G-Invariant Representations using Coorbits

Radu Balan

Applied Mathematics, Applied Statistics and Scientific Computing (AMSC) Program
Department of Mathematics and Norbert Wiener Center for Harmonic Analysis and Applications
University of Maryland, College Park, MD

April 10, 2025
Trinity College Dublin

Norbert Wiener Center for Harmonic Analysis and Applications

College of Computer, Mathematical, & Natural Sciences
AMSC | Applied Mathematics and Statistics, and Scientific Computation Program

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."

Collaborators:
Nadav Dym (Technion)
Daniel Levy (UMD, Princeton)
Naveed Haghani (UMD,APL-JHU)
Maneesh Singh (Verisk)
Efstratios Tsoukanis (UMD, Claremont Graduate Center - CA)
Matthias Wellershoff (UMD)

Preprints

Preprints:

1. R.B., Naveed Haghani, Maneesh Singh, "Permutation Invariant Representations with Applications to Graph Deep Learning", arXiv: 2203.07546 [math.FA] , [cs.LG]
2. R.B., Efstratios Tsoukanis, "Relationships between the Phase Retrieval Problem and Permutation Invariant Embeddings", arXiv:2306.13111 [math.FA] , [cs.IT] , [math.IT]
3. R.B., Efstratios Tsoukanis, "G-Invariant Representations using Coorbits: Bi-Lipschitz Properties", arXiv:2308.11784 [math.RT]
4. R.B., Efstratios Tsoukanis, "G-Invariant Representations using Coorbits: Injectivity Properties", arXiv:2310.16365 [math.RT]
5. R.B, Efstratios Tsoukanis, Matthias Wellershoff, "Stability of sorting based embeddings", arXiv:2410.05446 [math.FA]

Table of Contents:

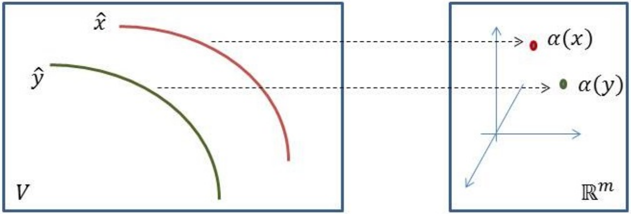
Table of Contents

High-Level View

In this talk, we discuss Euclidean embeddings of metric spaces induced by orthogonal representations of finite groups G acting on a linear space V with inner product.

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

$$a_0\mathbf{d}([x],[y]) \le \|\alpha([x]) - \alpha([y])\|_2 \le b_0\mathbf{d}([x],[y]).$$



The Program

Given a discrete group G acting unitarily on a normed real space V, we formulate four general problems

1. Construct injective embeddings of the quotient space V/G, $\alpha : \hat{V} \to \mathbb{R}^m$. The injectivity problem.
2. Construct/Obtain bi-Lipschitz properties for the Euclidean embedding $\alpha : \hat{V} \to \mathbb{R}^m$. The stability problem.
3. Develop algorithms for inversion $\alpha^{-1} : \mathbb{R}^m \to \hat{V}$. The recovery problem.
4. Analyze specific cases. Applications.

The Program

Given a discrete group G acting unitarily on a normed real space V, we formulate four general problems

1. Construct injective embeddings of the quotient space V/G, $\alpha : \hat{V} \to \mathbb{R}^m$. The injectivity problem.
2. Construct/Obtain bi-Lipschitz properties for the Euclidean embedding $\alpha : \hat{V} \to \mathbb{R}^m$. The stability problem.
3. Develop algorithms for inversion $\alpha^{-1} : \mathbb{R}^m \to \hat{V}$. The recovery problem.
4. Analyze specific cases. Applications.

The Program

Given a discrete group G acting unitarily on a normed real space V, we formulate four general problems

1. Construct injective embeddings of the quotient space V/G, $\alpha : \hat{V} \to \mathbb{R}^m$. The injectivity problem.
2. Construct/Obtain bi-Lipschitz properties for the Euclidean embedding $\alpha : \hat{V} \to \mathbb{R}^m$. The stability problem.
3. Develop algorithms for inversion $\alpha^{-1} : \mathbb{R}^m \to \hat{V}$. The recovery problem.
4. Analyze specific cases. Applications.

Today we discuss results about the first two problems: injectivity, bi-Lipschitz stability.

Table of Contents

I. 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.

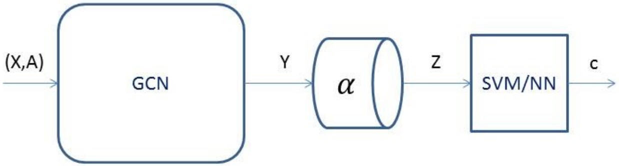
Contruct a map $f : (A, X) \rightarrow f(A, X)$ that performs:

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 Deep Learning with GCN/GNN

Our approach for these learning tasks (classification or regression) is based on the following scheme (see GCN[1] and equivariance[2]):



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.

