Geometry of Linear Neural Networks: Equivariance and Invariance under Permutation Groups

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

DMV Annual Meeting 2023
Section: Mathematics of Data Science

September 26, 2023
TU Ilmenau

www.alsattelberger.de/talks/
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 $(f_{L, \theta}, \ldots, f_{1, \theta})$

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

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

Training a network

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

$$\mathcal{L} : \mathbb{R}^N \xrightarrow{F} \mathcal{F} \xrightarrow{\ell_D} \mathbb{R}.$$ 

Example: For $\ell_D$ the squared error loss, this gives

$$\min_{\theta \in \mathbb{R}^N} \sum_{i=1}^{S} (f_{\theta}(\hat{x}_i) - \hat{y}_i)^2.$$ 

On function space: $\min_{M \in \mathcal{F}} \|M \hat{X} - \hat{Y}\|_{\text{Frob.}}^2$.

Critical points of $\mathcal{L}$

- **pure**: critical point of $\ell_D$  
- **spurious**: induced by parameterization
Linear convolutional networks (LCNs)

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

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

Geometry of linear convolutional networks [1]

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

**Theorem** [2]

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

1. $F(\theta_c) = 0$, or
2. $\theta_c$ is a regular point of $F$ and $F(\theta_c)$ is a smooth, interior point of $\mathcal{F}_{(d,s)}$.

In particular, $F(\theta_c)$ is a critical point of $\ell_D|_{\text{Reg}(\mathcal{F}^\circ_{(d,s)})}$.

This is known to be false for...

- linear fully-connected networks
- stride-one LCNs

---

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

$\mathcal{V}(p_0 p_2 - (p_0 + p_1)p_1) \subset \Delta_2$ a discrete statistical model

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

Fully connected linear neural networks

Example

\[ F : \mathbb{R}^{2 \times 4} \times \mathbb{R}^{3 \times 2} \rightarrow \mathbb{R}^{3 \times 4}, \quad (M_1, M_2) \mapsto M_2 \cdot M_1 \]

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

\[
\mathcal{M}_{2,3\times4}(\mathbb{R}) = \left\{ M \in \mathbb{R}^{3 \times 4} \mid \text{rank}(M) \leq 2 \right\}.
\]

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

For \( M = (m_{ij})_{i,j} \in \mathbb{C}^{m \times n} : \text{rank}(M) \leq r \iff \text{all } (r + 1) \times (r + 1) \text{ minors of } M \text{ vanish.} \)

Define

\[
\mathcal{M}_{r,m\times n} = \{ M \mid \text{rank}(M) \leq r \} \subset \mathbb{C}^{m \times n}.
\]

Well studied! \( \dim(\mathcal{M}_{r,m\times n}) = r(m + n - r), \ \mathcal{M}_{r,m\times n}(\mathbb{R}), \text{ singularities, } \ldots \)
Invariant functions

\[ f_\theta : \mathbb{R}^n \to \mathbb{R}^r \to \mathbb{R}^m \quad r < \min(m, n) \]

\( G = \langle \sigma_1, \ldots, \sigma_g \rangle \leq S_n \)

a permutation group, acting on \( \mathbb{R}^n \) by permuting the entries

induced action on \( M \): permuting its columns

Invariance under \( \sigma \in S_n \):

\[ f_\theta \circ \sigma \equiv f_\theta \]

Decomposing into cycles

The decomposition \( \sigma = \pi_1 \circ \cdots \circ \pi_k \) of \( \sigma \) into \( k \) disjoint cycles induces a partition

\[ \mathcal{P}(\sigma) = \{ A_1, \ldots, A_k \} \]

of the set \( [n] = \{1, \ldots, n\} \). \( A_1, \ldots, A_k \subset [n] \) pairwise disjoint sets

Example: The permutation \( \sigma = (1 \ 2 \ 3 \ 4 \ 5) = (1 \ 3 \ 4)(2 \ 5) \in S_5 \) 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

\[ M \overset{!}{\sim} P_{\sigma} \]

Let \( \sigma \in S_n \) 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.

