The Geometry of Linear Convolutional Networks

Kathlén Kohn

joint work with

Thomas Merkh
UCLA

Guido Montúfar
MPI MiS Leipzig & UCLA

Matthew Trager
Amazon Alexa AI, NYC

Linear Convolutional Network (LCN)

with 1D convolutions

= family of functions $\mathbb{R}^{d_0} \longrightarrow \mathbb{R}^{d_L}$

$$x \longmapsto Wx, \qquad \text{where } W = W_L \cdots W_1$$

and W_i = **convolutional matrix** in the i-th layer

$$= \begin{bmatrix} w_{i,0} & \cdots & w_{i,s_i} & \cdots & w_{i,k_i-1} & & & \\ & & w_{i,0} & & \cdots & & w_{i,k_i-1} & \\ & & & \ddots & & & \ddots & \\ & & & & w_{i,0} & & \cdots & w_{i,k_i-1} \end{bmatrix} \in \mathbb{R}^{d_i \times d_{i-1}}.$$

filter w_i of size k_i

stride s_i

Linear Convolutional Network (LCN)

with 1D convolutions

= family of functions $\mathbb{R}^{d_0} \longrightarrow \mathbb{R}^{d_L}$

$$x \longmapsto Wx, \qquad \text{where } W = W_L \cdots W_1$$

and W_i = **convolutional matrix** in the i-th layer

$$= \begin{bmatrix} w_{i,0} & \cdots & w_{i,s_i} & \cdots & w_{i,k_i-1} & & & & \\ & & w_{i,0} & & \cdots & & w_{i,k_i-1} & & \\ & & & \ddots & & & & \ddots & \\ & & & & & w_{i,0} & & \cdots & w_{i,k_i-1} \end{bmatrix} \in \mathbb{R}^{d_i \times d_{i-1}}.$$

Example: $d_i = 3, k_i = 3, s_i = 2 : W_i = \begin{bmatrix} w_{i,0} & w_{i,1} & w_{i,2} & & & & \\ & & w_{i,0} & w_{i,1} & w_{i,2} & & \\ & & & & w_{i,0} & w_{i,1} & w_{i,2} \end{bmatrix}$

Linear Convolutional Network (LCN)

with 1D convolutions

= family of functions $\mathbb{R}^{d_0} \longrightarrow \mathbb{R}^{d_L}$

$$x \longmapsto Wx, \quad \text{where } W = W_L \cdots W_1$$

and W_i = **convolutional matrix** in the i-th layer

$$= \begin{bmatrix} w_{i,0} & \cdots & w_{i,s_i} & \cdots & w_{i,k_i-1} & & & & \\ & & w_{i,0} & & \cdots & & w_{i,k_i-1} & & \\ & & & \ddots & & & & \ddots & \\ & & & & & w_{i,0} & & \cdots & w_{i,k_i-1} \end{bmatrix} \in \mathbb{R}^{d_i \times d_{i-1}}.$$

Example: $d_i = 3, k_i = 3, s_i = 2 : W_i = \begin{bmatrix} w_{i,0} & w_{i,1} & w_{i,2} & & & & \\ & & w_{i,0} & w_{i,1} & w_{i,2} & & \\ & & & & w_{i,0} & w_{i,1} & w_{i,2} \end{bmatrix}$

The **LCN architecture** is $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$
where $\boldsymbol{d} = (d_0, \ldots, d_L)$, $\boldsymbol{k} = (k_1, \ldots, k_L)$, $\boldsymbol{s} = (s_1, \ldots, s_L)$.

I The geometry of the function space

II Optimization

III Summary / Comparison to fully-connected networks

Expressivity

The **function space** of an LCN is

$$\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \left\{ W \in \mathbb{R}^{d_L \times d_0} \colon W = \prod_{i=1}^{L} W_i,\ W_i \in \mathbb{R}^{d_i \times d_{i-1}} \text{ convolutional} \right\}.$$

What is the impact of the architecture on the geometry of the function space?

Expressivity

The **function space** of an LCN is

$$\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \left\{ W \in \mathbb{R}^{d_L \times d_0} : \ W = \prod_{i=1}^{L} W_i, \ W_i \in \mathbb{R}^{d_i \times d_{i-1}} \text{ convolutional} \right\}.$$

What is the impact of the architecture on the geometry of the function space?

Obs.: W is a convolutional matrix with filter size $k = k_1 + \sum_{i=2}^{L}(k_i - 1)\prod_{m=1}^{i-1} s_m$ and stride $s = \prod_{i=1}^{L} s_i$.

Cor.: $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \subseteq \mathcal{M}_{(d_0,d_L),k,s}$

Expressivity

The **function space** of an LCN is

$$\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \left\{ W \in \mathbb{R}^{d_L \times d_0} : W = \prod_{i=1}^{L} W_i, \; W_i \in \mathbb{R}^{d_i \times d_{i-1}} \text{ convolutional} \right\}.$$

What is the impact of the architecture on the geometry of the function space?

Obs.: W is a convolutional matrix with filter size $k = k_1 + \sum_{i=2}^{L} (k_i - 1) \prod_{m=1}^{i-1} s_m$ and stride $s = \prod_{i=1}^{L} s_i$.

Cor.: $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \subseteq \mathcal{M}_{(d_0,d_L),k,s}$ ← vector space

An architecture $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ is **filling** if $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \mathcal{M}_{(d_0,d_L),k,s}$.