Our focus is on the α component.

[1] Kipf, T. N. and Welling, M., Semi-Supervised Classification with Graph Convolutional Networks, arXiv e-prints , arXiv:1609.02907 (Sep 2016).

[2] H. Maron, E. Fetaya, N. Segol, Y. Lipman, On the Universality of Invariant Networks, arXiv:1901.09342 [cs.LG] (May 2019).

II. Assignment Problems

The Graph Isomorphism Problem

Consider two graphs $G = (\mathcal{V}, \mathcal{E})$ and $\tilde{G} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ with n nodes. The graph isomorphism problem is the computational problem of determining whether these graphs are identical after a relabeling of nodes.

If A and $\tilde{A}$ denote their adjacency matrices, these graphs are isomorphic if and only if $\tilde{A} = \Pi A \Pi^T$ for some permutation matrix $\Pi \in \mathcal{S}_n$.

Current state-of-the-art (Wikipedia): Babai (2015,2017) presented a quasi-polynomial algorithm with running time $2^{O((\log n)^c)}$, for some fixed $c > 0$. Helfgott (2017) claims that one can take $c = 3$.

Similar problem can be stated for weighted graphs: $A, \tilde{A} \in \mathrm{Sym}(n)$ with nonnegative entries, isomorphic if and only if $\tilde{A} = \Pi A \Pi^T$ for some $\Pi \in \mathcal{S}_n$.

Graph Alignment Problems

Consider two $n \times n$ symmetric matrices A, B. The "vanilla" alignment problem for quadratic forms asks for the orthogonal matrix $U \in O(n)$ that minimizes

$$\|UAU^T - B\|_F^2 := trace((UAU^T - B)^2) = \|A\|_F^2 + \|B\|_F^2 - 2trace(UAU^T B).$$

The solution is well-known and depends on the eigendecomposition of matrices A, B: if $A = U_1 D_1 U_1^T$, $B = U_2 D_2 U_2^T$ then

$$U_{opt} = U_2 U_1^T \quad , \quad \|U_{opt} A U_{opt}^T - B\|_F^2 = \sum_{k=1}^{n} |\lambda_k - \mu_k|^2,$$

where $D_1 = diag(\lambda_k)$ and $D_2 = diag(\mu_k)$ are diagonal matrices with eigenvalues ordered monotonically.

Quadratic Assignment Problem (QAP)

The challenging case is when U is constrained to the permutation group as is the case in the *graph matching problem*. In this case, the optimization problem becomes

$$\min_{U \in \mathcal{S}_n} \|UAU^T - B\|_F$$

which turns into a QAP: $\max_{U \in \mathcal{S}_n} trace(UAU^T B)$.
This is equivalent to computing the natural distance $d(\hat{A}, \hat{B}) = \min_{P,Q \in \mathcal{S}_n} \|PAP^T - QBQ^T\|_F$ between the equivalence classes $\hat{A}, \hat{B} \in \widehat{\mathrm{Sym}(n)}$ induced by action $(\Pi, A) \mapsto \Pi A \Pi^T$.
How is this connected to the embedding problem? If one can design an *efficient* nearly isometric map $\Phi : Sym(n) \to \mathbb{R}^m$ so that
(1) $\Phi(PAP^T) = \Phi(A)$ for all $P \in \mathcal{S}_n$ and $A \in Sym(n)$, and

(2) $$(1-\delta) \min_{P \in \mathcal{S}_n} \|PAP^T - B\| \leq \|\Phi(A) - \Phi(B)\| \leq (1+\delta) \min_{P \in \mathcal{S}_n} \|PAP^T - B\|,$$

then the QAP solved efficiently up to a multiplicative factor.

Table of Contents

Problem Setup

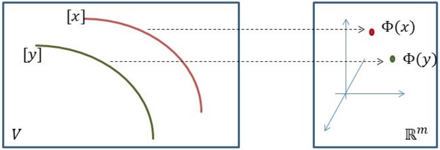
Consider a group $G \subset O(d)$ acting on the Euclidean space $V = \mathbb{R}^d$.

General problem

Construct an embedding map $\Phi : V \to \mathbb{R}^m$

1. Invariance: $\Phi(U_g x) = \phi(x)\ \forall g \in G, x \in V$
2. Injectivity: if $\Phi(x) = \Phi(y)$ then there exists $g \in G$ so that $y = U_g x$.
3. Φ is bi-Lipschitz on $(\hat{V} = V/G, \mathbf{d})$:

$$a_0 \inf_{u\in[x],v\in[y]} \|u - v\| \leq \|\Phi(x) - \Phi(y)\| \leq b_0 \inf_{u\in[x],v\in[y]} \|u - v\|.$$



Approaches

Over the past many years, several constructions have been proposed:

1. Invariant Polynomials: Hilbert, Noether, ..., Cahill[3], Bandeira[4]
2. Kernels: replace monomials by other kernels, e.g. $e^{i\omega x}$, e^{-x^2}, $\sigma(\langle x, a\rangle)$[5]
3. Sorting: extends the 1-D sorting, $x \mapsto \downarrow x$ [6],[7]

1+2: *sum pooling* layer; 3: extension of *max pooling* layer in deep nets[8], [9].

