Geometry of Equivariant Linear Neural Networks

Kathlén Kohn, *Anna-Laura Sattelberger*, Vahid Shahverdi
arxiv:2309.13736 [cs.LG]

Augsburger Mathematisches Kolloqium

January 30, 2024
Universität Augsburg

Motivation

Two questions

1. How do the network's properties affect the geometry of its **function space**?
   How to characterize **equivariance** or **invariance**?
2. How to **parameterize** equivariant and invariant networks?
   Which implications does it have for **network design**?

Training neural networks

Neural networks

A neural network F of depth L is a parameterized family of functions $(f_{L,\theta}, \dots, f_{1,\theta})$

$$F\colon\ \mathbb{R}^N \longrightarrow \mathcal{F}, \quad F(\theta) = f_{L,\theta} \circ \cdots \circ f_{1,\theta} =: f_\theta\,.$$

Each layer $f_{k,\theta}\colon \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ is a composition activation $\circ$ (affine-)linear.

Training a network

Given training data $\mathcal{D} = \{(\widehat{x}_i, \widehat{y}_i)_{i=1,\dots,S}\} \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$, the aim is to minimize the loss

$$\mathcal{L}\colon\ \mathbb{R}^N \xrightarrow{F} \mathcal{F} \xrightarrow{\ell_{\mathcal{D}}} \mathbb{R}\,.$$

Example: For $\ell_{\mathcal{D}}$ the squared error loss, this gives $\min_{\theta \in \mathbb{R}^N} \sum_{i=1}^{S} (f_\theta(\widehat{x}_i) - \widehat{y}_i)^2$.
On **function space**: $\min_{M \in \mathcal{F}} \|M\widehat{X} - \widehat{Y}\|^2_{\text{Frob}}$.

Critical points of $\mathcal{L}$

- ◇ **pure**: critical point of $\ell_{\mathcal{D}}$
- ◇ **spurious**: induced by parameterization

Linear convolutional networks (LCNs)

- linear: identity as activation function
- convolutional layers with filter $w \in \mathbb{R}^k$ and stride $s \in \mathbb{N}$:

$$\alpha_{w,s} \colon \mathbb{R}^d \to \mathbb{R}^{d'}, \quad (\alpha_{w,s}(x))_i = \sum_{j=0}^{k-1} w_j x_{is+j}\,.$$

Geometry of linear convolutional networks [1]

Function space $\mathcal{F}_{(\mathbf{d},\mathbf{s})}$ of LCN: semi-algebraic set, Euclidean-closed

Theorem [2]

Let $(\mathbf{d},\mathbf{s})$ be an LCN architecture with all strides > 1 and $N \geq 1 + \sum_i d_i s_i$. For almost all data $\mathcal{D} \in (\mathbb{R}^{d_0} \times \mathbb{R}^{d_L})^N$, every critical point θ_c of $\mathcal{L}$ satisfies one of the following:

1. $F(\theta_c) = 0$, or
2. θ_c is a regular point of F and $F(\theta_c)$ is a **smooth, interior point** of $\mathcal{F}_{(\mathbf{d},\mathbf{s})}$.
   In particular, $F(\theta_c)$ is a critical point of $\ell_{\mathcal{D}}|_{\mathrm{Reg}\left(\mathcal{F}^{\circ}_{(\mathbf{d},\mathbf{s})}\right)}$.

This is known to be false for...

- linear fully-connected networks
- stride-one LCNs

[1] K. Kohn, T. Merkh, G. Montúfar, M. Trager. Geometry of Linear Convolutional Networks. *SIAM J. Appl. Algebra Geom.*, 6(3):368–406, 2022.
[2] K. Kohn, G. Montúfar, V. Shahverdi, M. Trager. Function Space and Critical Points of Linear Convolutional Networks. Preprint arXiv:2304.0572, 2023.

