The geometry of neural networks

Kathlén Kohn

joint with

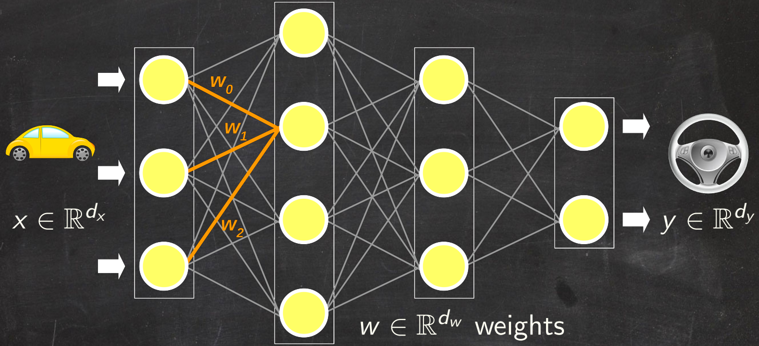
Joan Bruna
Center for Data Science
Courant Institute at NYU

Matthew Trager
Amazon Alexa AI, NYC

Guido Montúfar
MPI MiS Leipzig
UCLA

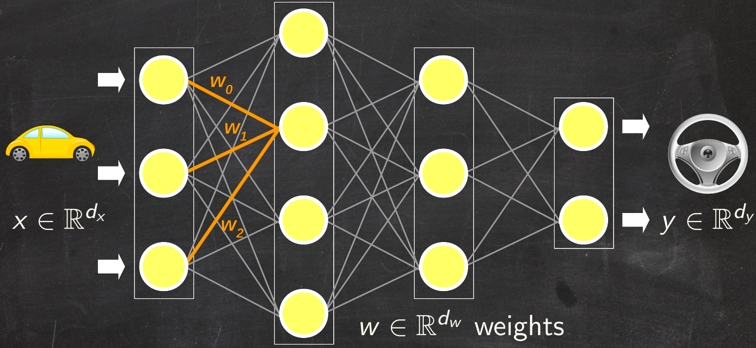
Thomas Merkh
UCLA

Neural Networks



A neural network is defined by a continuous mapping $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$.

Neural Networks

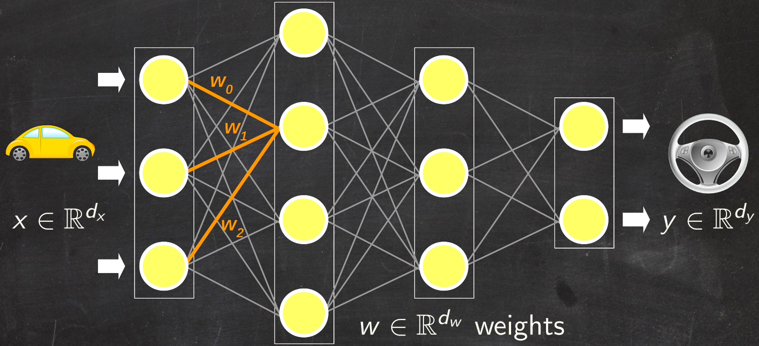


A neural network is defined by a continuous mapping $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$.

Definition $\mathcal{M}_\Phi := \left\{ \Phi(w, \cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y} \mid w \in \mathbb{R}^{d_w} \right\}$

is called the **neuromanifold** of Φ.

Neural Networks



A neural network is defined by a continuous mapping $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$.

Definition $\mathcal{M}_\Phi := \left\{ \Phi(w, \cdot) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y} \mid w \in \mathbb{R}^{d_w} \right\} \subset C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$

is called the **neuromanifold** of Φ.

Observation
1. Φ piecewise smooth $\Rightarrow \mathcal{M}_\Phi$ manifold with singularities
2. $\dim \mathcal{M}_\Phi \leq d_w$

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \dots W_1 x,$$

$$\text{where } w = (W_h, \dots, W_1) \text{ and } W_i \in \mathbb{R}^{d_i \times d_{i-1}},$$

(so $d_w = d_h d_{h-1} + \dots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \dots W_1 x,$$

$$\text{where } w = (W_h, \dots, W_1) \text{ and } W_i \in \mathbb{R}^{d_i \times d_{i-1}},$$

(so $d_w = d_h d_{h-1} + \dots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Example The neuromanifold of the linear network Φ is

$$\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq \min\{d_0, d_1, \dots, d_h\} \right\}.$$

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \ldots W_1 x,$$

$$\text{where } w = (W_h, \ldots, W_1) \text{ and } W_i \in \mathbb{R}^{d_i \times d_{i-1}},$$

(so $d_w = d_h d_{h-1} + \ldots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Example The neuromanifold of the linear network Φ is

$$\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq \underbrace{\min\{d_0, d_1, \ldots, d_h\}}_{=:r} \right\}.$$