[3] J. Cahill, A. Contreras, A.C. Hip, Complete Set of translation Invariant Measurements with Lipschitz Bounds, Appl. Comput. Harm. Anal. 49 (2020), 521–539.

[4] A. Bandeira, B. Blum-Smith, J. Kileel, J. Niles-Weed, A. Perry, A.S. Wein, Estimation under group actions: Recovering orbits from invariants, ACHA 66 (2023)

[5] D. Yarotsky, Universal approximations of invariant maps by neural networks, Constructive Approximation (2021)

[6] R. Balan, N. Haghani, M.Singh, Permutation Invariant Representations with Applications to Graph Deep Learning, arXiv:2203.07546

[7] J. Cahill, J.W. Iverson, D.G. Mixon, D. Packer, Group-invariant max filtering, arXiv:2205.14039.

[8] O. Vinyals, S. Bengio, M. Kudlur, Order Matters: Sequence to sequence for sets, Proc. ICLR 2016

Sorting based Representations and G-invariance

Assume V is a real d-dimensional Hilbert space and G a finite orthogonal group of size $N = |G|$, acting on V, $\{U_g, g \in G\}$.
Fix a generator $w \in V$ (call it, *window*, or *template*, or *wavelet*) and consider the nonlinear map induced by sorting its coorbit:

$$\phi_w : V \to \mathbb{R}^N \quad , \quad \phi_w(x) = \downarrow ((\langle x, U_g w\rangle)_{g\in G}) .$$

where $\downarrow (y) = (y_{\pi(i)})_{i\in[N]}$ is the non-increasing sorting operator: $y_{\pi(1)} \geq \cdots \geq y_{\pi(N)}$.
Key observations:

1. $\phi_w(U_g x) = \phi_w(x)$, ϕ is G-invariant.
2. ϕ_w is piecewise linear (in fact, $\phi_w(x) = \phi_x(w)$, and $(w,x) \mapsto \phi_w(x)$ is piecewise bilinear).

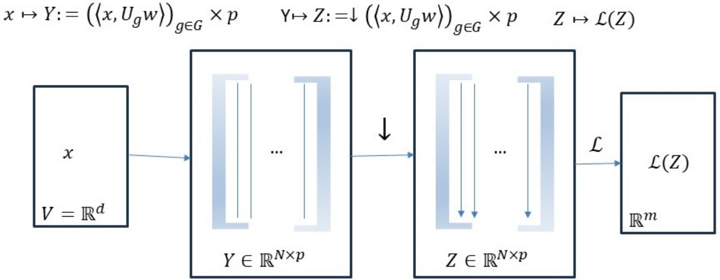
G-Invariant Coorbit Representations

For a collection $\mathbf{w} = (w_1, \cdots, w_p) \in V^p$ the sorted coorbit representation:

$$\Phi_{\mathbf{w}} : V \to \mathbb{R}^{N \times p} \quad , \quad \Phi_{\mathbf{w}}(x) = \left[\phi_{w_1}(x) | \cdots | \phi_{w_p}(x) \right] .$$

Pass through a linear operator $\mathcal{L} : \mathbb{R}^{N \times p} \to \mathbb{R}^m$, the G-invariant coorbit representation:

$$\Psi_{\mathbf{w},\mathcal{L}} : V \to \mathbb{R}^m \quad , \quad \Psi_{\mathbf{w},\mathcal{L}}(x) = \mathcal{L}(\Phi_{\mathbf{w}}(x))$$



In particular, if $S \subset [N] \times [p]$, $\Phi_{\mathbf{w},S} := \Psi_{\mathbf{w},1_S} = \Phi_{\mathbf{w}}|_S$.

G-Invariant Coorbit Representations

Special cases:
1. For $G = S_n$ and $V = \mathbb{R}^{n \times d}$ with action $(P, X) \mapsto PX$ [10] introduced the embedding $\beta_A(X) = \downarrow (XA)$, for *key* $A \in \mathbb{R}^{d \times D}$ and sorting operator acting independently in each column. This is of the type $\Psi_{\mathbf{w},\mathcal{L}}$ for $w_1 = \delta_1 \cdot a_1^T, \ldots,$ $w_D = \delta_1 \cdot a_D^T$, where $\delta_1 = (1, 0, \cdots, 0)^T$ and $A = [a_1 | \cdots | a_D]$, and $\mathcal{L}$ a restriction operator to an appropriate subset $S \subset [n!] \times [D]$ of size nD.

2. The *max filter* introduced in [11] for some template $w \in V$ is defined by $\langle\langle \cdot, w \rangle\rangle : V \to \mathbb{R}$, $\langle\langle x, w \rangle\rangle = \max_{g \in G} \langle x, U_g w \rangle$. Equivalent recasting: $\langle\langle x, w \rangle\rangle = \mathcal{L}(\Phi_w(X))$, for a restriction operator $\mathcal{L}$ to the subset $S = \{1\}$.