Expressivity

The **function space** of an LCN is

$$\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \left\{ W \in \mathbb{R}^{d_L \times d_0} \colon W = \prod_{i=1}^{L} W_i,\ W_i \in \mathbb{R}^{d_i \times d_{i-1}} \text{ convolutional} \right\}.$$

What is the impact of the architecture on the geometry of the function space?

Obs.: W is a convolutional matrix with filter size $k = k_1 + \sum_{i=2}^{L}(k_i - 1)\prod_{m=1}^{i-1} s_m$ and stride $s = \prod_{i=1}^{L} s_i$.

Cor.: $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \subseteq \mathcal{M}_{(d_0,d_L),k,s}$

An architecture $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ is **filling** if $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \mathcal{M}_{(d_0,d_L),k,s}$.

When does this happen?

Stride 1

$$\boldsymbol{s} = (1, \dots, 1)$$

We identify convolutional matrices with polynomials:

$$\begin{bmatrix} w_0 & w_1 & \cdots & w_{k-1} & & & \\ & w_0 & w_1 & \cdots & w_{k-1} & & \\ & & \ddots & & & \ddots & \\ & & & w_0 & w_1 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi}{\longmapsto}} w_0 \mathrm{x}^{k-1} + w_1 \mathrm{x}^{k-2}\mathrm{y} + \cdots + w_{k-1}\mathrm{y}^{k-1} \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_{k-1}$$

homogeneous polynomials of degree k-1

Note: $\pi(W_L \cdots W_1) = \pi(W_L) \cdots \pi(W_1)$.

Stride 1

$$\boldsymbol{s} = (1, \dots, 1)$$

We identify convolutional matrices with polynomials:

$$\begin{bmatrix} w_0 & w_1 & \cdots & w_{k-1} & & & \\ & w_0 & w_1 & \cdots & w_{k-1} & & \\ & & \ddots & & & \ddots & \\ & & & w_0 & w_1 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi}{\longmapsto}} w_0 x^{k-1} + w_1 x^{k-2} y + \cdots + w_{k-1} y^{k-1} \in \mathbb{R}[x, y]_{k-1}$$

Note: $\pi(W_L \cdots W_1) = \pi(W_L) \cdots \pi(W_1)$.

Theorem:

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \Leftrightarrow \pi(W)$ has at least $e := |\{k_i : k_i \text{ is even}\}|$ real roots (counting multiplicities)

Stride 1

$$\boldsymbol{s} = (1, \dots, 1)$$

We identify convolutional matrices with polynomials:

$$\begin{bmatrix} w_0 & w_1 & \cdots & w_{k-1} & & & \\ & w_0 & w_1 & \cdots & w_{k-1} & & \\ & & \ddots & & & \ddots & \\ & & & w_0 & w_1 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi}{\longmapsto}} w_0 x^{k-1} + w_1 x^{k-2} y + \cdots + w_{k-1} y^{k-1} \in \mathbb{R}[x, y]_{k-1}$$

Note: $\pi(W_L \cdots W_1) = \pi(W_L) \cdots \pi(W_1)$.

Theorem:

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \Leftrightarrow \pi(W)$ has at least $e := |\{k_i : k_i \text{ is even}\}|$ real roots (counting multiplicities)

2) $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a full-dimensional, semialgebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$.

↑ finite union of solution sets to finitely many polynomial equations and inequalities

Stride 1

$\boldsymbol{s} = (1, \ldots, 1)$

We identify convolutional matrices with polynomials:

$$\begin{bmatrix} w_0 & w_1 & \cdots & w_{k-1} & & & \\ & w_0 & w_1 & \cdots & w_{k-1} & & \\ & & \ddots & & & \ddots & \\ & & & w_0 & w_1 & \cdots & w_{k-1} \end{bmatrix} \underset{\pi}{\overset{\sim}{\longmapsto}} w_0 x^{k-1} + w_1 x^{k-2} y + \cdots + w_{k-1} y^{k-1} \in \mathbb{R}[x,y]_{k-1}$$

Note: $\pi(W_L \cdots W_1) = \pi(W_L) \cdots \pi(W_1)$.

Theorem:

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \Leftrightarrow \pi(W)$ has at least $e := |\{k_i : k_i \text{ is even}\}|$ real roots (counting multiplicities)
2) $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a full-dimensional, semialgebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$.
3) The architecture $(\boldsymbol{d}, \boldsymbol{k}, \boldsymbol{s})$ is filling (i.e., $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} = \mathcal{M}_{(d_0,d_L),k,s}$) $\Leftrightarrow$ $e \leq 1$.

2 even filter sizes

$$\boldsymbol{s} = (1, \dots, 1)$$

even

$$\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) \quad = \quad \{P \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_{k-1} : P \text{ has } \geq 2 \text{ real roots } \}$$

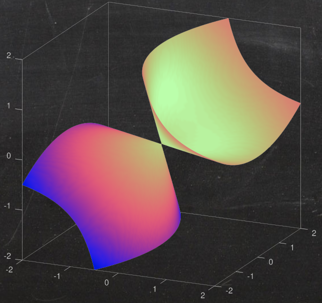
2 even filter sizes

$$\boldsymbol{s} = (1, \dots, 1)$$

$$\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \{P \in \mathbb{R}[x,y]_{k-1} : P \text{ has } \geq 2 \text{ real roots } \}$$

$$\mathbb{R}[x,y]_{k-1} \setminus \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \{P \in \mathbb{R}[x,y]_{k-1} : P \text{ has no real roots } \} = \{\text{positive polynomials}\} \cup \{\text{negative polynomials}\}$$

convex cones



The boundary of the function space

$s = (1, \ldots, 1)$

$P \in \mathbb{R}[x, y]_{k-1}$ has **real root multiplicity pattern**, short **rrmp**, $(\rho \mid \gamma) = (\rho_1, \ldots, \rho_r \mid \gamma_1, \ldots, \gamma_c)$ if it can be written as

multiplicities

$$P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad \text{where}$$

$p_i \in \mathbb{R}[x, y]_1$ and $q_j \in \mathbb{R}[x, y]_2$ are irreducible and pairwise linearly independent.