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \dots W_1 x,$$

where $w = (W_h, \dots, W_1)$ and $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$,

(so $d_w = d_h d_{h-1} + \ldots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Example The neuromanifold of the linear network Φ is

$$\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq \underbrace{\min\{d_0, d_1, \dots, d_h\}}_{=:r} \right\}.$$

1. If $r = \min\{d_0, d_h\}$, then $\mathcal{M}_\Phi = \mathbb{R}^{d_h \times d_0}$. "filling architecture"

filling

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \dots W_1 x,$$

$$\text{where } w = (W_h, \dots, W_1) \text{ and } W_i \in \mathbb{R}^{d_i \times d_{i-1}},$$

(so $d_w = d_h d_{h-1} + \ldots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Example The neuromanifold of the linear network Φ is

$$\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq \underbrace{\min\{d_0, d_1, \dots, d_h\}}_{=:r} \right\}.$$

1. If $r = \min\{d_0, d_h\}$, then $\mathcal{M}_\Phi = \mathbb{R}^{d_h \times d_0}$. "filling architecture"
2. If $r < \min\{d_0, d_h\}$, "non-filling architecture"
   then $\mathcal{M}_\Phi$ is a **determinantal variety**.

filling

non-filling

Linear Networks

A **linear network** is defined by a map $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ of the form

$$\Phi(w, x) = W_h W_{h-1} \dots W_1 x,$$

$$\text{where } w = (W_h, \dots, W_1) \text{ and } W_i \in \mathbb{R}^{d_i \times d_{i-1}},$$

(so $d_w = d_h d_{h-1} + \ldots + d_1 d_0$, $d_x = d_0$ and $d_y = d_h$).

Example The neuromanifold of the linear network Φ is

$$\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq \underbrace{\min\{d_0, d_1, \dots, d_h\}}_{=:r} \right\}.$$

1. If $r = \min\{d_0, d_h\}$, then $\mathcal{M}_\Phi = \mathbb{R}^{d_h \times d_0}$. "filling architecture"
2. If $r < \min\{d_0, d_h\}$, "non-filling architecture"
   then $\mathcal{M}_\Phi$ is a **determinantal variety**.
   Note: $\mathcal{M}_\Phi$ is neither convex nor smooth $(\mathrm{Sing}\, \mathcal{M}_\Phi = \{M \mid \mathrm{rk}(M) \leq r - 1\})$

filling

non-filling

Loss Landscapes

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

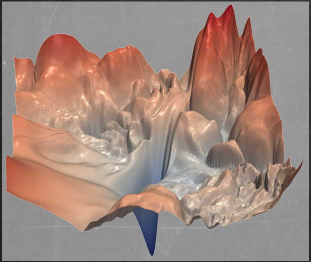
where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.

Loss Landscapes

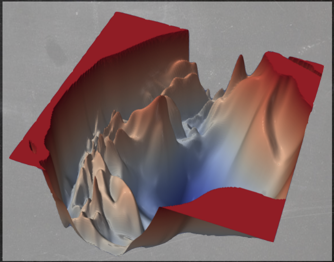
A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Visualizations of L



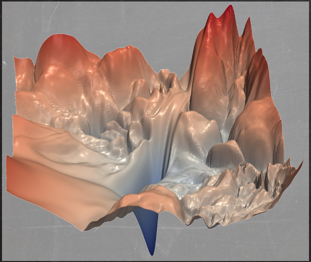
Source: Li, Hao, et al. "Visualizing the loss landscape of neural nets." Advances in Neural Information Processing Systems. 2018.

Loss Landscapes

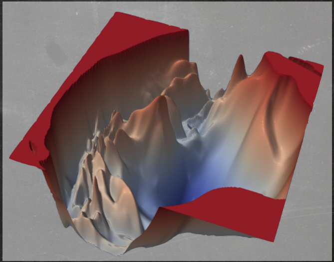
A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad\mu\quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Visualizations of L



Source: Li, Hao, et al. "Visualizing the loss landscape of neural nets." Advances in Neural Information Processing Systems. 2018.

Observation If $\varphi \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$, then $\mu^{-1}(\varphi) \subset \mathrm{Crit}(L)$.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

Recall: $\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq r \right\}$, where $r := \min\{d_0, d_1, \ldots, d_h\}$.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\mu} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\ell} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

Recall: $\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq r \right\}$, where $r := \min \left\{ d_0, d_1, \ldots, d_h \right\}$.

We characterize the connectivity of critical points for general losses:

critical points in $\mathcal{M}_\Phi$

critical M for $\ell|_{\mathcal{M}_\Phi}$
(e.g., global minimum)

$\longleftrightarrow$

critical points in $\mathbb{R}^{d_w}$

many *disconnected*
critical $W \in \mu^{-1}(M)$

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \quad \subset \quad \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

Recall: $\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \mathrm{rk}(M) \leq r \right\}$, where $r := \min \{d_0, d_1, \ldots, d_h\}$.