3. The operator $\Psi_{\mathbf{w},\mathcal{L}}$, $\Psi_{\mathbf{w},\mathcal{L}}(X) = \mathcal{L}(\Phi_{\mathbf{w}}(X))$ has been introduced in [12]

[10] R. Balan, N. Haghani, M.Singh, Permutation Invariant Representations with Applications to Graph Deep Learning, arXiv:2203.07546 (2022)

[11] J. Cahill, J. W. Iverson, D. G. Mixon, D. Packer, Group-invariant max filtering, arXiv:2205.14039 (2022)

[12] R.B, Efstratios Tsoukanis, Matthias Wellershoff, “Stability of sorting based embeddings”, arXiv:2410.05446 (2024)

Table of Contents

Main Results

Injectivity

Let $V_G = \{x \in V \;:\; U_g x = x \;,\; \forall g \in G\}$, $d_G = dim(V_G)$, $q \geq 0$ and for $g = (g_1, \cdots, g_n), h = (h_1, \cdots, h_n) \in G^n$ distinct, $\rho_n(q) = \max_{g,h} \gamma^q_{g,h}$ where

$$\gamma^q_{g,h} = semi.alg.dim.\; \{(x,y) \in V \times V \;:\; dim(span\{U_{g_k}x - U_{h_k}y, k \in [n]\}) = q\}$$

Theorem (R.B.,E.Tsoukanis '23-'25)

In any of the following cases

1. *Assume* $p \geq 2\,dim(V) - d_G$ *and set* $\mathbf{n} = (1, \cdots, 1) \in [N]^p$.
2. *Fix* $n \in [N]$ *and choose* $p > \max_{q \in [n]} \frac{1}{q}(\rho_n(q) - d_G - 1)$. *Set* $\mathbf{n} = (n, \cdots, n) \in [N]^p$.
3. *Choose* $p \geq 1$ *and* $\mathbf{n} = (n_1, \cdots, n_p) \in [N]^p$ *so that* $\max_{q_1 \in [n_1], \cdots, q_p \in [n_p]} \left(min_{i \in [p]} \rho_{n_i}(q_i) - (q_1 + \cdots + q_p)\right) \leq d_G$.

For a generic (w.r.t. Zariski topology) $\mathbf{w}$ *and for any* $S \subset [N] \times [p]$ *with* $|\{k : (k,i) \in S\}| \geq n_i$, *the map* $\Phi_{\mathbf{w},S} : (\widehat{V}, \mathbf{d}) \to (\mathbb{R}^{|S|}, \|\cdot\|_2)$ *is injective.*

Main Results (2)

Theorem (R.B, E.T., M. Wellershoff '24)

Consider the same setup as before. Assume $\mathbf{w} \in V^p$ *and* $\mathcal{L} : \mathbb{R}^{N\times p} \to \mathbb{R}^m$ *so that* $\Psi_{\mathbf{w},\mathcal{L}} : (\widehat{V}, \mathbf{d}) \to (\mathbb{R}^m, \|\cdot\|_2)$ *is* **injective**.

1. *Themap* $\Psi_{\mathbf{w},\mathcal{L}} : (\widehat{V}, \mathbf{d}) \to (\mathbb{R}^m, \|\cdot\|_2)$ *is bi-Lipschitz. Let* a_0, b_0 *denote its bi-Lipschitz constants.*
2. *If* $f : V \to H$ *is a Lipschitz continuous function so that* $f(U_g x) = f(x)$ *for all* g, x*, where* H *is a Hilbert space, then there exists a Lipschitz continuous function* $g : \mathbb{R}^m \to H$ *so that* $f = g \circ \Psi_{\mathbf{w},\mathcal{L}}$*, i.e.* $f(x) = g(\Psi_{\mathbf{w},\mathcal{L}}(x))$*. Furthermore,* $Lip(g) \leq Lip(f)/a_0$*.*
3. *Assume* $g : \mathbb{R}^m \to H$ *is a Lipschitz function with Lipschitz constant* $Lip(g)$*. Then* $f = g \circ \Psi_{\mathbf{w},\mathcal{L}} : V \to H$ *is G-invariant and Lipschitz, with Lipschitz constant* $Lip(f) \leq b_0 Lip(g)$*.*

Its proof is based on Kirszbraun's extension theorem.

Existing Results

Injectivity problem

Over the past 15 years or so, there have been works that recognized the difference between *generating polynomials* and *separating invariants*[13]

A seminal paper that resurfaces results on semi-algebraic sets is [14]. The method goes back to earlier works in phase retrieval[15].

More recently, in the context of G-invariance, [16],[17], or permutation invariance[18]

[13] Emilie Dufresne, Separating invariants and finite reflection groups, Advances in Mathematics 221 (2009), no. 6, 1979–1989.