⇒ If \( M \) is invariant under \( \sigma \), its rank is at most \( k \).
Example: rotation-invariance for $m \times m$ pictures

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

$\sigma \in S_n$: rotating an $m \times m$ picture clockwise by 90 degrees:

$$
\sigma : \mathbb{R}^{m \times m} \to \mathbb{R}^{m \times m}, \quad \begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1m} \\
  a_{21} & a_{22} & \cdots & a_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{m1} & a_{m2} & \cdots & a_{mm}
\end{pmatrix} \mapsto \begin{pmatrix}
  a_{m1} & a_{m-1,1} & \cdots & a_{11} \\
  a_{m2} & a_{m-1,2} & \cdots & a_{12} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{mm} & a_{m,m-1} & \cdots & a_{1m}
\end{pmatrix}
$$

Identify $\mathbb{R}^{m \times m} \cong \mathbb{R}^n$ via $A \mapsto (a_{1,1}, a_{1,m}, a_{m,m}, a_{m,1}, a_{1,2}, a_{2,m}, a_{m,m-1}, a_{m,1,1}, \ldots, a_{1,m-1}, a_{m-1,m}, a_{m,2}, a_{2,1}, a_{2,2}, a_{2,m-1}, a_{m-1,m-1}, a_{m-1,2}, \ldots, a_{m/2, m/2}, a_{m/2, m/2+1}, a_{m/2+1, m/2}, a_{m/2+1, m/2+1})^T$.

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

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

**N.B.:** $\sigma$-invariance of $f_\theta$ implies that columns 1–4, 5–8, \ldots, $(n - 3)$–$n$ of $M$ coincide.
Properties of $I_{r,m\times n}^G \subset M_{r,m\times n}$

$G = \langle \sigma_1, \ldots, \sigma_g \rangle \leq S_n$

a permutation group

$\sigma_i = \pi_{i,1} \circ \cdots \circ \pi_{i,k_i}, i = 1, \ldots, g$

decomposition into pairwise disjoint cycles $\pi_i$

Reduction to cyclic case

There exists $\sigma \in S_n$ such that

$I_{r,m\times n}^G = I_{r,m\times n}^\sigma$. Any $\sigma$ for which $P(\sigma)$ is the finest common coarsening of $P(\sigma_1), \ldots, P(\sigma_g)$ does the job!

Proposition

Let $G = \langle \sigma \rangle \leq S_n$ be cyclic, and $\sigma = \pi_1 \circ \cdots \circ \pi_k$ its decomposition into pairwise disjoint cycles $\pi_i$. The variety $I_{r,m\times n}^\sigma$ is isomorphic to the determinantal variety $M_{\min(r,k),m\times k}$ via a linear isomorphism $\psi_{P(\sigma)} : I_{r,m\times n}^\sigma \rightarrow M_{\min(r,k),m\times k}$. deleting repeated columns

Via that, we can determine $\dim(I_{r,m\times n}^\sigma)$, $\deg(I_{r,m\times n}^\sigma)$, and $\text{Sing}(I_{r,m\times n}^\sigma)$.

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

Let $\sigma = (1\,3\,4)(2\,5) \in S_5$ and hence $k = 2$. Any invariant matrix $M \in M_{2\times 5}(\mathbb{R})$ is of the form $\begin{pmatrix} a & c & a & c & a \\ b & d & b & d & b \end{pmatrix}$ 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

$$
\psi_{P(\sigma)} : \begin{pmatrix} a & \lambda a & a \\ b & \lambda b & b \end{pmatrix} \mapsto \begin{pmatrix} a & \lambda a \\ b & \lambda b \end{pmatrix}.
$$
Parameterizing invariance and network design

\[ S_n \ni \sigma = \pi_1 \circ \cdots \circ \pi_k, \quad \mathcal{P}(\sigma) = \{A_1, \ldots, A_k\} \]

Invariance of \( M \in \mathcal{M}_{m \times n} \): forces columns \( \{m_j\}_{j \in A_i} \) to coincide. For each \( i \), remember representative \( m_{A_i} \) and denote \( M_1 := (m_{A_1} \mid \cdots \mid m_{A_k}) \in \mathcal{M}_{m \times k} \).