We characterize the connectivity of critical points for general losses:

Theorem Let $M \in \mathcal{M}_\Phi$.

1. If $\mathrm{rk}(M) = r$, then $\mu^{-1}(M)$ has 2^b path-connected components

$$\text{where } b := \# \{ i \mid 0 < i < h, d_i = r \}.$$

critical points in $\mathcal{M}_\Phi$		**critical points in $\mathbb{R}^{d_w}$**
critical M for $\ell\vert_{\mathcal{M}_\Phi}$ (e.g., global minimum)	$\longleftrightarrow$	many *disconnected* critical $W \in \mu^{-1}(M)$


Linear Networks

A loss function on a linear network is of the form

$$L: \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\mu} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\ell} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

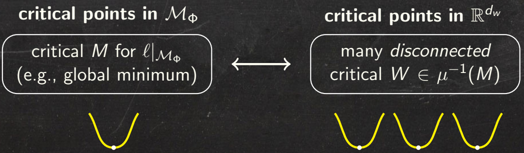
Recall: $\mathcal{M}_\Phi = \left\{ M \in \mathbb{R}^{d_h \times d_0} \mid \operatorname{rk}(M) \leq r \right\}$, where $r := \min \{d_0, d_1, \ldots, d_h\}$.

We characterize the connectivity of critical points for general losses:

Theorem Let $M \in \mathcal{M}_\Phi$.

1. If $\operatorname{rk}(M) = r$, then $\mu^{-1}(M)$ has 2^b path-connected components

$$\text{where } b := \# \{ i \mid 0 < i < h, d_i = r \}.$$



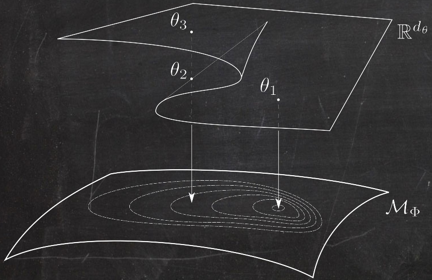
2. If $\operatorname{rk}(M) < r$, then $\mu^{-1}(M)$ is path-connected.

Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.

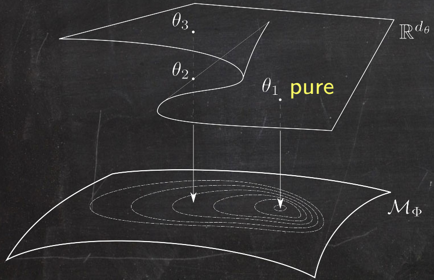


Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad\mu\quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



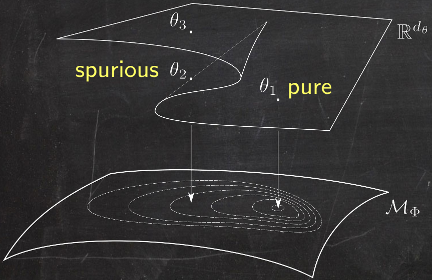
Definition
$w^* \in \mathrm{Crit}(L)$ is called **pure** if $\mu(w^*) \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$ and $\mu(w^*) \notin \mathrm{Sing}\,\mathcal{M}_\Phi$.

Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Definition

$w^* \in \mathrm{Crit}(L)$ is called **pure** if $\mu(w^*) \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$ and $\mu(w^*) \notin \mathrm{Sing}\,\mathcal{M}_\Phi$.

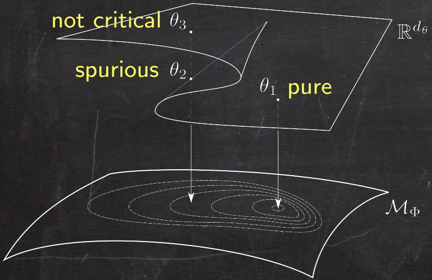
Otherwise $w^* \in \mathrm{Crit}(L)$ is called **spurious**.

Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Definition

$w^* \in \mathrm{Crit}(L)$ is called **pure** if $\mu(w^*) \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$ and $\mu(w^*) \notin \mathrm{Sing}\,\mathcal{M}_\Phi$.

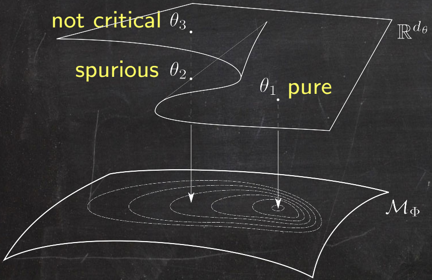
Otherwise $w^* \in \mathrm{Crit}(L)$ is called **spurious**.

Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{aligned} L : \mathbb{R}^{d_w} &\xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \xrightarrow{\quad \ell|_{\mathcal{M}_\Phi} \quad} \mathbb{R}, \\ w &\longmapsto \Phi(w, \cdot) \end{aligned}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Definition
$w^* \in \mathrm{Crit}(L)$ is called **pure** if $\mu(w^*) \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$ and $\mu(w^*) \notin \mathrm{Sing}\, \mathcal{M}_\Phi$.
Otherwise $w^* \in \mathrm{Crit}(L)$ is called **spurious**.

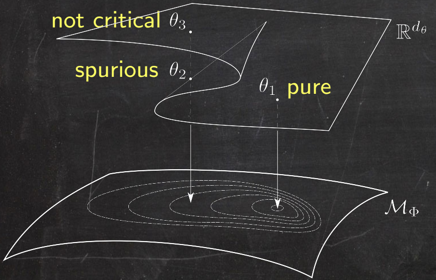
Proposition If the differential $D_{w^*}\mu$ at $w^* \in \mathrm{Crit}(L)$ has maximal rank (i.e., $\mathrm{rk}(D_{w^*}\mu) = \dim \mathcal{M}_\Phi$), then w^* is pure,

Pure & Spurious Critical Points

A **loss function** on a neural network $\Phi : \mathbb{R}^{d_w} \times \mathbb{R}^{d_x} \longrightarrow \mathbb{R}^{d_y}$ is of the form

$$\begin{array}{rcccl} L : \mathbb{R}^{d_w} & \xrightarrow{\quad\mu\quad} & \mathcal{M}_\Phi & \xrightarrow{\quad\ell|_{\mathcal{M}_\Phi}\quad} & \mathbb{R}, \\ w & \longmapsto & \Phi(w,\cdot) & & \end{array}$$

where ℓ is a functional defined on a subset of $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ containing $\mathcal{M}_\Phi$.



Definition
$w^* \in \mathrm{Crit}(L)$ is called **pure** if $\mu(w^*) \in \mathrm{Crit}(\ell|_{\mathcal{M}_\Phi})$ and $\mu(w^*) \notin \mathrm{Sing}\,\mathcal{M}_\Phi$.
Otherwise $w^* \in \mathrm{Crit}(L)$ is called **spurious**.

Proposition If the differential $D_{w^*}\mu$ at $w^* \in \mathrm{Crit}(L)$ has maximal rank (i.e., $\mathrm{rk}(D_{w^*}\mu) = \dim \mathcal{M}_\Phi$), then w^* is pure, and

w^* is a minimum for L (resp. saddle/maximum) $\Leftrightarrow$ $\mu(w^*)$ is a minimum for $\ell|_{\mathcal{M}_\Phi}$ (resp. saddle/maximum)

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \quad \subset \quad \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

For linear networks, the loss L often has "no bad minima",
i.e. every local minimum is global.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad\mu\quad} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad\ell\quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

For linear networks, the loss L often has "no bad minima",
i.e. every local minimum is global.

Proposition Let ℓ be smooth and convex.
L has non-global minima $\Leftrightarrow$ $\ell|_{\mathcal{M}_\Phi}$ has non-global minima.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

For linear networks, the loss L often has "no bad minima",
i.e. every local minimum is global.

Proposition Let ℓ be smooth and convex.
L has non-global minima $\Leftrightarrow$ $\ell|_{\mathcal{M}_\Phi}$ has non-global minima.

Corollary [Laurent & von Brecht '17]
If ℓ is smooth convex and $r = \min\{d_0, d_h\}$ (filling architecture),
then all local minima for L are global.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

For linear networks, the loss L often has "no bad minima",
i.e. every local minimum is global.

Proposition Let ℓ be smooth and convex.
L has non-global minima $\Leftrightarrow$ $\ell|_{\mathcal{M}_\Phi}$ has non-global minima.

Corollary [Laurent & von Brecht '17]
If ℓ is smooth convex and $r = \min\{d_0, d_h\}$ (filling architecture),
then all local minima for L are global.

Corollary [Baldi & Hornik '89, Kawaguchi '16]
If ℓ is a quadratic loss, then all local minima for L are global.

Linear Networks

A loss function on a linear network is of the form

$$L : \mathbb{R}^{d_h \times d_{h-1}} \times \ldots \times \mathbb{R}^{d_1 \times d_0} \xrightarrow{\quad \mu \quad} \mathcal{M}_\Phi \quad \subset \mathbb{R}^{d_h \times d_0} \xrightarrow{\quad \ell \quad} \mathbb{R},$$
$$(W_h, \ldots, W_1) \longmapsto W_h \cdots W_1$$

For linear networks, the loss L often has "no bad minima",
i.e. every local minimum is global.