[14] Dym Nadav, Steven J. Gortler. "Low dimensional invariant embeddings for universal geometric learning." arXiv preprint arXiv:2205.02956.

[15] R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, ACHA 20(2006)

[16] D. G. Mixon, D. Packer, Max filtering with reflection groups, arXiv:2212.05104

[17] R. Balan, E. Tsoukanis, G-invariant representations using coorbits: Injectivity properties, arXiv:2310.16365

[18] On the equivalence between graph isomorphism testing and function approximation with GNNs, Z. Chen, S. Villar, L. Chen, J. Bruna, NeurIPS 2019.

Existing Results (2)

Lipschitz and Bi-Lipschitz properties

Earlier results obtain Lipschitz/bi-Lipschitz properties on compacts, or certain classes of functions.

Global L/bi-L are harder to establish and typically rule out polynomial based embeddings.

So far only sorting based embeddings showed such global properties [19],[20], [21]

[19] R. Balan, E. Tsoukanis, G-invariant representations using coorbits: Bi-lipschitz properties, arXiv:2308.11784

[20] J. Cahill, J. W. Iverson, D. G. Mixon, Bilipschigz group invariants, arXiv:2305.17241

[21] D. G. Mixon, Y. Qaddura, Injectivity, stability, and positive definiteness of max filtering, arXiv:2212.11156

Sketch of Proof: Injectivity Result

Define the *"bad"* set of **w**'s that fail to separate all distinct classes:

$$\mathcal{F} = \{\mathbf{w} \in V^p \ , \ \exists x \not\sim y \ \Phi_{\mathbf{w}}(x) = \Phi_{\mathbf{w}}(y) \ \}.$$

The work is to embed $\mathcal{F}$ into a semi-algebraic set of semi-algebraic dimension strictly less than $pd = p\,dim(V)$.
This technique is called "lift-and-project'[22]: we construct a semi-algebraic vector bundle embedded into a certain Grassmanian vector bundle $\gamma^{\perp}_{n,k}$.
The bad set $\mathcal{F}$ is then indentified with a subset of the projection of this vector bundle into its second component.
The full result for $\Psi_{\mathbf{w},\mathcal{L}}$ follows from analyzing the semi-agebraic dimension of difference set $\{\Phi_{\mathbf{w}}(x) - \Phi_{\mathbf{w}}(y)\}$ and of the kernel of $\mathcal{L}$.

[22] R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, ACHA 20(2006)

Sketch of Proof: Lower Lipschitz bound

The proof is by contradiction. Consider the simpler case when $\mathcal{L}$ is given by restriction to a subset $S \subset [N] \times [p]$.

1. If lower Lipschitz constant vanishes, then it must vanish locally: there are $(x_n)_n, (y_n)_n$ such that

$$\lim_{n\to\infty} \frac{\|\Phi_{\mathbf{w},S}(x_n) - \Phi_{\mathbf{w},S}(y_n)\|^2}{\mathbf{d}([x_n],[y_n])^2} = 0$$

and

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} y_n = z_1, \ \|x_n\| = 1, \ \|y_n\| \leq 1, \ \|z_1\| = 1$$

and they are aligned with one another:

$$\|x_n - y_n\| = \min_{g\in G} \|x_n - U_g y_n\| \tag{4.1}$$

$$\|x_n - z_1\| = \min_{g\in G} \|x_n - U_g z_1\| \tag{4.2}$$

$$\|y_n - z_1\| = \min_{g\in G} \|y_n - U_g z_1\| \tag{4.3}$$

Lower Lipschitz bound

2. We construct inductively $z_2, z_3, ..., z_d$ such that for all $1 \leq k \leq d-1$:

$$\|z_{k+1}\| \ll \|z_k\|, \ \dim(\mathrm{span}(z_1, \ldots, z_k)) = k$$

and the local lower Lipschitz constant vanishes in a convex set $\{\sum_{r=1}^{k} a_r z_r \ , \ |a_r - 1| < \epsilon\}$.

3. For $k = d$ this construction defines a non-empty open set $\{\sum_{r=1}^{k} a_r z_r \ , \ |a_r - 1| < \epsilon\}$ where the local lower Lipschitz constant vanishes.

4. Finally, we can construct $u, v \neq 0$, so that $x = u + \sum_{r=1}^{d} z_r$ and $y = v + \sum_{r=1}^{d} z_r$ satisfy $x \neq y$ and yet

$$\Phi_{\mathbf{w},S}(x) = \Phi_{\mathbf{w},S}(y).$$

This contradicts the injectivity hypothesis.

Table of Contents

The Protein Dataset

Protein Dataset: PROTEINS_FULL[23] consists of 1113 proteins: 663 non-enzymes and 450 enzymes. 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 of size $r = 29$.