real roots

complex roots

The boundary of the function space

$s = (1, \dots, 1)$

$P \in \mathbb{R}[x, y]_{k-1}$ has **real root multiplicity pattern**, short **rrmp**, $(\rho \mid \gamma) = (\rho_1, \dots, \rho_r \mid \gamma_1, \dots, \gamma_c)$ if it can be written as

$$P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \qquad \text{where}$$

$p_i \in \mathbb{R}[x, y]_1$ and $q_j \in \mathbb{R}[x, y]_2$ are irreducible and pairwise linearly independent.

Theorem: Let $e := |\{k_i : k_i \text{ is even}\}| \geq 2$.

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \Leftrightarrow \pi(W)$ has rrmp $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$

The boundary of the function space

$\boldsymbol{s} = (1, \ldots, 1)$

$P \in \mathbb{R}[x,y]_{k-1}$ has **real root multiplicity pattern**, short **rrmp**, $(\rho \mid \gamma) = (\rho_1, \ldots, \rho_r \mid \gamma_1, \ldots, \gamma_c)$ if it can be written as

$$P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \qquad \text{where}$$

$p_i \in \mathbb{R}[x,y]_1$ and $q_j \in \mathbb{R}[x,y]_2$ are irreducible and pairwise linearly independent.

Theorem: Let $e := |\{k_i : k_i \text{ is even}\}| \geq 2$.

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \Leftrightarrow \pi(W)$ has rrmp $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$
2) $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is closed in Euclidean topology.

The boundary of the function space

$\boldsymbol{s} = (1, \dots, 1)$

$P \in \mathbb{R}[x, y]_{k-1}$ has **real root multiplicity pattern**, short **rrmp**, $(\rho \mid \gamma) = (\rho_1, \dots, \rho_r \mid \gamma_1, \dots, \gamma_c)$ if it can be written as

$$P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \qquad \text{where}$$

$p_i \in \mathbb{R}[x, y]_1$ and $q_j \in \mathbb{R}[x, y]_2$ are irreducible and pairwise linearly independent.

Theorem: Let $e := |\{k_i : k_i \text{ is even}\}| \geq 2$.

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \;\Leftrightarrow\; \pi(W)$ has rrmp $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$
2) $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is closed in Euclidean topology.
3) $W \in \partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \;\Leftrightarrow\; \pi(W)$ has $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$ & $|\{\rho_i : \rho_i \text{ is odd}\}| \leq e - 2$

↑ Euclidean boundary

The boundary of the function space

$\boldsymbol{s} = (1, \ldots, 1)$

$P \in \mathbb{R}[x, y]_{k-1}$ has **real root multiplicity pattern**, short **rrmp**, $(\rho \mid \gamma) = (\rho_1, \ldots, \rho_r \mid \gamma_1, \ldots, \gamma_c)$ if it can be written as

$$P = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \qquad \text{where}$$

$p_i \in \mathbb{R}[x, y]_1$ and $q_j \in \mathbb{R}[x, y]_2$ are irreducible and pairwise linearly independent.

Theorem: Let $e := |\{k_i : k_i \text{ is even}\}| \geq 2$.

1) $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \;\Leftrightarrow\; \pi(W)$ has rrmp $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$
2) $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is closed in Euclidean topology.
3) $W \in \partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \;\Leftrightarrow\; \pi(W)$ has $(\rho \mid \gamma)$ with $\sum \rho_i \geq e$ & $|\{\rho_i : \rho_i \text{ is odd}\}| \leq e - 2$
4) The Zariski closure of $\partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is the discriminant hypersurface.

↳ = {polynomials with (complex) double roots}

Example

$$\boldsymbol{k} = (2,2,2),\ \boldsymbol{s} = (1,1,1)$$

$$\begin{array}{ccccc} \begin{bmatrix} A & B & C & D \end{bmatrix} & = & \begin{bmatrix} a & b \end{bmatrix} & \begin{bmatrix} c & d & 0 \\ 0 & c & d \end{bmatrix} & \begin{bmatrix} e & f & 0 & 0 \\ 0 & e & f & 0 \\ 0 & 0 & e & f \end{bmatrix} \\ A\mathrm{x}^3 + B\mathrm{x}^2\mathrm{y} + C\mathrm{x}\mathrm{y}^2 + D\mathrm{y}^3 & = & (a\mathrm{x} + b\mathrm{y}) & (c\mathrm{x} + d\mathrm{y}) & (e\mathrm{x} + f\mathrm{y}) \end{array}$$

Example

$\boldsymbol{k} = (2,2,2), \ \boldsymbol{s} = (1,1,1)$

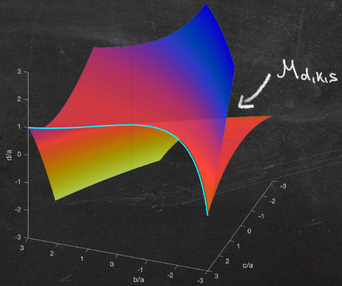
$$\begin{bmatrix} A & B & C & D \end{bmatrix} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} c & d & 0 \\ 0 & c & d \end{bmatrix} \begin{bmatrix} e & f & 0 & 0 \\ 0 & e & f & 0 \\ 0 & 0 & e & f \end{bmatrix}$$

$$Ax^3 + Bx^2y + Cxy^2 + Dy^3 = (ax+by) \quad (cx+dy) \quad (ex+fy)$$

Possible rrmp:

111	0
12	0
3	0
1	1

$\partial \mathcal{M}_{d,k,s}$ (rows 12, 3); $\mathcal{M}_{d,k,s}$ (rows 111, 12, 3)



Example

$$\boldsymbol{s} = (1, \ldots, 1)$$

$\boldsymbol{k} = (3,2,2)$	$\boldsymbol{k} = (4,2)$
$(ax^2 + bxy + cy^2) \cdot (dx + ey) \cdot (fx + gy)$	$(a'x^3 + b'x^2y + c'xy^2 + d'y^3) \cdot (e'x + f'y)$

$$= Ax^4 + Bx^3y + Cx^2y^2 + Dxy^3 + Ey^4$$

Both architectures have the same function space.

Example

$$\boldsymbol{s} = (1, \dots, 1)$$

$\boldsymbol{k} = (3, 2, 2)$	$\boldsymbol{k} = (4, 2)$
$(ax^2 + bxy + cy^2) \cdot (dx + ey) \cdot (fx + gy)$	$(a'x^3 + b'x^2y + c'xy^2 + d'y^3) \cdot (e'x + f'y)$

$$= Ax^4 + Bx^3y + Cx^2y^2 + Dxy^3 + Ey^4$$

Both architectures have the same function space.

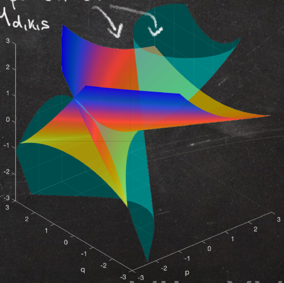
Possible rrmp:

1111	0
112	0
22	0
13	0
4	0

$\mathcal{M}_{d,k,s}$

11	1
2	1
0	11
0	2

$\partial \mathcal{M}_{d,k,s}$



Larger strides

$$\begin{bmatrix} w_0 \cdots w_s \cdots w_{k-1} & & & \\ & w_0 & \cdots & w_{k-1} & \\ & & \ddots & & \ddots & \\ & & & w_0 & \cdots & w_{k-1} \end{bmatrix} \xmapsto[\pi_s]{\sim} w_0 x^{s(k-1)} + w_1 x^{s(k-2)} y^s + \cdots + w_{k-1} y^{s(k-1)} \in \mathbb{R}[x^s, y^s]_{k-1}$$

Note: $\pi(W_2 W_1) = \pi_{s_1}(W_2)\pi(W_1)$.

Larger strides

$$\begin{bmatrix} w_0 \cdots w_s \cdots w_{k-1} & & & \\ & w_0 & \cdots & w_{k-1} & \\ & & \ddots & & \ddots \\ & & & w_0 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi_s}{\longmapsto}} w_0 x^{s(k-1)} + w_1 x^{s(k-2)} y^s + \cdots + w_{k-1} y^{s(k-1)} \in \mathbb{R}[x^s, y^s]_{k-1}$$

Note: $\pi(W_2 W_1) = \pi_{s_1}(W_2)\pi(W_1)$.

Theorem:
If $k_L > 1$ and $s_i > 1$ for some $i \leq L-1$, then $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a lower-dimensional semialgebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$. In particular, the architecture is not filling.

Larger strides

$$\begin{bmatrix} w_0 \cdots w_s \cdots w_{k-1} & & & \\ & w_0 & \cdots & w_{k-1} & \\ & & \ddots & & \ddots & \\ & & & w_0 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi_s}{\longmapsto}} w_0 x^{s(k-1)} + w_1 x^{s(k-2)} y^s + \cdots + w_{k-1} y^{s(k-1)} \in \mathbb{R}[x^s, y^s]_{k-1}$$

Note: $\pi(W_2 W_1) = \pi_{s_1}(W_2)\pi(W_1)$.

Theorem:
If $k_L > 1$ and $s_i > 1$ for some $i \leq L-1$, then $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a lower-dimensional semialgebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$. In particular, the architecture is not filling.

Lemma:

1) If $\boldsymbol{k} = (k_1, \ldots, k_{L-1}, 1)$ and $\boldsymbol{k}' = (k_1, \ldots, k_{L-1})$, then $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k}',\boldsymbol{s}})$.

Larger strides

$$\begin{bmatrix} w_0 \cdots w_s \cdots w_{k-1} & & & \\ & w_0 & \cdots & w_{k-1} & \\ & & \ddots & & \ddots \\ & & & w_0 & \cdots & w_{k-1} \end{bmatrix} \overset{\sim}{\underset{\pi_s}{\longmapsto}} w_0 x^{s(k-1)} + w_1 x^{s(k-2)} y^s + \cdots + w_{k-1} y^{s(k-1)} \in \mathbb{R}[x^s, y^s]_{k-1}$$

Note: $\pi(W_2 W_1) = \pi_{s_1}(W_2)\pi(W_1)$.

Theorem:
If $k_L > 1$ and $s_i > 1$ for some $i \leq L-1$, then $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$ is a lower-dimensional semialgebraic subset of $\mathcal{M}_{(d_0,d_L),k,s}$. In particular, the architecture is not filling.