Proposition Let ℓ be smooth and convex.
L has non-global minima $\Leftrightarrow \ell|_{\mathcal{M}_\Phi}$ has non-global minima.

Corollary [Laurent & von Brecht '17]
If ℓ is smooth convex and $r = \min\{d_0, d_h\}$ (filling architecture),
then all local minima for L are global.

Corollary [Baldi & Hornik '89, Kawaguchi '16]
If ℓ is a quadratic loss, then all local minima for L are global.
(even in the non-filling case!)

The Quadratic Loss

Fixed data matrices $X \in \mathbb{R}^{d_0 \times s}$ and $Y \in \mathbb{R}^{d_h \times s}$ define a **quadratic loss**

$$\begin{aligned} \ell_{X,Y} : \mathbb{R}^{d_h \times d_0} &\longrightarrow \mathbb{R}, \\ M &\longmapsto \|MX - Y\|_F^2 \end{aligned}$$

The Quadratic Loss

Fixed data matrices $X \in \mathbb{R}^{d_0 \times s}$ and $Y \in \mathbb{R}^{d_h \times s}$ define a **quadratic loss**

$$\begin{aligned} \ell_{X,Y} : \mathbb{R}^{d_h \times d_0} &\longrightarrow \mathbb{R}, \\ M &\longmapsto \|MX - Y\|_F^2 \end{aligned}$$

Observation If $XX^T = \mathrm{I}_{d_0}$ ("whitened data"), then

$$\ell_{X,Y}(M) = \|M - YX^T\|_F^2 + \text{const.}$$

The Quadratic Loss

Fixed data matrices $X \in \mathbb{R}^{d_0 \times s}$ and $Y \in \mathbb{R}^{d_h \times s}$ define a **quadratic loss**

$$\begin{aligned} \ell_{X,Y} : \mathbb{R}^{d_h \times d_0} &\longrightarrow \mathbb{R}, \\ M &\longmapsto \|MX - Y\|_F^2 \end{aligned}$$

Observation If $XX^T = \mathrm{I}_{d_0}$ ("whitened data"), then

$$\ell_{X,Y}(M) = \|M - YX^T\|_F^2 + \mathrm{const.}$$

Minimizing $\ell_{X,Y}$ on the determinantal variety $\mathcal{M}_\Phi = \{M \mid \mathrm{rk}(M) \leq r\}$ is equivalent to minimizing the Euclidean distance of YX^T to $\mathcal{M}_\Phi$.

Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

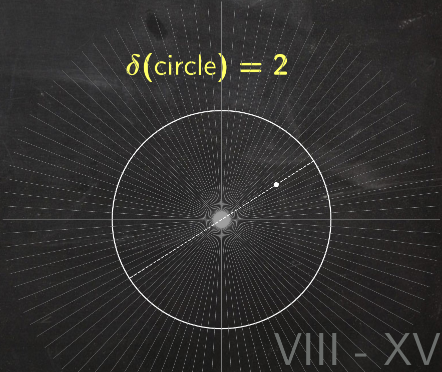
δ is called the **ED degree** of $\mathcal{Z}$.

Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

δ is called the **ED degree** of $\mathcal{Z}$.



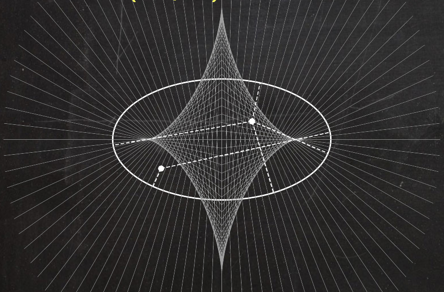
Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

δ is called the **ED degree** of $\mathcal{Z}$.

δ(ellipse) = 4



δ(circle) = 2

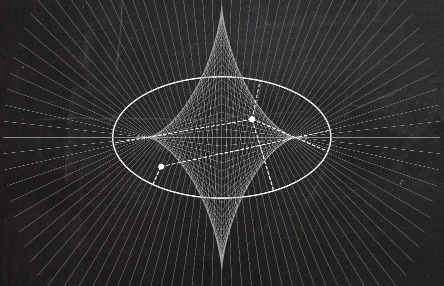
Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min\limits_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.
δ is called the **ED degree** of $\mathcal{Z}$.

The other $q \in \mathbb{R}^N$ form a complex hypersurface, called **ED discriminant** of $\mathcal{Z}$.

$\delta(\text{ellipse}) = 4$



$\delta(\text{circle}) = 2$

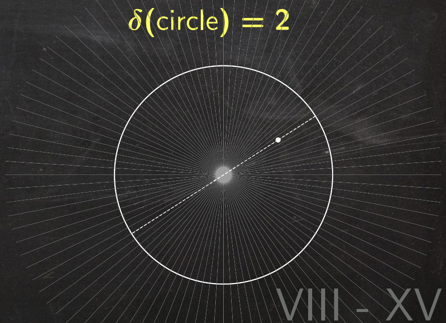
Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

δ is called the **ED degree** of $\mathcal{Z}$.