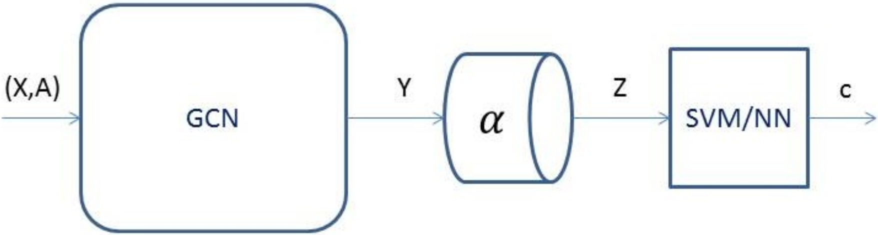
Task: the task is classification of each protein into *enzyme* or *non-enzyme*.

[23] P.D. Dobson, A.J. Doig, "Distinguishing Enzyme Structures from Non-enzymes without Alignments", J. Mol. Biol. 330, 771-783, 2003.

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.

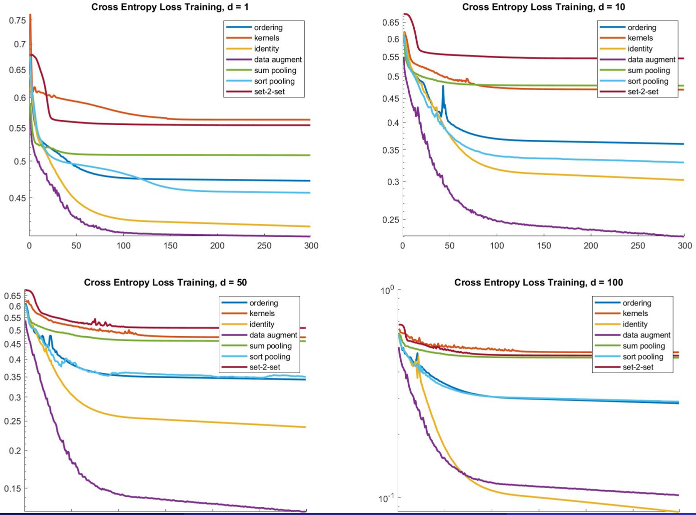
The following 7 α modules have been tested:

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.][24]

[24] Vinyals, O., Bengio, S. Kudlur, M., Order Matters: Sequence to sequence for sets, ICLR 2016.