Lemma:

1) If $\boldsymbol{k} = (k_1, \ldots, k_{L-1}, 1)$ and $\boldsymbol{k}' = (k_1, \ldots, k_{L-1})$, then $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k}',\boldsymbol{s}})$.
2) If $\boldsymbol{s} = (s_1, \ldots, s_L)$ and $\boldsymbol{s}' = (s_1, \ldots, s_{L-1}, 1)$, then $\pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}) = \pi(\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}'})$.

D-dimensional convolutions

stride 1

- input x: tensor of order D
- filter w: tensor of format $k^{(1)} \times \ldots \times k^{(D)}$
- **convolutional tensor** W of order $2D$

D-dimensional convolutions

stride 1

- input x: tensor of order D
- filter w: tensor of format $k^{(1)} \times \ldots \times k^{(D)}$
- **convolutional tensor** W of order $2D$ can be identified with a polynomial

$$\pi(W) \in \mathbb{R}[x_1, y_1, \ldots, x_D, y_D]_{(k^{(1)}-1,\ldots,k^{(D)}-1)}$$

  that is homogeneous of degree $k^{(j)} - 1$ in each pair x_j, y_j.

Note: $\pi(W_L \circ \cdots \circ W_1) = \pi(W_L) \cdots \pi(W_1)$.

D-dimensional convolutions

stride 1

- input x: tensor of order D
- filter w: tensor of format $k^{(1)} \times \ldots \times k^{(D)}$
- **convolutional tensor** W of order $2D$ can be identified with a polynomial

$$\pi(W) \in \mathbb{R}[\mathrm{x}_1, \mathrm{y}_1, \ldots, \mathrm{x}_D, \mathrm{y}_D]_{(k^{(1)}-1, \ldots, k^{(D)}-1)}$$

that is homogeneous of degree $k^{(j)} - 1$ in each pair $\mathrm{x}_j, \mathrm{y}_j$.

Note: $\pi(W_L \circ \cdots \circ W_1) = \pi(W_L) \cdots \pi(W_1)$.

↑ degree $k^{(j)}-1 := \sum_{i=1}^{L}(k_i^{(j)}-1)$ in x_j, y_j

↑ degree $k_L^{(j)}-1$ in x_j, y_j

↑ degree $k_1^{(j)}-1$ in x_j, y_j

D-dimensional convolutions

stride 1

- input x: tensor of order D
- filter w: tensor of format $k^{(1)} \times \ldots \times k^{(D)}$
- **convolutional tensor** W of order $2D$ can be identified with a polynomial

$$\pi(W) \in \mathbb{R}[x_1, y_1, \ldots, x_D, y_D]_{(k^{(1)}-1,\ldots,k^{(D)}-1)}$$

that is homogeneous of degree $k^{(j)} - 1$ in each pair x_j, y_j.

Note: $\pi(W_L \circ \cdots \circ W_1) = \pi(W_L) \cdots \pi(W_1)$.

Theorem:
Given an LCN with $D > 1$, $L > 1$ and non-trivial filter sizes, the function space is a lower-dimensional semialgebraic subset of $\pi^{-1}\mathbb{R}[x_1, y_1, \ldots, x_D, y_D]_{(k^{(1)}-1,\ldots,k^{(D)}-1)}$.
In particular, the architecture is not filling.

I The geometry of the function space

II Optimization

III Summary / Comparison to fully-connected networks

Critical points of the loss

Assume: 1D convolutions with stride 1

A **loss** of an LCN is a function $\mathcal{L} = \ell \circ \mu$ where

- $\mu : (W_1, \dots, W_L) \mapsto W = W_L \cdots W_1$ and
- ℓ is a smooth function on $\mathcal{M}_{(d_0,d_L),k,s}$.

Critical points of the loss

Assume: 1D convolutions with stride 1

A **loss** of an LCN is a function $\mathcal{L} = \ell \circ \mu$ where

- $\mu : (W_1, \dots, W_L) \mapsto W = W_L \cdots W_1$ and
- ℓ is a smooth function on $\mathcal{M}_{(d_0,d_L),k,s}$.

How are critical points in function space and critical points in parameter space related?

Critical points of the loss

Assume: 1D convolutions with stride 1

A **loss** of an LCN is a function $\mathcal{L} = \ell \circ \mu$ where

- $\mu : (W_1, \ldots, W_L) \mapsto W = W_L \cdots W_1$ and
- ℓ is a smooth function on $\mathcal{M}_{(d_0, d_L), k, s}$.

How are critical points in function space and critical points in parameter space related?

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

Critical points of the loss

Assume: 1D convolutions with stride 1

A **loss** of an LCN is a function $\mathcal{L} = \ell \circ \mu$ where

- $\mu : (W_1, \ldots, W_L) \mapsto W = W_L \cdots W_1$ and
- ℓ is a smooth function on $\mathcal{M}_{(d_0, d_L), k, s}$.

How are critical points in function space and critical points in parameter space related?

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in ?$

Fibers in parameter space

Scaling equivalence classes:

$$(W_1, \dots, W_L) \sim (W_1', \dots, W_L') \text{ if } \exists \alpha_1, \dots, \alpha_L \in \mathbb{R} : \alpha_1 \cdots \alpha_L = 1, W_i' = \alpha_i W_i$$

Proposition: Let $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \setminus \{0\}$. Then

1) $\mu^{-1}(W)$ consists of finitely many scaling equivalence classes.
2) Either all parameters in the same equivalence class are critical or none are.