The other $q \in \mathbb{R}^N$ form a complex hypersurface, called **ED discriminant** of $\mathcal{Z}$.



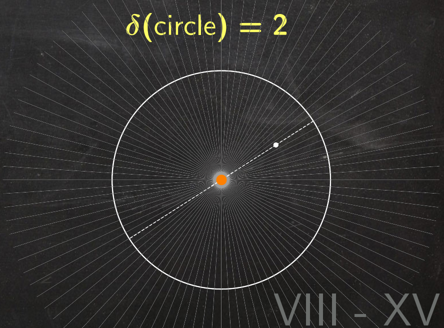
Euclidean Distance to Varieties

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety
(i.e., the common zero locus of some set of polynomials).

There is a constant $\delta \in \mathbb{Z}_{>0}$ such that for almost all $q \in \mathbb{R}^N$ the minimization problem $\min_{z \in \mathcal{Z}} \|z - q\|_2^2$ has δ complex critical points.

δ is called the **ED degree** of $\mathcal{Z}$.

The other $q \in \mathbb{R}^N$ form a complex hypersurface, called **ED discriminant** of $\mathcal{Z}$.



Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ **determinantal variety**

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m \times n}$ be of full rank with pairwise distinct singular values.

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m \times n}$ be of full rank with pairwise distinct singular values.

1. $\min_{M \in \mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m \times n}$ be of full rank with pairwise distinct singular values.

1. $\min\limits_{M \in \mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.

   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m\times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m\times n}$ be of full rank with pairwise distinct singular values.

1. $\min\limits_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.

   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$
2. All critical points are real.

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m\times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m\times n}$ be of full rank with pairwise distinct singular values.

1. $\min_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.
   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$
2. All critical points are real.
   $\Rightarrow$ ED discriminant has codimension 2 over $\mathbb{R}$

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m\times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m\times n}$ be of full rank with pairwise distinct singular values.

1. $\min\limits_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.
   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$
2. All critical points are real.
   $\Rightarrow$ ED discriminant has codimension 2 over $\mathbb{R}$
   In fact: ED discriminant $=$ $\{$ matrices with ≥ 2 coinciding singular values $\}$

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m\times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m\times n}$ be of full rank with pairwise distinct singular values.

1. $\min_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.
   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$
2. All critical points are real.
   $\Rightarrow$ ED discriminant has codimension 2 over $\mathbb{R}$
   In fact: ED discriminant $=$ { matrices with ≥ 2 coinciding singular values }
3. $\min_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has unique local minimum

Eckart-Young Theorem

$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m\times n}$ **determinantal variety**

EY Theorem

Let $Q \in \mathbb{R}^{m\times n}$ be of full rank with pairwise distinct singular values.

1. $\min_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has $\binom{\min\{m,n\}}{r}$ complex critical points.

   $\Rightarrow$ ED degree $\delta(\mathcal{M}_r) = \binom{\min\{m,n\}}{r}$
2. All critical points are real.

   $\Rightarrow$ ED discriminant has codimension 2 over $\mathbb{R}$

   In fact: ED discriminant $=$ { matrices with ≥ 2 coinciding singular values }
3. $\min_{M\in\mathcal{M}_r} \|M - Q\|_F^2$ has unique local minimum

Corollary [Baldi & Hornik '89, Kawaguchi '16]

If ℓ is a quadratic loss, then all local minima for the loss $L = \ell \circ \mu$ on a linear network are global. (even in the non-filling case!)

Linear Networks Can Have Bad Local Minima

Linear Networks Can Have Bad Local Minima

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety.

There is a constant $\delta^{\mathrm{gen}} \in \mathbb{Z}_{>0}$ such that for almost all linear coordinate changes $f : \mathbb{R}^N \to \mathbb{R}^N$ the ED degree of $f(\mathcal{Z})$ is δ^{gen}.

Linear Networks Can Have Bad Local Minima

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety.

There is a constant $\delta^{\mathrm{gen}} \in \mathbb{Z}_{>0}$ such that for almost all linear coordinate changes $f : \mathbb{R}^N \to \mathbb{R}^N$ the ED degree of $f(\mathcal{Z})$ is δ^{gen}.

δ^{gen} is called the **generic ED degree** of $\mathcal{Z}$.

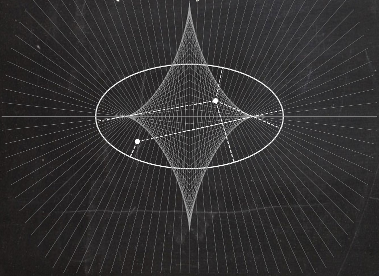
Linear Networks Can Have Bad Local Minima

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety.

