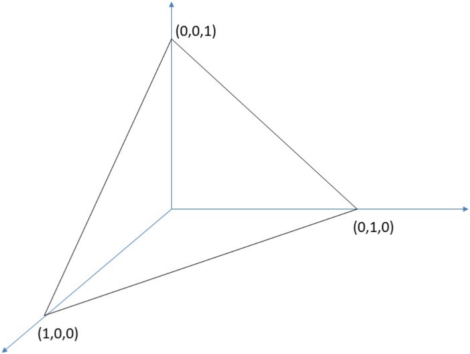
$$\beta_A : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times D} \quad , \quad \beta_A(X) = \downarrow (XA)$$

where $\downarrow(\cdot)$ sorts decreasingly each column of $\cdot$, independently.
The matrix $A \in \mathbb{R}^{d\times D}$ is called the *key* of encoder β_A.
The key is called *universal* if $\widehat{\beta_A} : \widehat{\mathbb{R}^{n\times d}} \to \mathbb{R}^{n\times D}$ is injective.

Intuition behind universality of keys

Consider the case
$n = 2$, $d = 3$

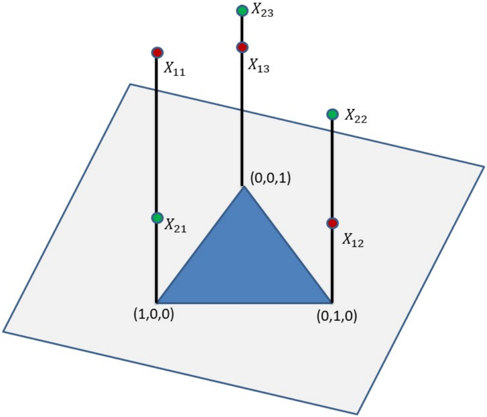
$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix}$$



Intuition behind universality of keys

Consider the case
$n = 2$, $d = 3$

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix}$$



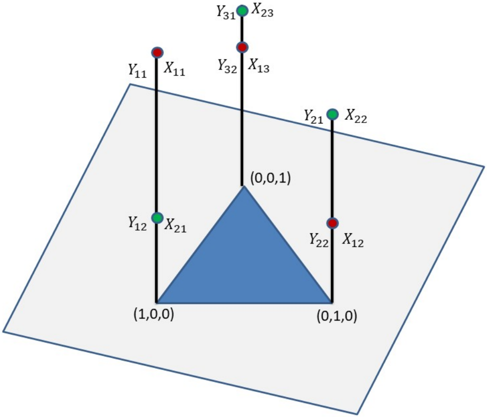
Intuition behind universality of keys

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix}$$

$$Y = \downarrow X$$

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \end{bmatrix}$$

Information lost!

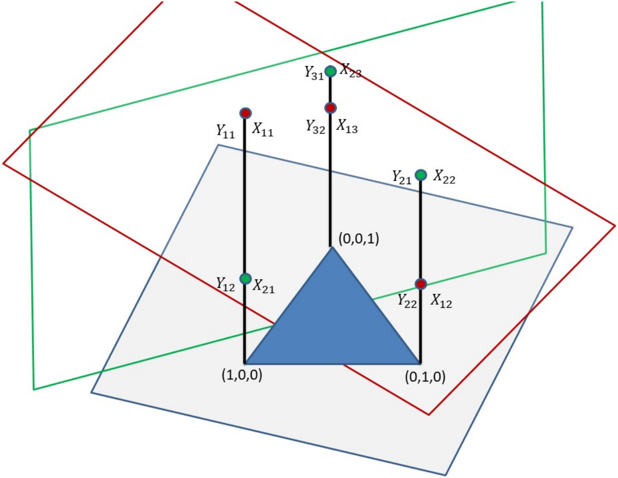


Intuition behind universality of keys

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix}$$

$$Y = \downarrow X$$

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \end{bmatrix}$$

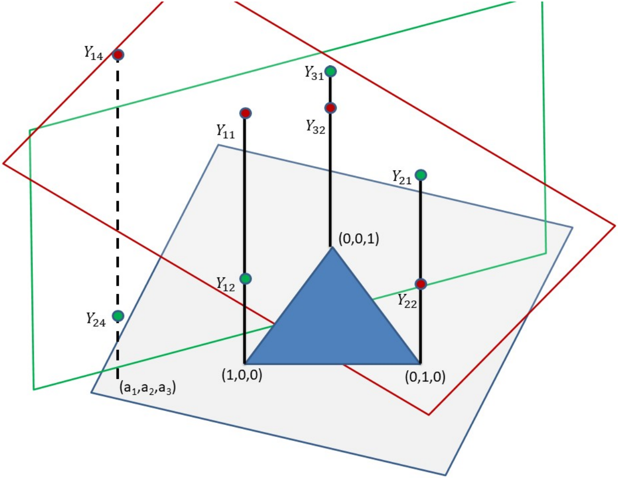


Intuition for this encoder

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix}$$

$$Y = \downarrow \begin{bmatrix} X & Xa \end{bmatrix}$$

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} & Y_{14} \\ Y_{21} & Y_{22} & Y_{23} & Y_{24} \end{bmatrix}$$



Three results (1)