Algebraic geometry for machine learning

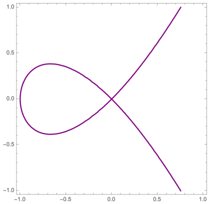
Natural points of entry

- ◇ algebraic vision [3]
- ◇ geometry of function spaces

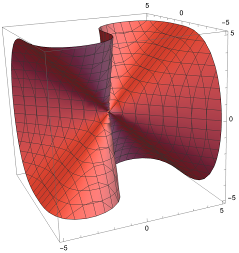
Algebraic varieties

subsets of $\mathbb{C}^n$ obtained as common **zero set of polynomials** $p_1, \ldots, p_N \in \mathbb{C}[x_1, \ldots, x_n]$

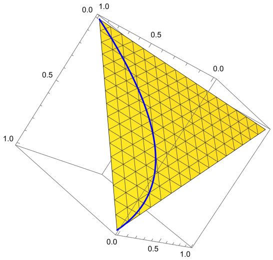
Drawing real points of algebraic varieties



$\mathcal{V}(y^2 - x^2(x+1))$
a nodal curve

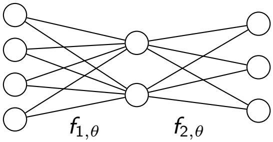


$\mathcal{V}(x^2y - y^3 - z^3)$
a cubic surface



$\mathcal{V}(p_0p_2 - (p_0 + p_1)p_1) \cap \Delta_2$
a discrete statistical model

[3] J. Kileel and K. Kohn. Snapshot of Algebraic Vision. Preprint arXiv:2210.11443, 2022.



Example

$F\colon\ \mathbb{R}^{2\times 4}\times\mathbb{R}^{3\times 2}\longrightarrow\mathbb{R}^{3\times 4},\quad (M_1,M_2)\mapsto M_2\cdot M_1$

parameter space: $\mathbb{R}^N=\mathbb{R}^{2\times 4}\times\mathbb{R}^{3\times 2},\ f_{1,\theta}=M_1,\ f_{2,\theta}=M_2$

Its function space $\mathcal{F}$ is the set of real points of the determinantal variety

$$\mathcal{M}_{2,3\times 4}(\mathbb{R}) \;=\; \left\{M\in\mathbb{R}^{3\times 4}\mid \operatorname{rank}(M)\le 2\right\}.$$

The determinantal variety $\mathcal{M}_{r,m\times n}$

For $M=(m_{ij})_{i,j}\in\mathbb{C}^{m\times n}$: $\operatorname{rank}(M)\le r\;\Leftrightarrow$ all $(r+1)\times(r+1)$ minors of M vanish.
Define polynomials in m_{ij}

$$\mathcal{M}_{r,m\times n}\;=\;\{M\mid \operatorname{rank}(M)\le r\}\subset\mathbb{C}^{m\times n}.$$

Well studied! $\dim(\mathcal{M}_{r,m\times n})\;=\;r\cdot(m+n-r)$, $\mathcal{M}_{r,m\times n}(\mathbb{R})$, singularities, ...

Invariant functions

$f_\theta: \mathbb{R}^n \to \mathbb{R}^r \to \mathbb{R}^m$ — $r < \min(m, n)$

$G = \langle \sigma_1, \ldots, \sigma_g \rangle \leq \mathcal{S}_n$ — a permutation group, acting on $\mathbb{R}^n$ by permuting the entries
induced action on M: permuting its columns

Invariance under $\sigma \in \mathcal{S}_n$: $f_\theta \circ \sigma \equiv f_\theta$

Decomposing into cycles

The decomposition $\sigma = \pi_1 \circ \cdots \circ \pi_k$ of σ into k disjoint cycles induces a partition $\boxed{\mathcal{P}(\sigma) = \{A_1, \ldots, A_k\}}$ of the set $[n] = \{1, \ldots, n\}$. $A_1, \ldots, A_k \subset [n]$ pairwise disjoint sets

Example: The permutation $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{smallmatrix}\right) = (1\,3\,4)(2\,5) \in \mathcal{S}_5$ induces the partition $\mathcal{P}(\sigma) = \{\{1, 3, 4\}, \{2, 5\}\}$ of $[5] = \{1, 2, 3, 4, 5\}$. For $\eta = (1\,4\,3)(2\,5) \neq \sigma$: $\mathcal{P}(\eta) = \mathcal{P}(\sigma)$.