There is a constant $\delta^{\text{gen}} \in \mathbb{Z}_{>0}$ such that for almost all linear coordinate changes $f : \mathbb{R}^N \to \mathbb{R}^N$ the ED degree of $f(\mathcal{Z})$ is δ^{gen}.

δ^{gen} is called the **generic ED degree** of $\mathcal{Z}$.

$\boldsymbol{\delta}$**(ellipse) = 4**



$\delta^{\text{gen}}(\text{circle})$
$= \delta(\text{ellipse})$
$= 4$

$\boldsymbol{\delta}$**(circle) = 2**

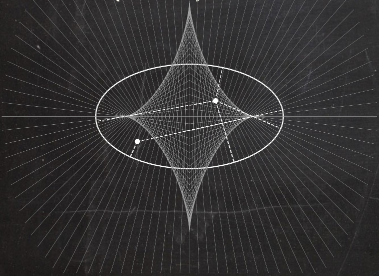
Linear Networks Can Have Bad Local Minima

Let $\mathcal{Z} \subset \mathbb{R}^N$ be an algebraic variety.

There is a constant $\delta^{\text{gen}} \in \mathbb{Z}_{>0}$ such that for almost all linear coordinate changes $f : \mathbb{R}^N \to \mathbb{R}^N$ the ED degree of $f(\mathcal{Z})$ is δ^{gen}.

δ^{gen} is called the **generic ED degree** of $\mathcal{Z}$.

$\boldsymbol{\delta}(\textbf{ellipse}) = \mathbf{4}$



$\delta^{\text{gen}}(\text{circle})$
$= \delta(\text{ellipse})$
$= 4$

$\boldsymbol{\delta}(\textbf{circle}) = \mathbf{2}$

Equivalently: δ^{gen} is the ED degree of $\mathcal{Z}$
under the perturbed Euclidean distance $\|f(\cdot)\|_2$.

Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

1. $\delta(\mathcal{M}_1) = 3$

Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

1. $\delta(\mathcal{M}_1) = 3 \quad < \quad 39 = \delta^{\mathrm{gen}}(\mathcal{M}_1)$

Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

1. $\delta(\mathcal{M}_1) = 3 \quad < \quad 39 = \delta^{\mathrm{gen}}(\mathcal{M}_1)$
2. under almost all perturbed Euclidean distances $\|f(\cdot)\|_2$, the ED discriminant of $\mathcal{M}_1$ is a hypersurface over $\mathbb{R}$

Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

1. $\delta(\mathcal{M}_1) = 3 \quad < \quad 39 = \delta^{\mathrm{gen}}(\mathcal{M}_1)$
2. under almost all perturbed Euclidean distances $\|f(\cdot)\|_2$, the ED discriminant of $\mathcal{M}_1$ is a hypersurface over $\mathbb{R}$

$\Rightarrow$ different number of real critical points in different open regions of $\mathbb{R}^{3\times 3}$

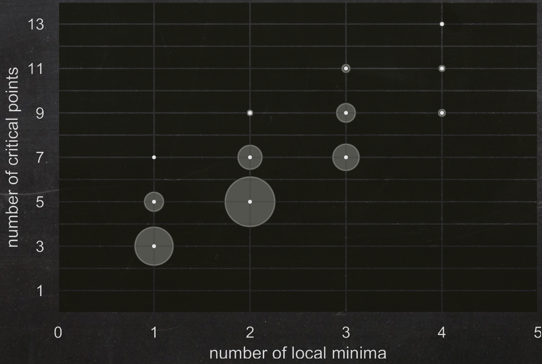
Linear Networks Can Have Bad Local Minima

Example $\mathcal{M}_1 = \{M \mid \mathrm{rk}(M) \leq 1\} \subset \mathbb{R}^{3\times 3}$

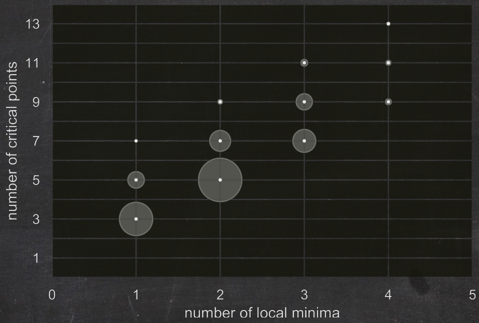
1. $\delta(\mathcal{M}_1) = 3 \quad < \quad 39 = \delta^{\mathrm{gen}}(\mathcal{M}_1)$
2. under almost all perturbed Euclidean distances $\|f(\cdot)\|_2$, the ED discriminant of $\mathcal{M}_1$ is a hypersurface over $\mathbb{R}$

$\Rightarrow$ different number of real critical points in different open regions of $\mathbb{R}^{3\times 3}$

3. Also: different number of local minima in different open regions of $\mathbb{R}^{3\times 3}$, not all of them global !