Fibers in parameter space

Scaling equivalence classes:

$$(W_1, \ldots, W_L) \sim (W_1', \ldots, W_L') \text{ if } \exists \alpha_1, \ldots, \alpha_L \in \mathbb{R} : \alpha_1 \cdots \alpha_L = 1, W_i' = \alpha_i W_i$$

Proposition: Let $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \setminus \{0\}$. Then

1) $\mu^{-1}(W)$ consists of finitely many scaling equivalence classes.
2) Either all parameters in the same equivalence class are critical or none are.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c},$$
$$p_i \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_1, \; q_j \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_2$$

aggregate into factors

$$\pi(W_L) \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_{k_L - 1}, \ldots, \pi(W_1) \in \mathbb{R}[\mathrm{x}, \mathrm{y}]_{k_1 - 1}$$

Fibers in parameter space

Scaling equivalence classes:

$$(W_1, \dots, W_L) \sim (W_1', \dots, W_L') \text{ if } \exists \alpha_1, \dots, \alpha_L \in \mathbb{R} : \alpha_1 \cdots \alpha_L = 1, W_i' = \alpha_i W_i$$

Proposition: Let $W \in \mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}} \setminus \{0\}$. Then

1) $\mu^{-1}(W)$ consists of finitely many scaling equivalence classes.
2) Either all parameters in the same equivalence class are critical or none are.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad p_i \in \mathbb{R}[x,y]_1,\ q_j \in \mathbb{R}[x,y]_2$$

$\sum \rho_i$ balls of size 1,
$\sum \gamma_j$ balls of size 2,

aggregate into factors ↓ / ↓ place into

$\pi(W_L) \in \mathbb{R}[x,y]_{k_L-1}, \dots, \pi(W_1) \in \mathbb{R}[x,y]_{k_1-1}$ — bins of sizes $k_L - 1, \dots, k_1 - 1$

Numerical experiments

square loss & gradient descent

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in \,?$

Numerical experiments

square loss & gradient descent

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in ?$

$\boldsymbol{k} = (4, 2)$

target	%	initialization 1111|0 solution	%	mean loss
1111|0	5.28	1111|0	100	3.04e-15
11|1	72.6	112|0	15.5	0.228
		11|1	83.2	1.94e-15
		2|1	1.36	0.54
0|11	22.1	112|0	7.85	0.347
		2|1	92.2	0.231

interior of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
$\partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
complement of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

Numerical experiments

square loss & gradient descent

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in$?

$\boldsymbol{k} = (4, 2)$

		initialization 1111|0		
target	%	solution	%	mean loss
1111|0	5.28	1111|0	100	3.04e-15
11|1	72.6	112|0	15.5	0.228
		11|1	83.2	1.94e-15
		2|1	1.36	0.54
0|11	22.1	112|0	7.85	0.347
		2|1	92.2	0.231

interior of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
$\partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
complement of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

not in $\mathrm{Crit}(\ell|_{\mathcal{M}_{d,k,s}})$
i.e., critical point induced by parametrization μ

Numerical experiments

square loss & gradient descent

If $W \in \mathrm{Crit}(\ell|_{\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}})$, then $\mu^{-1}(W) \subset \mathrm{Crit}(\mathcal{L})$.

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in ?$

$\boldsymbol{k} = (4, 2)$

		initialization 1111|0		
target	%	solution	%	mean loss
1111|0	5.28	1111|0	100	3.04e-15
11|1	72.6	112|0	15.5	0.228
		11|1	83.2	1.94e-15
		2|1	1.36	0.54
0|11	22.1	112|0	7.85	0.347
		2|1	92.2	0.231

$\boldsymbol{k} = (3, 2, 2)$ ← same function space

		initialization 1111|0		
target	%	solution	%	mean loss
1111|0	4.82	1111|0	99.6	4.68e-15
		13|0	0.429	0.71
11|1	72.9	112|0	27.1	0.221
		22|0	1.28	0.992
		13|0	25.8	0.798
		11|1	45.5	1.78e-15
		2|1	0.381	0.446
0|11	22.3	112|0	11.2	0.374
		22|0	25.5	0.855
		13|0	7.1	0.895
		2|1	56.2	0.224

interior of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
$\partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$
complement of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

not in $\mathrm{Crit}(\ell|_{\mathcal{M}_{d,k,s}})$
i.e., critical point induced by parametrization μ

Critical points of the loss

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in \,?$

Critical points of the loss

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in$?

For $(\rho \mid \gamma)$ we define $\Delta_{(\rho|\gamma)} := \{W \in \mathcal{M}_{(d_0,d_L),k,s} : \pi(W) \text{ has rrmp } (\rho \mid \gamma)\}$.

Critical points of the loss

If $(W_1, \dots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in \,?$

For $(\rho \mid \gamma)$ we define $\Delta_{(\rho|\gamma)} := \{W \in \mathcal{M}_{(d_0,d_L),k,s} : \pi(W) \text{ has rrmp } (\rho \mid \gamma)\}$.

Theorem: Let $W = W_L \cdots W_1 \in \Delta_{(\rho|\gamma)}$.

1) If $(W_1, \dots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.

Critical points of the loss

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in \,?$

For $(\rho \mid \gamma)$ we define $\Delta_{(\rho|\gamma)} := \{W \in \mathcal{M}_{(d_0,d_L),k,s} : \pi(W) \text{ has rrmp } (\rho \mid \gamma)\}$.

Theorem: Let $W = W_L \cdots W_1 \in \Delta_{(\rho|\gamma)}$.

1) If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.

2) If $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$ and no $\pi(W_i)$ has a double root, then
$$(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L}).$$

Critical points of the loss

If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W = W_L \cdots W_1 \in$?