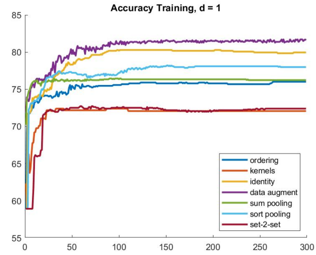
Enzyme Classification Example

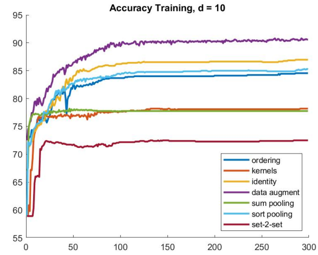
Training Loss: X Entropy

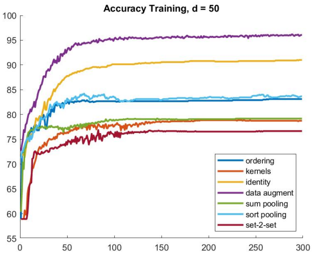


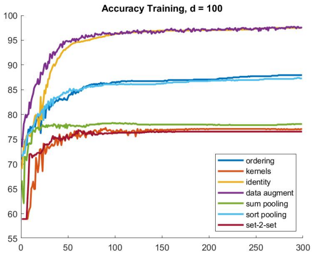
Enzyme Classification Example

Accuracy on Training set



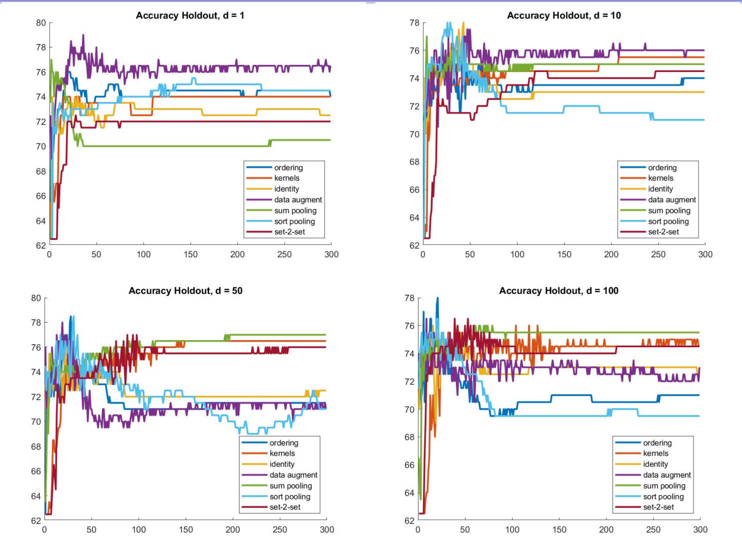






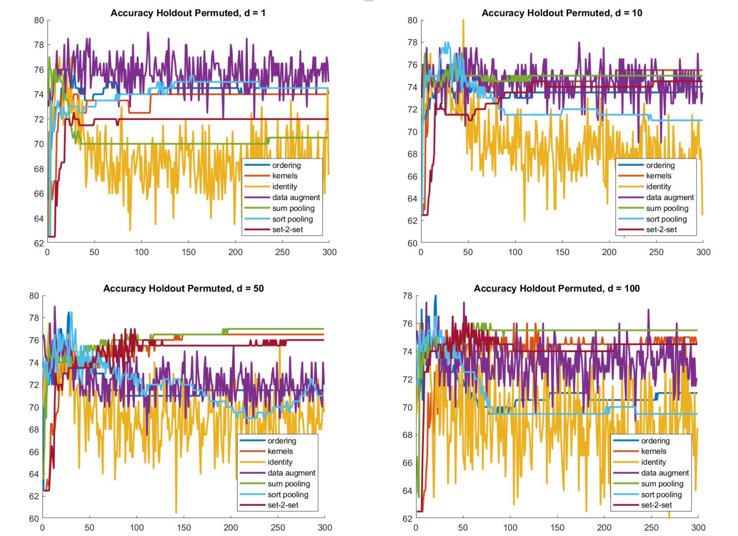
Enzyme Classification Example

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.][25] obtains an accuracy of 77-80% using an SVM based classifier.

[25] P.D. Dobson, A.J. Doig, "Distinguishing Enzyme Structures from Non-enzymes without Alignments", J. Mol. Biol. 330, 771-783, 2003.

The QM9 Dataset

Dataset: QM9[26] 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.

[26] R. Ramakrishnan, P.O. Dral, M. Rupp, and OA. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1(1):1-7, 2014.

The Network

Training has been done over 300 epochs with a batch size of 128. Loss function: Mean-Square Error (MSE).

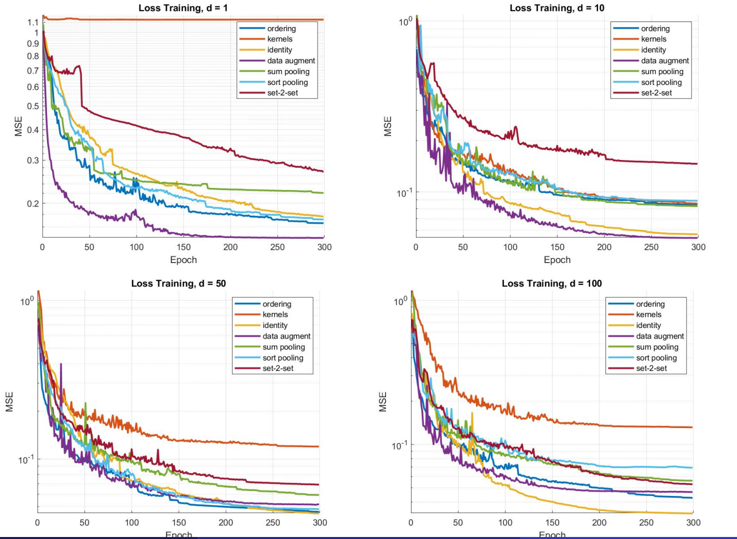
The same 7 α modules have been tested:

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.][27]

[27] Vinyals, O., Bengio, S. Kudlur, M., Order Matters: Sequence to sequence for sets, ICLR 2016