Existence of Universal Keys

Theorem

Consider the metric space $(\widehat{\mathbb{R}^{n\times d}}, d)$. *Set* $D = 1 + (d-1)n!$ *and let* $A \in \mathbb{R}^{d\times D}$ *be a matrix whose columns form a full spark frame. Then the key* A *is universal and the induced map* $\hat{\beta}_A : \widehat{\mathbb{R}^{n\times d}} \to \mathbb{R}^{n\times D}$, $\hat{\beta}_A(\hat{X}) = \downarrow(XA)$ *is injective. Furthermore,* $\hat{\beta}_A$ *is bi-Lipschitz with constants* $a_0 = \min_{J\subset[D],|J|=d} s_d(A[J])$ *and* $b_0 = s_1(A)$, *where* $s_1(A)$ *denotes the largest singular value of* A, $A[J]$ *denotes the submatrix of* A *formed by columns indexed by* J, *and* $s_d(A[J])$ *denotes the* d^{th} *singular value (in this case, the smallest) of* $A[J]$. *Specifically, for any* $X, Y \in \mathbb{R}^{n\times d}$,

$$a_0 d(\hat{X}, \hat{Y}) \leq \|\beta_A(X) - \beta_A(Y)\| \leq b_0 d(\hat{X}, \hat{Y}) \tag{3.1}$$

where all norms are Frobenius norms.

Three results (2)

Bi-Lipschitz Property of Universal Keys

Theorem

Assume the key $A \in \mathbb{R}^{d \times D}$ *is universal, i.e., the induced map* $\hat{\beta}_A : \widehat{\mathbb{R}^{n \times d}} \to \mathbb{R}^{n \times D}$, $X \mapsto \beta_A(X) = \downarrow (XA)$ *is injective. Then* $\hat{\beta}_A$ *is bi-Lipschitz, that is, there are constants* $a_0 > 0$ *and* $b_0 > 0$ *so that for all* $X, Y \in \mathbb{R}^{n \times d}$,

$$a_0\, d(\hat{X}, \hat{Y}) \leq \|\beta_A(X) - \beta_A(Y)\| \leq b_0\, d(\hat{X}, \hat{Y}) \tag{3.2}$$

where all are Frobenius norms. Furthermore, an estimate for b_0 *is provided by the largest singular value of A,* $b_0 = s_1(A)$.

Three results (3)

Dimension Reduction

Theorem

Assume $A \in \mathbb{R}^{d\times D}$ is a universal key for $\widehat{\mathbb{R}^{n\times d}}$ with $D \geq 2d$. Then, for $m \geq 2nd$, a generic linear operator $B : \mathbb{R}^{n\times D} \to \mathbb{R}^m$ with respect to Zariski topology on $\mathbb{R}^{n\times D\times m}$, the map

$$\hat{\beta}_{A,B} : \widehat{\mathbb{R}^{n\times d}} \to \mathbb{R}^{2nd} \; , \; \hat{\beta}_{A,B}(\hat{X}) = B\left(\hat{\beta}_A(\hat{X})\right) \tag{3.3}$$

is bi-Lipschitz. In particular, almost every full-rank linear operator $B : \mathbb{R}^{n\times D} \to \mathbb{R}^{2nd}$ produces such a bi-Lipschitz map.

This result is compatible with a Whitney embedding theorem with the important caveat that the Whitney embedding result applies to smooth manifolds, whereas $\widehat{\mathbb{R}^{n\times d}}$ is not a manifold.

Highlights of proofs

Universal keys

The upper bound is imediate. For lower bound, fix $X, Y \in \mathbb{R}^{n\times d}$:

$$\|\beta_A(X) - \beta_A(Y)\|_2^2 = \sum_{k=1}^{D} \| \downarrow (Xa_k) - \downarrow (Ya_k)\|_2^2 = \sum_{k=1}^{D} \|P_k X a_k - Q_k Y a_k\|_2^2$$

$$\overset{\Pi_k := Q_k^T P_k}{=} \sum_{k=1}^{D} \|(\Pi_k X - Y) a_k\|_2^2$$

Highlights of proofs

Universal keys

The upper bound is imediate. For lower bound, fix $X, Y \in \mathbb{R}^{n\times d}$:

$$\|\beta_A(X)-\beta_A(Y)\|_2^2 = \sum_{k=1}^{D} \|\downarrow(Xa_k)-\downarrow(Ya_k)\|_2^2 = \sum_{k=1}^{D} \|P_kXa_k - Q_kYa_k\|_2^2$$

$$\overset{\Pi_k:=Q_k^TP_k}{=} \sum_{k=1}^{D} \|(\Pi_kX-Y)a_k\|_2^2 \geq \sum_{j=1}^{d} \|(\Pi_{k_j}X-Y)a_{k_j}\|_2^2$$

so that $\Pi_{k_1} = \cdots = \Pi_{k_d} = \Pi_0$ (pigeonhole principle: needs $D > (d-1)n!$).