Characterizing invariance $\quad MP_\sigma \overset{!}{=} M$

Let $\sigma \in \mathcal{S}_n$ and $\mathcal{P}(\sigma) = \{A_1, \ldots, A_k\}$ its induced partition. A matrix $M = (m_1 | \cdots | m_n)$ is invariant under $\sigma = \pi_1 \circ \cdots \circ \pi_k$ if and only if for each i, the columns $\{m_j\}_{j \in A_i}$ **coincide**.

$\Rightarrow$ If M is invariant under σ, its rank is at most k.

Example: rotation-invariance for $p \times p$ pictures

Setup: $n = p^2$ an even square number, $f_\theta : \mathbb{R}^n \to \mathbb{R}^n$ linear

$\sigma \in \mathcal{S}_n$: rotating a $p \times p$ picture clockwise by 90 degrees:

$$\sigma : \ \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}, \quad \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pp} \end{pmatrix} \mapsto \begin{pmatrix} a_{p1} & a_{p-1,1} & \dots & a_{11} \\ a_{p2} & a_{p-1,2} & \dots & a_{12} \\ \vdots & \vdots & \ddots & \vdots \\ a_{pp} & a_{p,p-1} & \dots & a_{1p} \end{pmatrix}$$

Identify $\mathbb{R}^{p\times p} \cong \mathbb{R}^n$ via $A \mapsto (a_{1,1}, a_{1,p}, a_{p,p}, a_{p,1}, a_{1,2}, a_{2,p}, a_{p,p-1}, a_{p-1,1}, \dots, a_{1,p-1}, a_{p-1,p}, a_{p,2}, a_{2,1}, a_{2,2}, a_{2,p-1}, a_{p-1,p-1}, a_{p-1,2}, \dots, a_{\frac{p}{2},\frac{p}{2}}, a_{\frac{p}{2},\frac{p}{2}+1}, a_{\frac{p}{2}+1,\frac{p}{2}}, a_{\frac{p}{2}+1,\frac{p}{2}+1})^\top$.

Under this identification, σ acts on $\mathbb{R}^n$ by the $n \times n$ block matrix

$$\begin{pmatrix} 0 & 0 & 0 & 1 & & & & & \\ 1 & 0 & 0 & 0 & & & & & \\ 0 & 1 & 0 & 0 & & & & & \\ 0 & 0 & 1 & 0 & & & & & \\ & & & & \dots & & & & \\ & & & & & 0 & 0 & 0 & 1 \\ & & & & & 1 & 0 & 0 & 0 \\ & & & & & 0 & 1 & 0 & 0 \\ & & & & & 0 & 0 & 1 & 0 \end{pmatrix}.$$

N.B.: σ-invariance of f_θ implies that columns 1–4, 5–8, ..., $(n-3)$–n of M **coincide**.

Properties of $\mathcal{I}^G_{r,m\times n} \subset \mathcal{M}_{r,m\times n}$

$G = \langle \sigma_1, \ldots, \sigma_g \rangle \leq \mathcal{S}_n$	a permutation group
$\sigma_i = \pi_{i,1} \circ \cdots \circ \pi_{i,k_i},\ i = 1, \ldots, g$	decomposition into pairwise disjoint cycles π_i

Reduction to cyclic case

There exists $\sigma \in \mathcal{S}_n$ such that $\boxed{\mathcal{I}^G_{r,m\times n} = \mathcal{I}^\sigma_{r,m\times n}}$. Any σ for which $\mathcal{P}(\sigma)$ is the **finest common coarsening** of $\mathcal{P}(\sigma_1), \ldots, \mathcal{P}(\sigma_g)$ does the job!

Proposition

Let $G = \langle \sigma \rangle \leq \mathcal{S}_n$ be cyclic, and $\sigma = \pi_1 \circ \cdots \circ \pi_k$ its decomposition into pairwise disjoint cycles π_i. The variety $\mathcal{I}^\sigma_{r,m\times n}$ is isomorphic to the determinantal variety $\mathcal{M}_{\min(r,k),m\times k}$ via a linear isomorphism $\psi_{\mathcal{P}(\sigma)} \colon \mathcal{I}^\sigma_{r,m\times n} \to \mathcal{M}_{\min(r,k),m\times k}$. deleting repeated columns

Via that, we can determine $\dim(\mathcal{I}^\sigma_{r,m\times n})$, $\deg(\mathcal{I}^\sigma_{r,m\times n})$, and $\mathrm{Sing}(\mathcal{I}^\sigma_{r,m\times n})$.