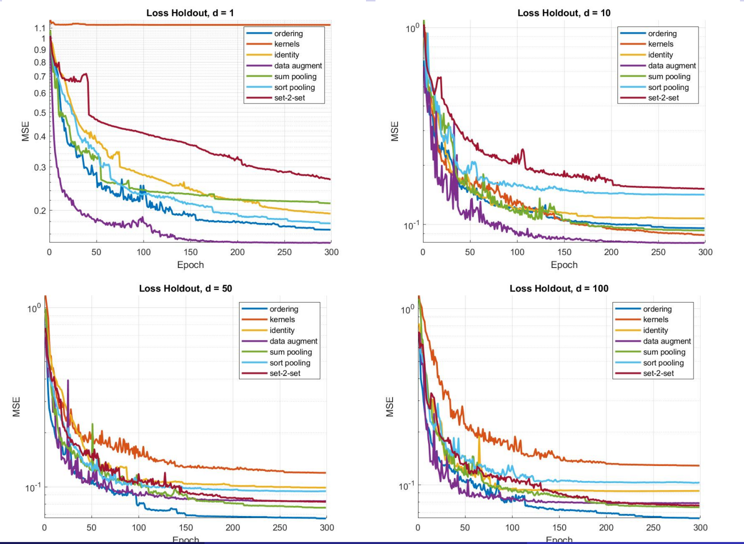
QM9 Regression Example

Training MSE



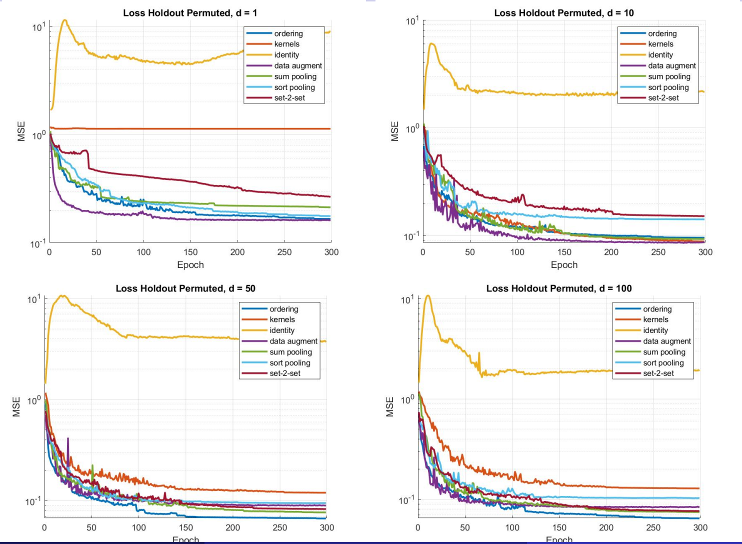
QM9 Regression Example

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 $0.229eV$

Thank you!

Questions?

Table of Contents

A Universal Embedding

Consider the map

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

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.

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, $L^{-1}P^T A \in \mathcal{A}_{d,1}(XPL)$. 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:

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

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

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

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

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

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

cont'd

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

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

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

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:

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

$$L \equiv \begin{bmatrix} A_{1,1}(\Xi_1 - \Pi_1) & A_{2,1}(\Xi_1 - \Pi_2) & \cdots & A_{d,1}(\Xi_1 - \Pi_d) \\ A_{1,2}(\Xi_2 - \Pi_1) & A_{2,2}(\Xi_2 - \Pi_2) & \cdots & A_{d,2}(\Xi_2 - \Pi_d) \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,D-d}(\Xi_{D-d} - \Pi_1) & A_{2,D-d}(\Xi_{D-d} - \Pi_2) & \cdots & A_{d,D-d}(\Xi_{D-d} - \Pi_d \end{bmatrix}$$

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

Planar Rotations

Consider the action on $\mathbb{R}^2$ of the cyclic group $< U > \simeq \mathbb{Z}_N$ generated by a planar rotation by $a = \frac{2\pi}{N}$, $U = [cos(a) \;\; -sin(a); sin(a) \;\; cos(a)]$. The quotient space $\mathbb{R}^2 / \sim$ is topolgically equivalent to a 2 dim. cone in $\mathbb{R}^3$, by identifying the positive x-semiaxis with the half-line of angle a.

Recall $\rho_n(q) = \max_{g,h} \gamma_{g,h}^q$ where

$$\gamma_{g,h}^q = semi.alg.dim. \{(x,y) \in V \times V \; : \; dim(span\{U_{g_k}x - U_{h_k}y, k \in [n]\}) = q\}$$

Explicit computation:

$$\rho_1(q) = \begin{cases} 2, & q = 0, \\ 4, & q = 1, \\ -1, & q \geq 2. \end{cases} \quad \rho_2(q) = \begin{cases} 2, & q = 0, \\ 3, & q = 1 \;\&\; N \text{ odd}, \\ 4, & q = 1 \;\&\; N \text{ even}, \\ 4, & q = 2 \\ -1, & q \geq 3. \end{cases}$$

Planar Rotations (2)

Theorem

1. *For any* $w \in \mathbb{R}^2$*, the map* $\Phi_w : \widehat{\mathbb{R}^2} \to \mathbb{R}^N$ *is never injective.*
2. *For any* $w_1, w_2 \in \mathbb{R}^2$ *and* $S = \{(1,1),(1,2)\}$ *(the max filter case), the map* $\Phi_{\mathbf{w},S} : \widehat{\mathbb{R}^2} \to \mathbb{R}^2$ *is never injective.*
3. *If* $w_1, w_2, w_3 \in \mathbb{R}^2$ *are linearly independent and* $angle(w_i, w_j) < \frac{2\pi}{N}$ *then for either* $S = \{(1,1),(1,2),(1,3)\}$ *(the max filter, or* $\mathbf{n} = (1,1,1)$ *configuration), or* $S = \{(1,1),(2,1),(1,2)\}$ *(a* $\mathbf{n} = (2,1)$ *configuration), generically, the map* $\Phi_{\mathbf{w},S} : \widehat{\mathbb{R}^2} \to \mathbb{R}^3$ *is injective and bi-Lipschitz.*

A careful analysis of our main theorem would guarantee the embedding $\Phi_{\mathbf{w},S} : \widehat{\mathbb{R}^2} \to \mathbb{R}^4$ is injective (and hence bi-Lipschitz) for certain $\mathbf{n} = (2,1,1)$ configurations, or any $\mathbf{n} = (1,1,1,1)$ configuration.

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.