Highlights of proofs

Universal keys

The upper bound is imediate. For lower bound, fix $X, Y \in \mathbb{R}^{n\times d}$:

$$\|\beta_A(X) - \beta_A(Y)\|_2^2 = \sum_{k=1}^{D} \| \downarrow(Xa_k) - \downarrow(Ya_k)\|_2^2 = \sum_{k=1}^{D} \|P_k X a_k - Q_k Y a_k\|_2^2$$

$$\overset{\Pi_k := Q_k^T P_k}{=} \sum_{k=1}^{D} \|(\Pi_k X - Y)a_k\|_2^2 \geq \sum_{j=1}^{d} \|(\Pi_{k_j} X - Y)a_{k_j}\|_2^2$$

so that $\Pi_{k_1} = \cdots = \Pi_{k_d} = \Pi_0$ (pigeonhole principle: needs $D > (d-1)n!$). Then:

$$\|\beta_A(X) - \beta_A(Y)\|_2^2 \geq \sum_{j=1}^{d} \|(\Pi_0 X - Y)a_{k_j}\|_2^2 \overset{full\ \ spark}{\geq} s_d(A[J])^2 \|\Pi_0 X - Y\|^2$$

$$\geq s_d(A[J])^2 \min_{\Pi \in \mathcal{S}_n} \|\Pi X - Y\|^2 = s_d(A[J])^2 d(\hat{X}, \hat{Y})^2$$

Highlights of proofs

Bi-Lipschitz Property

The proof resembles the treatment of phase retrieval problem:

1. Homogeneity and compactness reduce the problem to local analysis.
2. The encoder is "locally" linearized. The failure of local lower Lipschitz bound implies a certain behavior for a Quadratically Constrained Ratio of Quadratics (QCRQ).
3. QCRQ has a minimizer:inf $\Rightarrow$ min. [Teboulle&al.]
   This step took most of time and lots of (self)convincing !
4. Contradiction to injectivity assumption.

Highlights of proofs

Dimension Reduction

The proof follows the approach in [Cahill&al.], [Dufresne]:

$$0 = B(\beta_A(X)) - B(\beta_A(Y)) \Rightarrow \beta_A(X) - \beta_A(Y) \in \ker(B)$$

Need to show: $\beta_A(X) - \beta_A(Y) = 0$, or, $Ran(\Delta) \cap \ker(B) = \{0\}$, where

$$\Delta : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times D} \;,\; \Delta(X,Y) = \beta_A(X) - \beta_A(Y).$$

In the polynomial case, [Cahill&al.] exploit arguments from algebraic geometry. Here the problem is simpler since $Ran(\Delta)$ is included in a finite union of linear subspaces of dimension at most $2nd$.

By a dimension argument it follows that the target space for B must be of dimension at least $2nd$ to obtain an injective embedding. In this case, generically, $Ran(\Delta)$ and $\ker(B)$ intersect transversally.

Towards universal keys

The arXiv preprint provides necessary and sufficient conditions for a key to be universal.

Open Problem: Given (n, d) find the smallest dimension D so that there exists a universal key $A \in \mathbb{R}^{d\times D}$ for $\mathbb{R}^{n\times d}$.

So far we obtained (joint with Daniel Levy (UMD)):

n	d	D-d
2	2	1
3	2	2
4	2	2
5	2	3
6	2	≥ 4

Open Problem: If a universal key exists for a triple (n, d, D) then is it true that universal keys are generic in $\mathbb{R}^{d\times D}$?

Table of Contents

The Protein Dataset

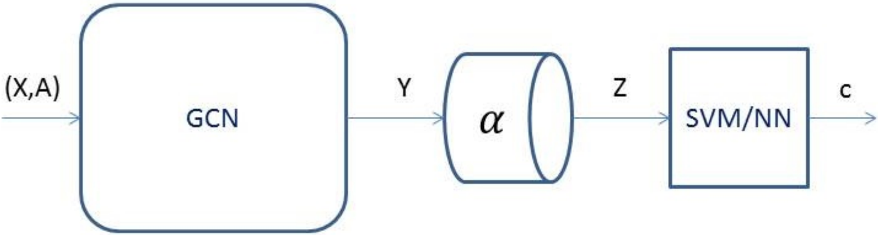
Protein Dataset: 663 non-enzymes and 450 enzymes out of 1113 proteins. Each graph associated to one protein: nodes represent amino acids and edges represent the bonds between them. Number of nodes (aminoacids): varying between 20 and 620 with average of 39. Input feature vectors os size $r = 29$.

Task: the task is classification of each protein into *enzyme* or *non-enzyme*.

The Deep Network Architecture

Architecture: ReLU activation and

- GCN with $L = 3$ layers and 29 input feature vectors, and 50 hidden nodes in each layer; no dropouts, no batch normalization. output of GCN: $d = 1, 10, 50, 100$.
- Mid-layer component: α
- Fully connected NN with dense 3-layers and 150 internal units; no dropouts, with batch normalization.



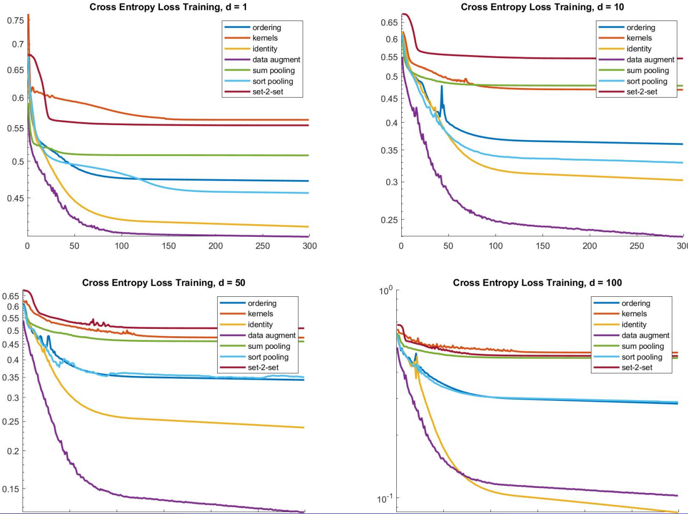
The Network

Training has been done over 300 epochs with a batch size of 128. Loss function: binary cross-entropy.
The following 7 α modules have been tested:

1. identity: $\alpha(X) = X$; no permutation invariance.
2. data augmentation: $\alpha(X) = X$ BUT the training data set has been augmented with 4 random permutatons of each graph.
3. ordering: $\alpha(X) = \downarrow(XA)$, $A = [I\ 1]$
4. kernels: $\alpha(X) = (\sum_{k=1}^{n} exp(-\|x_k - a_j\|^2))_{1\leq j\leq m=5nd}$
5. sumpooling: $\alpha(X) = 1^T X$
6. sort-pooling: sorted by last column
7. set-to-set: introduced in [Vinyals&al.]

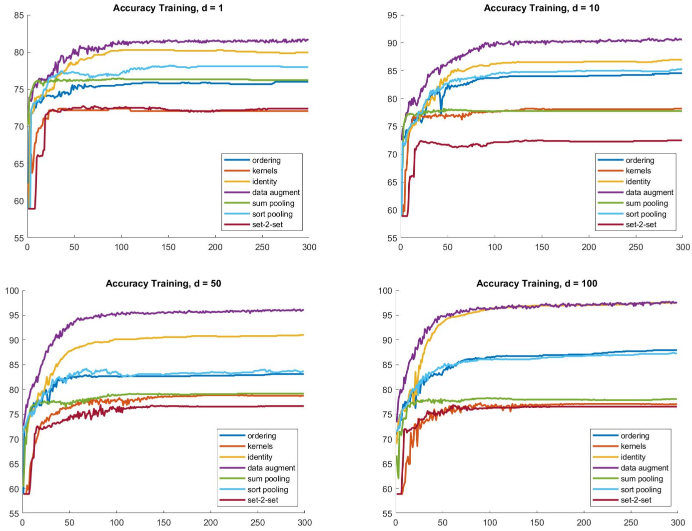
Enzyme Classification Example

Training Loss: X Entropy



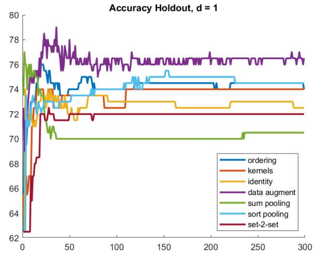
Enzyme Classification Example

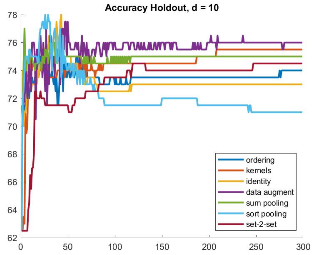
Accuracy on Training set

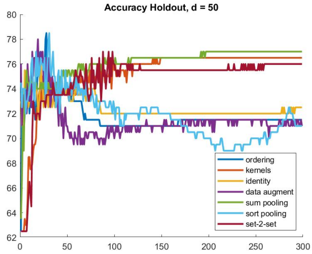


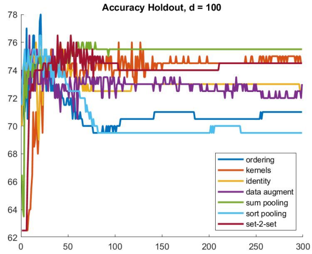
Enzyme Classification Example

Accuracy on Holdout data



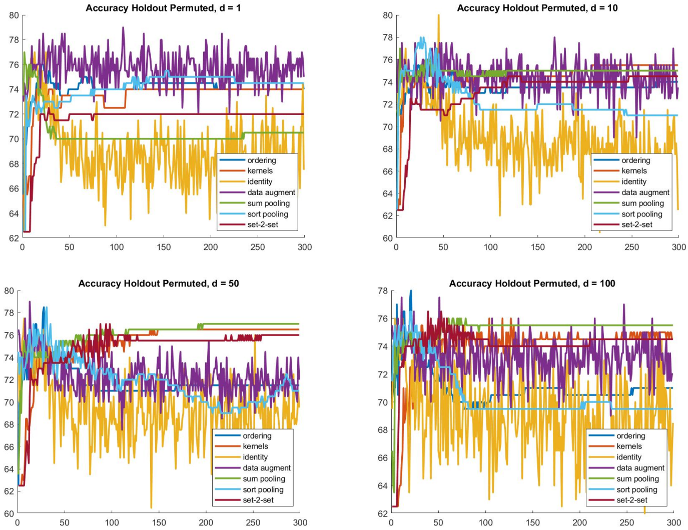






Enzyme Classification Example

Accuracy on Holdout data with nodes randomly permuted



Performance Results: Accuracy

d = 50	ordering	kernels	identity	data augment	sum-pooling	sort-pooling	set-2-set
Training	83.1	78.8	91	96	79.2	83.7	76.7
Holdout	71.5	76.5	72.5	71	77	71	76
Holdout Perm	71.5	76.5	69.5	72	77	71	76

Table: Accuracy ACC(%) for enzyme/non-enzyme classification of the seven algorithms on PROTEINS_FULL dataset after 300 epochs for embedding dimension $d = 50$

For comparison: [Dobson&al.] obtain an accuracy of 77-80% using an SVM based classifier.