Example ($m = 2,\ n = 5,\ r = 1$)

Let $\sigma = (1\,3\,4)(2\,5) \in \mathcal{S}_5$ and hence $k = 2$. Any invariant matrix $M \in \mathcal{M}_{2\times 5}(\mathbb{R})$ is of the form $\left(\begin{smallmatrix} a & c & a & a & c \\ b & d & b & b & d \end{smallmatrix}\right)$ for some $a, b, c, d \in \mathbb{R}$. The rank constraint $r = 1$ imposes that $(c, d) = \lambda \cdot (a, b)^\top$ for some $\lambda \in \mathbb{R}$, where we assume that $(a, b) \neq (0, 0)$. Then

$$\boxed{\psi_{\mathcal{P}(\sigma)} \colon \begin{pmatrix} a & \lambda a & a & a & \lambda a \\ b & \lambda b & b & b & \lambda b \end{pmatrix} \mapsto \begin{pmatrix} a & \lambda a \\ b & \lambda b \end{pmatrix}.}$$

Parameterizing invariance & network design

$\mathcal{S}_n \ni \sigma = \pi_1 \circ \cdots \circ \pi_k,\ \mathcal{P}(\sigma) = \{A_1, \ldots, A_k\}$

Invariance of $M \in \mathcal{M}_{m\times n}$: forces columns $\{m_j\}_{j\in A_i}$ to coincide. For each i, remember representative m_{A_i} so that $\boxed{\psi_{\mathcal{P}(\sigma)}(M) = (m_{A_1} \,|\cdots|\, m_{A_k}) \in \mathcal{M}_{m\times k}\,.}$

Parameterization

Any σ-invariant $M \in \mathcal{M}_{m\times n}$ of rank k factorizes as $M = \psi_{\mathcal{P}(\sigma)}(M) \cdot (e_{i_1}|\cdots|e_{i_n})$.
i-th standard unit vector in column j for all $j \in A_i$

Fibers of multiplication map

Let $r \leq \min(m, n)$. Denote by $\mu\colon\ \mathcal{M}_{m\times r} \times \mathcal{M}_{r\times n},\ (A, B) \mapsto A \cdot B$. If $\operatorname{rank}(M) = r$ and $M = \mu(A, B)$ for some A, B, then the fiber of μ over M is

$$\mu^{-1}(M) = \left\{\left(AT^{-1}, TB\right) \mid T \in \mathrm{GL}_n(\mathbb{C})\right\} \subset \mathcal{M}_{m\times r} \times \mathcal{M}_{r\times n}\,.$$

$\mathcal{S}_n \ni \sigma$	permutation splitting into disjoint cycles $\pi_1 \circ \cdots \circ \pi_k$
$\mathcal{P}(\sigma)$	induced partition $\{A_1, \ldots, A_k\}$ of $[n]$
$E_{\mathcal{P}(\sigma)}$	the $k \times n$ matrix with e_i in column j for all $j \in A_i$

Proposition

Let M be invariant under σ and of rank k. **Any** factorization $M = A \cdot B$ is of the form

$$\boxed{(A, B) \in \left\{ \left(\psi_{\mathcal{P}(\sigma)}(M) \cdot T^{-1}, T \cdot E_{\mathcal{P}(\sigma)} \right) \mid T \in \mathrm{GL}_n \right\}.}$$

This parameterization imposes a **weight sharing property** on the encoder!

Proposition

Let $\sigma \in S_n$ consist of k disjoint cycles and let $r \leq k$. Consider the linear autoencoder $\mathbb{R}^n \to \mathbb{R}^r \to \mathbb{R}^n$ with fully-connected dense decoder $\mathbb{R}^r \to \mathbb{R}^n$ and encoder $\mathbb{R}^n \to \mathbb{R}^r$, with σ-weight sharing on the encoder. Its function space is $\mathcal{I}_{r,n\times n}(\mathbb{R})$.