**Parameterization**

Any \( \sigma \)-invariant \( M \in \mathcal{M}_{m \times n} \) of rank \( k \) factorizes as \( M = M_1 \cdot (e_{i_1} \mid \cdots \mid e_{i_n}) \), \( e_{i_j} \in \mathbb{R}^k \). \( i \)-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 \( p: \mathcal{M}_{m \times r} \times \mathcal{M}_{r \times n} \to A \cdot B \). If \( \text{rank}(M) = r \) and \( M = p(A, B) \) for some \( A, B \), then the fiber of \( p \) over \( M \) is

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

**Learning invariant linear functions with autoencoders**

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

\[ (A, B) \in \left\{ (M_1 T^{-1}, T (e_{i_1} \mid \cdots \mid e_{i_n})) \mid T \in \text{GL}_n \right\} . \]

This parameterization imposes a **weight sharing property** on the encoder!
Euclidean distance (ED) degree

Motivation: complexity during and after training

1. For an arbitrary learned function, find a nearest invariant function.

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 $\deg_{\mathrm{ED}}(\mathcal{X})$.

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

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

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

- $\deg_{\mathrm{ED}}(\mathcal{M}_{r,m\times n}(\mathbb{R})) = \binom{\min(m, n)}{r}$,
- $\deg_{\mathrm{ED}}(\mathcal{I}_{r,m\times n}^\sigma(\mathbb{R})) = \deg_{\mathrm{ED}}(\mathcal{M}_{\min(r,k),m\times k}(\mathbb{R})) = \binom{\min(m, k)}{\min(r, k)}$.

Equivariant linear autoencoders

\[ f_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^r \rightarrow \mathbb{R}^n \quad r < n \]

\[ G = \langle \sigma \rangle \leq S_n \quad \text{a cyclic permutation group generated by a single } \sigma \in S_n \]

Equivariance under \( \sigma \): \( f_\theta \circ \sigma \equiv \sigma \circ f_\theta \).

For matrices: \( M \) equivariant if \( MP_\sigma = P_\sigma M \).

In- and output

- \( n = m^2 \): \( m \times m \) image with real pixels
- \( n = m^3 \): cubic 3D scenery

Characterizing \( \mathcal{E}_{r,n \times n}^\sigma \)

- \( \text{dim}: \checkmark \)
- \( \text{deg}: \checkmark \)
- \( \text{Sing}: \checkmark \)
- \( \text{ED degree: under construction!} \)

Exploiting similarity transforms of the form

\[
P_\sigma = \begin{pmatrix}
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
\end{pmatrix}
\sim \tau_1
\begin{pmatrix}
0 & 0 & 1 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
\end{pmatrix}
\sim \tau_2
\begin{pmatrix}
1 & \zeta_3 & \zeta_3^2 \\
\zeta_3 & 1 & -1 \\
\end{pmatrix}.
\]
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 invariance implies rank constraint & weight sharing property
4. determining the complexity during and post training ED degree of real varieties

Future work

- full characterization of equivariance non-cyclic permutation groups
- generalization to other groups e.g. non-discrete groups
- variation of the network architecture more layers, non-linear activation functions

Thank you for your attention!
Characterizing invariance

\[ S_n \ni \sigma = \pi_1 \circ \cdots \circ \pi_k \]

decomposition of \( \sigma \) into \( k \) pairwise disjoint cycles

\[ \psi_{\mathcal{P}(\sigma)} : \mathcal{I}_{r,m \times n}^\sigma \cong \mathcal{M}_{\min(r,k),m \times k} \]

linear isomorphism

Properties of \( \mathcal{I}_{r,m \times n}^\sigma \)

\[
\dim(\mathcal{I}_{r,m \times n}^\sigma) = \min(r,k) \cdot (m + k - \min(r,k)),
\]

\[
\deg(\mathcal{I}_{r,m \times n}^\sigma) = \prod_{i=0}^{k-\min(r,k)-1} \frac{(m+i)! \cdot i!}{(\min(r,k)+i)! \cdot (m-(\min(r,k)+i)!)},
\]

\[
\Sing(\mathcal{I}_{r,m \times n}^\sigma) = \psi_{\mathcal{P}(\sigma)}^{-1}(\mathcal{M}_{\min(r,k)-1,m \times k}).
\]

Euclidean distance degree

\[
\deg_{\text{ED}}\left(\mathcal{I}_{r,m \times n}(\mathbb{R})^G\right) = \begin{pmatrix} \min(m,k) \\ \min(r,k) \end{pmatrix}.
\]
Example

Let $m = n = 5$, $r = 2$ and $\sigma = (1\ 3\ 4)(2\ 5) \in S_5$. If a matrix $M = AB \in I_{2,5\times 5}$ is invariant under $\sigma$, the encoder factor $B$ has to fulfill the following weight sharing property:

Figure: The $\sigma$-weight sharing property imposed on the encoder.
Consider the permutation \( \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 4 & 1 & 2 \end{pmatrix} = (1 \, 3 \, 4)(2 \, 5) \in S_5 \). Then

\[
P_\sigma = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \sim_T \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 \\ 1 \end{pmatrix} \sim_T \begin{pmatrix} 1 & \zeta_3 & \zeta_3^2 \\ \zeta_3 & \zeta_3^2 & 1 \\ \zeta_3^2 & 1 & -1 \end{pmatrix}
\]

with

\[
T_1 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad T_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \zeta_3 & \zeta_3^2 \\ 1 & \zeta_3^2 & \zeta_3 \\ 0 \\ 1 & 1 & -1 \end{pmatrix} \in \text{GL}_5(\mathbb{C}),
\]

where \( \zeta_3 \) denotes the primitive 3rd root of unity \( \exp^{2\pi i / 3} \).

\textbf{N.B.}: \( T_2 \) is block diagonal with Vandermonde matrix blocks \( V(1, \zeta_3, \zeta_3^3) \) and \( V(1, -1) \).
For a subvariety $\mathcal{X} \subset \mathcal{M}_{m \times n}$ and any $T \in \text{GL}_n(\mathbb{C})$, we denote by $\mathcal{X} \cdot T$ the image of $\mathcal{X}$ under the linear isomorphism

$$\cdot T : \mathcal{M}_{m \times n} \rightarrow \mathcal{M}_{m \times n}, \quad M \mapsto MT.$$ 

**Lemma**

Let $\mathcal{X} \subset \mathcal{M}_{m \times n}$ be a subvariety and let $T \in \text{GL}_n(\mathbb{C})$. Then, $\dim(\mathcal{X} \cdot T) = \dim \mathcal{X}$, $\deg(\mathcal{X} \cdot T) = \deg \mathcal{X}$, $\text{Sing}(\mathcal{X} \cdot T) = \text{Sing}(\mathcal{X}) \cdot T$, and $(\mathcal{X} \cdot T) \cap \mathcal{M}_{r, m \times n} = (\mathcal{X} \cap \mathcal{M}_{r, m \times n}) \cdot T$ for any $r \leq \min(m, n)$.

**Notation:** For $T \in \text{GL}_n(\mathbb{C})$ and $M \in \mathcal{M}_{n \times n}$, denote $M^{\sim T} := T^{-1}MT$.

**Observation**

A matrix $M$ commutes with a matrix $P$ if and only if $P^{\sim T}$ commutes with $M^{\sim T}$, and $MP = M$ if and only if $M^{\sim T}P^{\sim T} = M^{\sim T}$ if and only if $MTP^{\sim T} = MT$. 


Characterizing equivariance

Proposition

There is a one-to-one correspondence between the irreducible components of $E_{\sigma, n \times n}$ and the integer solution vectors $\mathbf{r} = (r_l, m)$ of

$$\sum_{l \geq 1} \sum_{m \in (\mathbb{Z}/l\mathbb{Z})^r} r_l, m = \mathbf{r},$$

where $0 \leq r_l, m \leq d_l$. $d_l$ the dimension of the eigenspace of $P_\sigma$ of the eigenvalue $\zeta_k = e^{2\pi i / l}$

Properties of $E_{\sigma, n \times n}$

$$\dim \left( E_{\sigma, n \times n} (\mathbb{C}) \right) = \max_{\mathbf{r} = (r_l, m)} \left\{ \sum_{l \geq 1} \sum_{m \in (\mathbb{Z}/l\mathbb{Z})^r} (2d_k - r_l, m) \cdot r_l, m \right\},$$

$$\deg \left( E_{\sigma, n \times n} (\mathbb{C}) \right) = \prod_{l \geq 1} \prod_{m \in (\mathbb{Z}/l\mathbb{Z})^r} \prod_{i=0}^{d_k - r_l, m - 1} \frac{(d_k + i)! \cdot i!}{(r_l, m + i)! \cdot (d_k - r_l, m + i)!},$$

Sing $\left( E_{\sigma, n \times n} (\mathbb{K}) \right) = E_{\sigma, n-1 \times n} (\mathbb{K})$. $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$