The QM9 Dataset

Dataset: Consists of about 134,000 isomers of organic molecules made up of CHONF, each containing 10-29 atoms. see http://quantum-machine.org/datasets/ Nodes corresponds to atoms; each feature vector containins geometry (x,y,z coordinates), partial charge per atom (Mulliken charge), and atom type.

Task: the task is regression: predict a physical feature (electron energy gap $\Delta\varepsilon$) computed for each molecule.

Architecture: ReLU activation and

- GCN with $L = 3$ layers and 50 hidden nodes in each layer; no dropouts, no batch normalization; zero padding to $m = 29$ number of rows. output of GCN: $d = 1, 10, 50, 100$.
- Mid-layer component: α
- Fully connected NN with dense 3-layers and 150 internal units in each of the two hidden layers; no dropouts, with batch normalization.

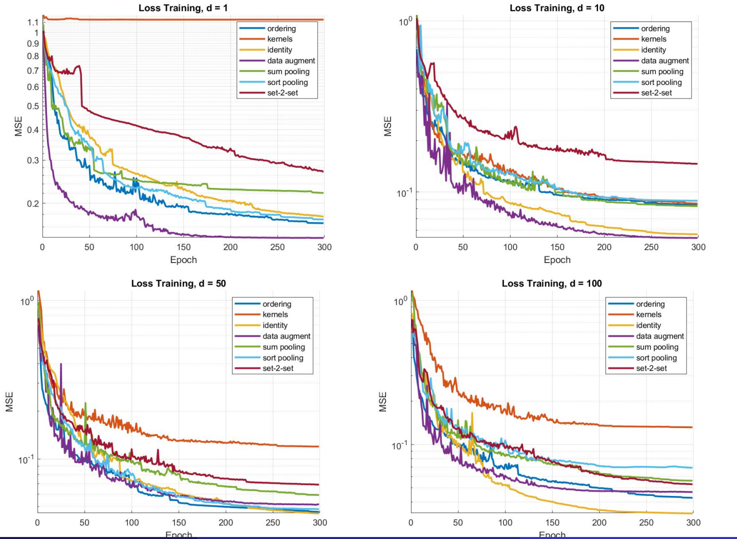
The Network

Training has been done over 300 epochs with a batch size of 128. Loss function: Mean-Square Error (MSE).
The same 7 α modules have been tested:

1. identity: $\alpha(X) = X$; no permutation invariance.
2. data augmentation: $\alpha(X) = X$ BUT the training data set has been augmented with 4 random permutatons of each graph.
3. ordering: $\alpha(X) = \downarrow (XA)$, $A = [I \ 1]$
4. kernels: $\alpha(X) = (\sum_{k=1}^{n} exp(-\|x_k - a_j\|^2))_{1\leq j\leq m=5nd}$
5. sumpooling: $\alpha(X) = 1^T X$
6. sort-pooling: sorted by last column
7. set-to-set: introduced in [Vinyals&al.]