Example

Let $m = n = 5$, $r = 2$ and $\sigma = (1\,3\,4)(2\,5) \in \mathcal{S}_5$. If a matrix $M = AB \in \mathcal{I}^{\sigma}_{2,5\times 5}$ is invariant under σ, the encoder factor B has to fulfill the following weight sharing property:

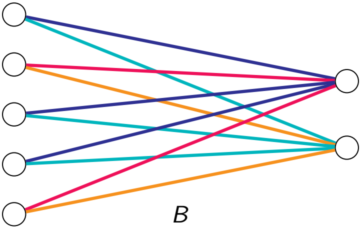


Figure: The σ-weight sharing property imposed on the encoder.

Motivation: complexity during and after training

1. For an arbitrary learned function, find a nearest invariant function.
2. Training invariant networks: determine pure critical points for Euclidean loss.

Definition

The **Euclidean distance (ED) degree** of an algebraic variety $\mathcal{X}$ in $\mathbb{R}^N$ is the number of complex critical points of the squared Euclidean distance from $\mathcal{X}$ to a general point outside the variety. It is denoted by $\mathrm{EDdegree}(\mathcal{X})$.

Examples: $\mathrm{EDdegree}(\text{circle}) = 2\,,\ \ \mathrm{EDdegree}(\text{ellipse}) = 4\,.$

ED degree of $\mathcal{M}_{r,m\times n}(\mathbb{R})$ and $\mathcal{I}^{\sigma}_{r,m\times n}(\mathbb{R})$

Let $\sigma = \pi_1 \circ \cdots \circ \pi_k \in \mathcal{S}_n$ and $r \leq \min(m,n)$. Then

- ◇ $\mathrm{EDdegree}(\mathcal{M}_{r,m\times n}(\mathbb{R})) = \binom{\min(m,n)}{r}\,,$
- ◇ $\mathrm{EDdegree}\left(\mathcal{I}^{G}_{r,m\times n}(\mathbb{R})\right) = \mathrm{EDdegree}\left(\mathcal{M}_{\min(r,k),m\times k}(\mathbb{R})\right) = \binom{\min(m,k)}{\min(r,k)}\,.$

[4] J. Draisma, E. Horobeţ, G. Ottaviani, B. Sturmfels, R. R. Thomas. The Euclidean Distance Degree of an Algebraic Variety. *Found. Comp. Math.*, 16:99–149, 2016.

$$f_\theta: \mathbb{R}^n \longrightarrow \mathbb{R}^r \longrightarrow \mathbb{R}^n \qquad r < n$$

$G = \langle\sigma\rangle \leq \mathcal{S}_n$ a **cyclic** permutation group generated by a single $\sigma \in \mathcal{S}_n$

Equivariance under σ: $f_\theta \circ \sigma \equiv \sigma \circ f_\theta$.

For matrices: M equivariant iff $MP_\sigma = P_\sigma M$. commutator of P_σ

In- and output

- $n = p^2$: $p \times p$ image with real pixels
- $n = p^3$: cubic 3D scenery

Finding good bases

Exploiting similarity transforms of the form

$$P_\sigma = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \overset{\sim T_1}{\mapsto} \left(\begin{array}{ccc|cc} 0 & 0 & 1 & & \\ 1 & 0 & 0 & \multicolumn{2}{c}{0} \\ 0 & 1 & 0 & & \\ \hline & 0 & & 0 & 1 \\ & & & 1 & 0 \end{array}\right) \overset{\sim T_2}{\mapsto} \begin{pmatrix} 1 & & & & \\ & \zeta_3 & & & \\ & & \zeta_3^2 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix}.$$

permutation matrix — block circulant matrix — diagonal matrix

Second base change involves complex Vandermonde matrices. EDdegree not preserved!