For $(\rho \mid \gamma)$ we define $\Delta_{(\rho|\gamma)} := \{W \in \mathcal{M}_{(d_0,d_L),k,s} : \pi(W) \text{ has rrmp } (\rho \mid \gamma)\}$.

Theorem: Let $W = W_L \cdots W_1 \in \Delta_{(\rho|\gamma)}$.

1) If $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$, then $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.

2) If $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$ and no $\pi(W_i)$ has a double root, then $(W_1, \ldots, W_L) \in \mathrm{Crit}(\mathcal{L})$.

3) For the square loss with generic training data, 2) becomes an "if and only if".

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad p_i \in \mathbb{R}[x,y]_1,\ q_j \in \mathbb{R}[x,y]_2$$

ρ_i balls of size 1 and color i,
γ_j balls of size 2 and color $-j$,

aggregate into factors ↓

↓ place into

$$\pi(W_L) \in \mathbb{R}[x,y]_{k_L-1}, \ldots, \pi(W_1) \in \mathbb{R}[x,y]_{k_1-1}$$

s.t. no $\pi(W_i)$ has a double root

bins of sizes $k_L - 1, \ldots, k_1 - 1$
s.t. no bin has 2 balls of the same color

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad p_i \in \mathbb{R}[x,y]_1,\ q_j \in \mathbb{R}[x,y]_2$$

$$\downarrow$$

$$\pi(W_L) \in \mathbb{R}[x,y]_{k_L-1}, \ldots, \pi(W_1) \in \mathbb{R}[x,y]_{k_1-1}$$

s.t. no $\pi(W_i)$ has a double root

ρ_i balls of size 1 and color i,

γ_j balls of size 2 and color $-j$,

$$\downarrow$$

bins of sizes $k_L - 1, \ldots, k_1 - 1$

s.t. no bin has 2 balls of the same color

We call such $(\rho \mid \gamma)$ **compatible** with the architecture.

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad p_i \in \mathbb{R}[x,y]_1,\ q_j \in \mathbb{R}[x,y]_2$$

$$\downarrow$$

$$\pi(W_L) \in \mathbb{R}[x,y]_{k_L-1}, \ldots, \pi(W_1) \in \mathbb{R}[x,y]_{k_1-1}$$

s.t. no $\pi(W_i)$ has a double root

ρ_i balls of size 1 and color i,
γ_j balls of size 2 and color $-j$,

$$\downarrow$$

bins of sizes $k_L - 1, \ldots, k_1 - 1$
s.t. no bin has 2 balls of the same color

We call such $(\rho \mid \gamma)$ **compatible** with the architecture.

3) For each compatible $(\rho \mid \gamma)$, determine all $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c}, \quad p_i \in \mathbb{R}[\mathrm{x},\mathrm{y}]_1,\ q_j \in \mathbb{R}[\mathrm{x},\mathrm{y}]_2$$

$\downarrow$

$\pi(W_L) \in \mathbb{R}[\mathrm{x},\mathrm{y}]_{k_L-1}, \ldots, \pi(W_1) \in \mathbb{R}[\mathrm{x},\mathrm{y}]_{k_1-1}$
s.t. no $\pi(W_i)$ has a double root

ρ_i balls of size 1 and color i,
γ_j balls of size 2 and color $-j$,

$\downarrow$

bins of sizes $k_L - 1, \ldots, k_1 - 1$
s.t. no bin has 2 balls of the same color

We call such $(\rho \mid \gamma)$ **compatible** with the architecture.

3) For each compatible $(\rho \mid \gamma)$, determine all $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.
4) For each W obtained in 3), find all compatible factorizations $W = W_L \cdots W_1$.

Finding all critical points in parameter space

for the square loss with generated data

1) List all rrmp $(\rho \mid \gamma)$ of polynomials of degree $k-1$.
2) For each $(\rho \mid \gamma)$, decide if polynomials with rrmp $(\rho \mid \gamma)$ can be factored according to the architecture such that no factor has a double root.

$$\pi(W) = p_1^{\rho_1} \cdots p_r^{\rho_r} q_1^{\gamma_1} \cdots q_c^{\gamma_c},$$
$$p_i \in \mathbb{R}[\mathrm{x},\mathrm{y}]_1,\ q_j \in \mathbb{R}[\mathrm{x},\mathrm{y}]_2$$

ρ_i balls of size 1 and color i,
γ_j balls of size 2 and color $-j$,

$$\downarrow \qquad\qquad \downarrow$$

$$\pi(W_L) \in \mathbb{R}[\mathrm{x},\mathrm{y}]_{k_L-1}, \ldots, \pi(W_1) \in \mathbb{R}[\mathrm{x},\mathrm{y}]_{k_1-1}$$
s.t. no $\pi(W_i)$ has a double root

bins of sizes $k_L - 1, \ldots, k_1 - 1$
s.t. no bin has 2 balls of the same color

We call such $(\rho \mid \gamma)$ **compatible** with the architecture.

3) For each compatible $(\rho \mid \gamma)$, determine all $W \in \mathrm{Crit}(\ell|_{\Delta_{(\rho|\gamma)}})$.
4) For each W obtained in 3), find all compatible factorizations $W = W_L \cdots W_1$.

In particular:
$\mathrm{Crit}(\mathcal{L})$ consists of finitely many scaling equivalence classes.