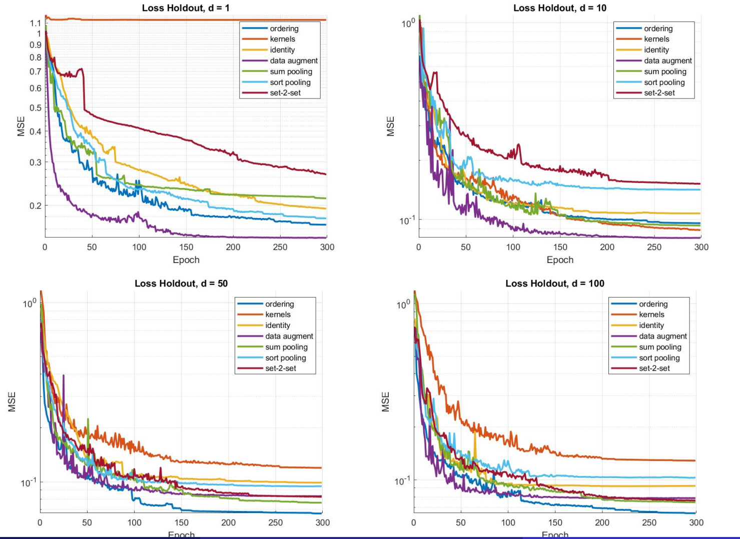
QM9 Regression Example

Training MSE



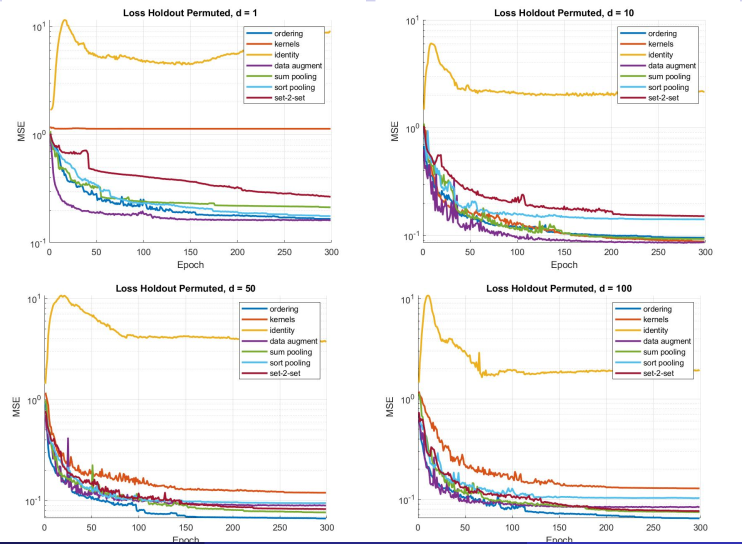
QM9 Regression Example

Validation MSE



QM9 Regression Example

Validation MSE with Random Permutations



Performance Results: MAE

d = 100	ordering	kernels	identity	data augment	sum-pooling	sort-pooling	set-2-set
Training	0.155	0.269	0.139	0.164	0.178	0.199	0.173
Holdout	0.187	0.267	0.227	0.206	0.201	0.239	0.201
Holdout Perm	0.187	0.267	1.086	0.213	0.201	0.239	0.201

Table: Mean Absolute Error (MAE) for regression of the electron energy gap $\Delta\varepsilon = LUMO - HOMO$ (eV) of the seven algorithms on QM9 dataset after 300 epochs for embedding dimension $d = 100$

For comparison:

- chemical accuracy is $0.043eV$
- the best ML method [Gilmer&al.] achieves MAE of $0.053eV$
- Coulomb method [Rupp&al.] achieves MAE of $0.229eV$

Bibliography

[1] Vinyals, O., Bengio, S. Kudlur, M., Order Matters: Sequence to sequence for sets, ICLR 2016.
[2] Sutskever, I., Vinyals, O., and Le, Q. V., Sequence to Sequence Learning with Neural Networks, arXiv e-prints , arXiv:1409.3215 (Sep 2014).
[3] Bello, I., Pham, H., Le, Q. V., Norouzi, M., and Bengio, S., Neural Combinatorial Optimization with Reinforcement Learning, arXiv e-prints , arXiv:1611.09940 (Nov 2016).
[4] Williams, R. J., Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning 8(3-4), 229-256 (1992).
[5] Kool, W., van Hoof, H., and Welling, M., Attention, Learn to Solve Routing Problems, arXiv e-prints , arXiv:1803.08475 (Mar 2018).

Bibliography

[6] Dai, H., Khalil, E. B., Zhang, Y., Dilkina, B., and Song, L., Learning Combinatorial Optimization Algorithms over Graphs, arXiv e-prints , arXiv:1704.01665 (Apr 2017).
[7] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al., Human-level control through deep reinforcement learning, Nature 518(7540), 529 (2015).
[8] Dai, H., Dai, B., and Song, L., Discriminative embeddings of latent variable models for structured data, in International conference on machine learning, 2702-2711 (2016).
[9] Nowak, A., Villar, S., Bandeira, A. S., and Bruna, J., Revised Note on Learning Algorithms for Quadratic Assignment with Graph Neural Networks, arXiv e-prints , arXiv:1706.07450 (Jun 2017).

Bibliography

[10] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G., The graph neural network model, IEEE Transactions on Neural Networks 20(1), 61-80 (2008).
[11] Li, Z., Chen, Q., and Koltun, V., Combinatorial Optimization with Graph Convolutional Networks and Guided Tree Search, arXiv e-prints , arXiv:1810.10659 (Oct 2018).
[12] Kipf, T. N. and Welling, M., Semi-Supervised Classification with Graph Convolutional Networks, arXiv e-prints , arXiv:1609.02907 (Sep 2016).
[13] Kingma, D. P. and Ba, J., Adam: A Method for Stochastic Optimization, arXiv e-prints , arXiv:1412.6980 (Dec 2014).
[14] H. Derksen, G. Kemper, Computational Invariant Theory, Springer 2002.

Bibliography

[15] J. Cahill, A. Contreras, A.C. Hip, Complete Set of translation Invariant Measurements with Lipschitz Bounds, arXiv:1903.02811 (2019).
[16] M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Poczos, R. Salakhutdinov, A.J. Smola, Deep Sets, arXiv:1703.06114
[17] H. Maron, E. Fetaya, N. Segol, Y. Lipman, On the Universality of Invariant Networks, arXiv:1901.09342 [cs.LG] (May 2019).
[18] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam and P. Vandergheynst, "Geometric Deep Learning: Going beyond Euclidean data," in IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 18-42, July 2017, doi: 10.1109/MSP.2017.2693418.
[19] S. Ravanbaksh, J. Schneider, B. Poczos, Equivariance through parameter sharing, ICML 2017.

W. Li, W. Liao, "Stable super-resolution limit and smallest singular value of restricted Fourier matrices", Applied and Computational Harmonic Analysis, vol. 51, 118-156, 2021.
[21] P.D. Dobson, A.J. Doig, "Distinguishing Enzyme Structures from Non-enzymes without Alignments", J. Mol. Biol. 330, 771-783, 2003.

Thank you!

Questions?