Finding good bases

After a real, orthogonal base change Q_σ, the rotation $\sigma \in \mathcal{S}_9$ is represented by

$$\mathrm{I}_3 \oplus (-\mathrm{I}_2) \oplus \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Matrices that commute with it:

$$\left(\begin{array}{ccc|cc|cccc} \alpha_{11} & \alpha_{12} & \alpha_{13} & & & & & & \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \multicolumn{2}{c|}{0} & \multicolumn{4}{c}{0} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & & & & & & \\ \hline \multicolumn{3}{c|}{0} & \beta_{12} & \beta_{22} & \multicolumn{4}{c}{0} \\ & & & \beta_{21} & \beta_{23} & & & & \\ \hline & & & & & \gamma_1 & -\gamma_2 & \delta_1 & -\delta_2 \\ \multicolumn{3}{c|}{0} & \multicolumn{2}{c|}{0} & \gamma_2 & \gamma_1 & \delta_2 & \delta_1 \\ & & & & & \epsilon_1 & -\epsilon_2 & \eta_1 & -\eta_2 \\ & & & & & \epsilon_2 & \epsilon_1 & \eta_2 & \eta_1 \end{array}\right).$$

Realization map

$$\mathcal{R}\colon\ \mathbb{C} \longrightarrow \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \,\middle|\, a, b \in \mathbb{R} \right\}, \quad z \mapsto \begin{pmatrix} \mathfrak{Re}(z) & -\mathfrak{Im}(z) \\ \mathfrak{Im}(z) & \mathfrak{Re}(z) \end{pmatrix}.$$

scaled rotation matrix

Characterizing equivariance

Proposition

There is a one-to-one correspondence between the irreducible components of $\mathcal{E}^{\sigma}_{r,n\times n}(\mathbb{R})$ that contain a matrix of rank r and the non-negative integer solutions $\mathbf{r} = (r_{l,m})$ of

$$\boxed{r_{1,1} + r_{2,1} + \sum_{l\geq 3} \sum_{\substack{m\,\in\,(\mathbb{Z}/l\mathbb{Z})^{\times},\\ \frac{1}{2}<\frac{m}{l}<1}} 2\cdot r_{l,m} \;=\; r\,, \quad \text{where} \;\; 0 \leq r_{l,m} \leq d_l\,.}$$

d_l the dimension of the eigenspace of P_σ of the eigenvalue $\zeta_l = e^{2\pi i/l}$

The irreducible component $\mathcal{E}^{\sigma,\mathbf{r}}_{r,n\times n}(\mathbb{R})$ corresponding to such an integer solution $\mathbf{r}$ after the real orthogonal base change Q_σ is

$$\mathcal{M}_{r_{1,1},d_1\times d_1}(\mathbb{R}) \times \mathcal{M}_{r_{2,1},d_2\times d_2}(\mathbb{R}) \times \prod_{l\geq 3} \prod_{\substack{m\,\in\,(\mathbb{Z}/l\mathbb{Z})^{\times},\\ \frac{1}{2}<\frac{m}{l}<1}} \mathcal{R}(\mathcal{M}_{r_{l,m},d_l\times d_l}(\mathbb{C}))\,.$$

Via that: dim ✓ deg ✓ EDdegree ✓ Sing ✓

Consequence

Equivariant linear functions can not be parameterized by a single neural network! One needs to parameterize each irreducible component of $\mathcal{E}^{\sigma}_{r,n\times n}$ separately.

Weight sharing on de- and encoder

The real irreducible component $(\mathcal{E}^{\sigma;\mathbf{r}}_{3,9\times 9})^{\sim Q_\sigma}$ with $\mathbf{r} = (1,0,1)$ is

$$\mathcal{M}_{1,3\times 3}(\mathbb{R}) \times \mathcal{M}_{0,2\times 2}(\mathbb{R}) \times \mathcal{R}\left(\mathcal{M}_{1,2\times 2}(\mathbb{C})\right) .$$

Every matrix in this component can be obtained as product of a 9×3 and a 3×9 matrix of the form $\quad * \in \mathbb{R},\ \star \in \mathbb{C}$

$$\begin{pmatrix} * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & \mathcal{R}(\star & \star) & \\ 0 & 0 & 0 & 0 & 0 & & & & \end{pmatrix}^\top \cdot \begin{pmatrix} * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & & \mathcal{R}(\star & \star) & \\ 0 & 0 & 0 & 0 & 0 & & & & \end{pmatrix} .$$