Example

Compatible architectures with factorizations:

$\rho\|\gamma$	1111|0	112|0	22|0	13|0	4|0	11|1	2|1	0|2	0|11
$\boldsymbol{k}=(3,2,2)$	$p_1p_2 \cdot p_3 \cdot p_4$	$p_1p_2 \cdot p_3 \cdot p_3$	$p_1p_2 \cdot p_1 \cdot p_2$	$p_1p_2 \cdot p_2 \cdot p_2$	—	$q_1 \cdot p_1 \cdot p_2$	$q_1 \cdot p_1 \cdot p_1$	—	—
$\boldsymbol{k}=(4,2)$	$p_1p_2p_3 \cdot p_4$	$p_1p_2p_3 \cdot p_3$	—	—	—	$p_1q_1 \cdot p_2$	$p_1q_1 \cdot p_1$	—	—

$\boldsymbol{k}=(4,2)$

		initialization 1111|0		
target	%	solution	%	mean loss
1111|0	5.28	1111|0	100	3.04e-15
11|1	72.6	112|0	15.5	0.228
		11|1	83.2	1.94e-15
		2|1	1.36	0.54
0|11	22.1	112|0	7.85	0.347
		2|1	92.2	0.231

$\boldsymbol{k}=(3,2,2)$

		initialization 1111|0		
target	%	solution	%	mean loss
1111|0	4.82	1111|0	99.6	4.68e-15
		13|0	0.429	0.71
11|1	72.9	112|0	27.1	0.221
		22|0	1.28	0.992
		13|0	25.8	0.798
		11|1	45.5	1.78e-15
		2|1	0.381	0.446
0|11	22.3	112|0	11.2	0.374
		22|0	25.5	0.855
		13|0	7.1	0.895
		2|1	56.2	0.224

interior of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

$\partial\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

complement of $\mathcal{M}_{\boldsymbol{d},\boldsymbol{k},\boldsymbol{s}}$

Invariants of gradient flow

If the initialization of gradient descent in known, there are only finitely many points in $\mathrm{Crit}(\mathcal{L})$ that gradient descent can reach.

Invariants of gradient flow

If the initialization of gradient descent in known, there are only finitely many points in $\mathrm{Crit}(\mathcal{L})$ that gradient descent can reach.

Theorem: Let $w_i \in \mathbb{R}^{k_i}$ be the filter of W_i. If $\omega(t) = (w_1(t), \ldots, w_L(t))$ is an integral curve for the negative gradient field of $\mathcal{L}$ (i.e., $\dot{\omega}(t) = -\nabla\mathcal{L}(\omega(t))$), then

$$\delta_{ij}(t) := \|w_i(t)\|^2 - \|w_j(t)\|^2 \quad \text{for } 1 \leq i, j \leq L$$

remain constant for all $t \in \mathbb{R}$.

Invariants of gradient flow

If the initialization of gradient descent in known, there are only finitely many points in $\mathrm{Crit}(\mathcal{L})$ that gradient descent can reach.

Theorem: Let $w_i \in \mathbb{R}^{k_i}$ be the filter of W_i. If $\omega(t) = (w_1(t), \ldots, w_L(t))$ is an integral curve for the negative gradient field of $\mathcal{L}$ (i.e., $\dot{\omega}(t) = -\nabla\mathcal{L}(\omega(t))$), then

$$\delta_{ij}(t) := \|w_i(t)\|^2 - \|w_j(t)\|^2 \quad \text{for } 1 \leq i,j \leq L$$

remain constant for all $t \in \mathbb{R}$.

known from initialization

Corollary: Let $\delta_{ij} \in \mathbb{R}$ be fixed for $1 \leq i,j \leq L$.
For any $(W_1, \ldots, W_L)$, there are only finitely many $(\alpha_1, \ldots, \alpha_L) \in \mathbb{R}^L$ such that $\alpha_1 \cdots \alpha_L = 1$ and the invariants of $(\alpha_1 W_1, \ldots, \alpha_L W_L)$ are the δ_{ij}.

I The geometry of the function space

II Optimization

III Summary / Comparison to fully-connected networks

fully-connected linear network

function space $\mathcal{M}$
$= \{$ rank-bounded matrices $\}$
$=$ algebraic variety
defined by polynomial equations

LCN

function space $\mathcal{M}$
$= \{$ polynomials with certain factorizations $\}$
$=$ semialgebraic set
defined by polynomial equations & inequalities

fully-connected linear network

function space $\mathcal{M}$
$= \{$ rank-bounded matrices $\}$
$=$ algebraic variety
defined by polynomial equations

If ℓ convex:
$\mathcal{L}$ has non-global minima
$\Leftrightarrow \ell|_{\mathcal{M}}$ has non-global minima.

LCN

function space $\mathcal{M}$
$= \{$ polynomials with certain factorizations $\}$
$=$ semialgebraic set
defined by polynomial equations & inequalities

$\mathcal{L}$ can have non-global minima
even if $\mathcal{M}$ is a vector space
and ℓ is convex.

fully-connected linear network

function space $\mathcal{M}$
$= \{$ rank-bounded matrices $\}$
$=$ algebraic variety
defined by polynomial equations

If ℓ convex:
$\mathcal{L}$ has non-global minima
$\Leftrightarrow \ell|_{\mathcal{M}}$ has non-global minima.

Reason:
critical points induced by the parametrization μ are always saddles

LCN

function space $\mathcal{M}$
$= \{$ polynomials with certain factorizations $\}$
$=$ semialgebraic set
defined by polynomial equations & inequalities

$\mathcal{L}$ can have non-global minima
even if $\mathcal{M}$ is a vector space
and ℓ is convex.

Reason:
critical points induced by the parametrization μ can be non-global minima

due to different structure of the fibers $\mu^{-1}(W)$