The Embedding Problem

Notations (2)

Definition

Fix $X \in \mathbb{R}^{n\times d}$. A matrix $A \in \mathbb{R}^{d\times D}$ is called admissible *for X if $\beta_A^{-1}(\beta_A(X)) = \hat{X}$. In other words, if $Y \in \mathbb{R}^{n\times d}$ so that $\downarrow(XA) = \downarrow(YA)$ then there is $\Pi \in \mathcal{S}_n$ sot that $Y = \Pi X$.*

We denote by $\mathcal{A}_{d,D}(X)$ (or $\mathcal{A}(X)$) the set of admissible keys for X.

Definition

Fix $A \in \mathbb{R}^{d\times D}$. A data matrix $X \in \mathbb{R}^{n\times d}$ is said separated by A *if $A \in \mathcal{A}(X)$.*

We let $\mathcal{S}(A)$ denote the set of data matrices separated by A.
The key A is universal iff $\mathcal{S}(A) = \mathbb{R}^{n\times d}$.

Genericity Results for $d \geq 2$

Admissible keys

Theorem

Let $X \in \mathbb{R}^{n\times d}$. For any $D \geq d+1$ the set $\mathcal{A}_{d,D}(X)$ of admissible keys for X is dense in $\mathbb{R}^{d\times D}$ with respect to Euclidean topology, and it is generic with respect to Zariski topology. In particular, $\mathbb{R}^{d\times D} \setminus \mathcal{A}_{d,D}(X)$ has Lebesgue measure 0, i.e., almost every key is admissible for X.

Proof

It is sufficient to consider the case $D = d+1$. Also, it is sufficient to analyze the case $A = [I_d \;\; b]$ and to show that a generic $b \in \mathbb{R}^d$ defines an admissible key. The vector $b \in \mathbb{R}^d$ does **not** define an admissible key if there are $\Xi, \Pi_1, \cdots, \Pi_d \in S_n$ so that for $Y = [\Pi_1 x_1, \cdots, \Pi_d x_d]$,

$$Yb = \Xi X b \;\; \text{but} \;\; Y - \Pi X \neq 0 \;,\; \forall \Pi \in \mathcal{S}_n$$

Define the linear operator

Genericity Results for $d \geq 2$

Admissible keys

Proof - cont'd
Let

$$\mathcal{P} = \left\{ (\Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^d \;\; \forall \Pi \in \mathcal{S}_n, \exists k \in [d] \; s.t. \; (\Pi - \Pi_k) x_k \neq 0 \right\}$$

Then

$$\{ b \in \mathbb{R}^d : [I_d \; b] \text{ not admissible for } X \} = \bigcup_{(\Xi; \Pi_1, \cdots, \Pi_d) \in \mathcal{S}_n \times \mathcal{P}} \ker(B(\Xi; \Pi_1, \cdots, \Pi$$

It is now sufficient to show that each null space has dimension less than d. Indeed, the alternative would mean $B(\Xi; \Pi_1, \cdots, \Pi_d) = 0$ but this would imply $(\Pi_1, \cdots, \Pi_d) \notin \mathcal{P}$. $\square$

Non-Universality of vector keys

Insufficiency of a single vector key

The following is a no-go result, which shows that there is no universal single vector key for data matrices tall enough.

Proposition

If $d \geq 2$ and $n \geq 3$,

$$\bigcup_{X \in \mathbb{R}^{n\times d}} \{ b \in \mathbb{R}^d \; : \; A = [I_d \;\; b] \text{ not admissible for} X \} = \mathbb{R}^d.$$

Consequently,

$$\bigcap_{X \in \mathbb{R}^{n\times d}} \mathcal{A}_{d,d+1}(X) = \emptyset.$$

On the other hand, for $n = 2$, $d = 2$, any vector $b \in \mathbb{R}^2$ with $b_1 b_2 \neq 0$ defines a universal key $A = [I_2 \;\; b]$.

Non-Universality of vector keys

Insufficiency of a single vector key - cont'd

Proof

To show the result, it is sufficient to consider a counterexample for $n = 3$, $d = 2$, with key $b = [1, 1]^T$.

$$X = \begin{bmatrix} 1 & -1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \quad , \quad Y = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

Then $Xb = [0, -1, 1]^T$ and $Yb = [1, 0, -1]^T$, yet $X \not\sim Y$. Thus $[I_2 \; b]$ is not admissible for X.

Then note if $a \in \mathbb{R}^d$ so that $[I_d \; a]$ is admissible for X then for any $P \in S_d$ and L an invertible $d \times d$ diagonal matrix, $L^{-1}P^TA \in \mathcal{A}_{d,1}(XPL)$. This shows how for any $b \in \mathbb{R}^2$, one can construct $X \in \mathbb{R}^{3\times 2}$ so that $b \notin \mathcal{A}_{2,1}(X)$.

For $n > 3$ or $d > 2$, proof follows by embedding this example.

Genericity Results for $d \geq 2$

Admissible Data Matrices

Theorem

Assume $a \in \mathbb{R}^d$ is a vector with non-vanishing entries, i.e., $a_1 a_2 \cdots a_d \neq 0$. Then for any $n \geq 1$, $\mathcal{S}([I_d\ \ a])$ is dense in $\mathbb{R}^{n\times d}$ and includes an open dense set with respect to Zariski topology. In particular, $\mathbb{R}^{n\times d} \setminus \mathcal{S}([I_d\ \ a])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the vector key a.