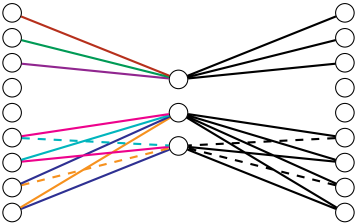


Figure: Weight-sharing of the encoder and decoder matrices. Edges of the same color share the same weight—and differ by sign, in case one of the edges is dashed.

Training on MNIST

MNIST	60.000 images of handwritten digits, size 28×28 each
$\mathbb{R}^{784} \to \mathbb{R}^{r} \to \mathbb{R}^{784}$	linear autoencoder, bottleneck $r = 99$
$\sigma \in \mathcal{S}_{784}$	permutation of pixels: translating to the right

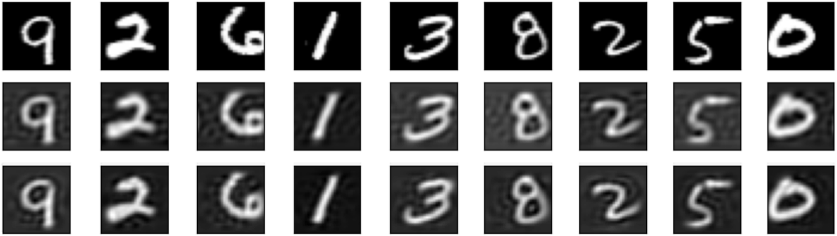


Figure: *Top row:* Nine samples from the MNIST [5] test dataset, shifted horizontally randomly by up to six pixels. *Middle row:* Output of a linear equivariant autoencoder designed to be equivariant under horizontal translations. The network architecture is determined by the integer vector $\mathbf{r}$. *Bottom row:* Output of a dense linear autoencoder with $r = 99$ without equivariance imposed.

[5] L. Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Pro. Mag.*, 29(6):141–142, 2012.

Training on MNIST

Irreducible components

$\mathcal{E}^{\sigma}_{99,784\times 784}$ has **many** irreducible components: 72,425,986,088,826
Choose component $\mathcal{E}^{\sigma,\mathbf{r}}_{99,784\times 784}$ corresponding to

$$\begin{aligned}\mathbf{r} &= (r_{1,1}, r_{28,27}, r_{14,13}, r_{28,25}, r_{7,6}, r_{28,23}, r_{14,11}, r_{4,3}, r_{7,5}, r_{28,19}, r_{14,9}, r_{28,17}, r_{7,4}, r_{28,15}, r_{2,1}) \\ &= (13, 10, 9, 8, 7, 5, 3, 1, 0, 0, 0, 0, 0, 0, 0)\,.\end{aligned}$$

Training loss

	Equivariant	equal-rank equivariant	high-pass equivariant	non-equivariant
Loss	0.0082	0.0206	0.1063	**0.0057**

Table: Comparison of average square loss values per pixel between linear equivariant and non-equivariant autoencoders on the MNIST test dataset.

Efficiency of equivariant architecture

Significant drop in number of parameters compared to general dense linear autoencoder!

$2 \cdot 99 \cdot 784 = 155{,}232 \longrightarrow 5{,}544 = 2 \cdot (28 \cdot 13 + 2 \cdot 28 \cdot (10 + 9 + 8 + 7 + 5 + 3 + 1))$

Implementations in `Python`

Soon to be available at https://github.com/vahidshahverdi/Equivariant

Conclusion

Key points: algebraic geometry helps for...

1. a thorough study of function spaces of linear neural networks.
   fully connected, convolutional
2. understanding the training process.
   locating critical points of the loss
3. the design of neural networks.
   rank constraint, weight sharing properties
4. determining the complexity during and post training.
   ED degree of real varieties

Future work

- full characterization of equivariance
  non-cyclic permutation groups
- variation of the network architecture
  more layers, non-linear activation functions

Tack för uppmärksamheten!