Genericity Results for $d \geq 2$

Admissible Data Matrices

Theorem

Assume $a \in \mathbb{R}^d$ is a vector with non-vanishing entries, i.e., $a_1 a_2 \cdots a_d \neq 0$. Then for any $n \geq 1$, $\mathcal{S}([I_d \;\; a])$ is dense in $\mathbb{R}^{n\times d}$ and includes an open dense set with respect to Zariski topology. In particular, $\mathbb{R}^{n\times d} \setminus \mathcal{S}([I_d \;\; a])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the vector key a.

Corollary

Assume $A \in \mathbb{R}^{d\times(D-d)}$ is a matrix such that at least one column has non-vanishing entries. Then for any $n \geq 1$, $\mathcal{S}([I_d \;\; A])$ is dense in $\mathbb{R}^{n\times d}$ and is generic with respect to Zariski topology. In particular, $\mathbb{R}^{n\times d} \setminus \mathcal{S}([I_d \;\; A])$ has Lebesgue measure 0, i.e., almost every data matrix X is separated by the matrix key $[I_d \;\; A]$.

Proof that $\mathcal{S}([I_d\ A])$ is generic

The case $D > d$

Assume $A \in \mathbb{R}^{d\times(D-d)}$ satisfies $A_{1,k}A_{2,k}\cdots A_{d,k} \neq 0$ for some $k \in [D-d]$. The set of non-separated data matrices $X \in \mathbb{R}^{n\times d}$ (i.e., the complement of $\mathcal{S}([I_d\ A])$) factors as follows:

$$\mathbb{R}^{n\times d}\setminus\mathcal{S}([I_d\ A]) = \bigcup_{(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d)\in(\mathcal{S}_n)^D} \left(ker\ L(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d;A\right.$$

$$\left.\setminus \bigcup_{\Pi\in\mathcal{S}_n} \ker\ M(\Pi,\Pi_1,\cdots,\Pi_d)\right) \quad (*)$$

where, with $A = [a_1,\cdots,a_{D-d}]$, $X = [x_1,\cdots,x_d]$:

$$L(\Xi_1,\cdots,\Xi_{D-d};\Pi_1,\cdots,\Pi_d;A):\mathbb{R}^{n\times d}\to\mathbb{R}^{n\times D-d} \ ,\ (L((\ldots)X)_k=[(\Xi_k-\Pi_1)x_1,\cdots,(\Xi_k-\Pi_d)x_d]a_k \ ,\ k\in[D-$$

$$M(\Pi,\Pi_1,\cdots,\Pi_d):\mathbb{R}^{n\times d}\to\mathbb{R}^{n\times d} \quad ,\quad M(\Pi,\Pi_1,\cdots,\Pi_d)X=[(\Pi-\Pi_1)x_1,\cdots,(\Pi-\Pi_d)x_d]$$

Proof that $\mathcal{S}(A)$ is generic

cont'd

1. The outer union can be reduced by noting that on the "diagonal" Δ,

$$\Delta = \{(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^D \quad , \quad \Pi_1 = \Pi_2 = \cdots = \Pi_d\}$$

$$M(\Pi_1, \Pi_1, \cdots, \Pi_d) = 0 \rightarrow \bigcup_{\Pi \in \mathcal{S}_n} \ker M(\Pi, \Pi_1, \cdots, \Pi_d) = \mathbb{R}^{n\times d}$$

2. If $(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d) \in (\mathcal{S}_n)^D \setminus \Delta$ then for every $k \in [D-d]$ there is $j \in [d]$ such that $\Xi_k - \Pi_j \neq 0$. In particular choose the k column of A that is non-vanishing. Let $x_j \in \mathbb{R}^n$ so that $(\Xi_k - \Pi_j)x_j \neq 0$. Consider the matrix $X = [0, \cdots, 0, x_j, 0, \cdots, 0]$ where x_j is the only non identically 0 column. Claim: $X \notin \ker L(\Xi_1, ..., \Pi_d; A)$. Indeed, the resulting k column of $L()X$ is $A_{j,k}(\Xi_k - \Pi_j)x_j \neq 0$. It follows that

$$\dim \ker L(\Xi_1, \cdots, \Xi_{D-d}; \Pi_1, \cdots, \Pi_d; A) < nd$$

Hence $\mathbb{R}^{n\times d} \setminus \mathcal{S}([I_d \;\; A])$ is a finite union of subsets of closed linear spaces properly included in $\mathbb{R}^{n\times d}$. This proves the theorem. $\square$

Additional Relations

Note the following relationship and matrix representation of X when matrices are column-stacked:

$$M(\Pi,\Pi_1,\cdots,\Pi_d) = L(\Pi,\cdots,\Pi;\Pi_1,\cdots,\Pi_d;I)$$

$$L \equiv \left[\begin{array}{cccc} A_{1,1}(\Xi_1-\Pi_1) & A_{2,1}(\Xi_1-\Pi_2) & \cdots & A_{d,1}(\Xi_1-\Pi_d) \\ A_{1,2}(\Xi_2-\Pi_1) & A_{2,2}(\Xi_2-\Pi_2) & \cdots & A_{d,2}(\Xi_2-\Pi_d) \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,D-d}(\Xi_{D-d}-\Pi_1) & A_{2,D-d}(\Xi_{D-d}-\Pi_2) & \cdots & A_{d,D-d}(\Xi_{D-d}-\Pi_d \end{array}\right.$$

a $n(D-d)\times nd$ matrix.