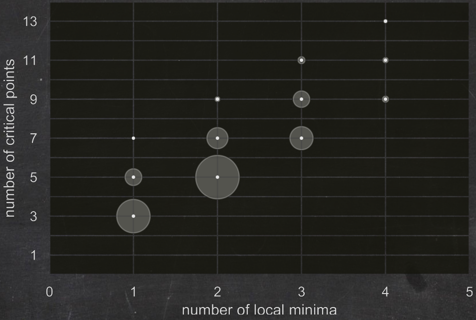


Linear Networks Can Have Bad Local Minima



		# real critical points						
		1	**3**	**5**	**7**	**9**	**11**	**13**
	1	0	476	120	1	0	0	0
# local minima	**2**	0	0	805	190	10	0	0
	3	0	0	0	228	116	21	0
	4	0	0	0	0	16	12	5

Linear Networks Can Have Bad Local Minima



		# real critical points						
		1	**3**	**5**	**7**	**9**	**11**	**13**
# local minima	**1**	0	476	120	1	0	0	0
	2	0	0	805	190	10	0	0
	3	0	0	0	228	116	21	0
	4	0	0	0	0	16	12	5

All determinantal varieties behave like this !

Linear Networks Can Have Bad Local Minima

Remark Closed formula for generic ED degree of
$\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ involving only m, n, r difficult to derive.

Linear Networks Can Have Bad Local Minima

Remark Closed formula for generic ED degree of $\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ involving only m, n, r difficult to derive.

For $r = 1$,

$$\delta^{\mathrm{gen}}(\mathcal{M}_1) = \sum_{s=0}^{m+n} (-1)^s (2^{m+n+1-s} - 1)(m+n-s)! \left[\sum_{\substack{i+j=s \\ i \leq m,\ j \leq n}} \frac{\binom{m+1}{i}\binom{n+1}{j}}{(m-i)!(n-j)!} \right]$$

Linear Networks Can Have Bad Local Minima

Remark Closed formula for generic ED degree of $\mathcal{M}_r = \{M \mid \mathrm{rk}(M) \leq r\} \subset \mathbb{R}^{m \times n}$ involving only m, n, r difficult to derive.

For $r = 1$,

$$\delta^{\mathrm{gen}}(\mathcal{M}_1) = \sum_{s=0}^{m+n} (-1)^s (2^{m+n+1-s} - 1)(m+n-s)! \left[\sum_{\substack{i+j=s \\ i \leq m,\ j \leq n}} \frac{\binom{m+1}{i}\binom{n+1}{j}}{(m-i)!(n-j)!} \right]$$

$$\delta(\mathcal{M}_1) = \min\{m, n\}$$

Take Away

- determinantal varieties are examples of neuromanifolds
- for linear networks with smooth convex losses:

	quadratic loss	other loss
filling	no bad min.	no bad min.
non-filling	no bad min.	**bad min.**

convex optimization on vector space

special embedding of determinantal varieties

- future extensions to
  - convolutional networks
    (ongoing work with T. Merkh, G. Montúfar, M. Trager)
  - networks with polynomial activation functions or
  - ReLU networks (using semi-algebraic sets)

Linear Convolutional Networks

- 1D convolutional layers with 1 filter having stride size 1 correspond to circulant matrices $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix}$

Linear Convolutional Networks

- 1D convolutional layers with 1 filter having stride size 1 correspond to circulant matrices $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix}$
- applying several layers (i.e. multiplying circulant matrices) corresponds to polynomial multiplication

Linear Convolutional Networks

- 1D convolutional layers with 1 filter having stride size 1 correspond to circulant matrices $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix}$
- applying several layers (i.e. multiplying circulant matrices) corresponds to polynomial multiplication

Theorem
A network architecture is filling (i.e. the neuromanifold is a vector space) if and only if there is at most 1 filter of even width.

Linear Convolutional Networks

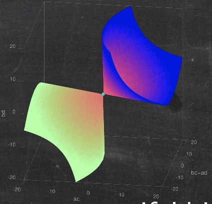
- 1D convolutional layers with 1 filter having stride size 1 correspond to circulant matrices $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix}$
- applying several layers (i.e. multiplying circulant matrices) corresponds to polynomial multiplication

Theorem
A network architecture is filling (i.e. the neuromanifold is a vector space) if and only if there is at most 1 filter of even width.

- In the non-filling case, the neuromanifold is a semi-algebraic set whose boundary is contained in the discriminant hypersurface of polynomials.

Linear Convolutional Networks

- 1D convolutional layers with 1 filter having stride size 1 correspond to circulant matrices $\begin{bmatrix} a & b & 0 \\ 0 & a & b \\ b & 0 & a \end{bmatrix}$
- applying several layers (i.e. multiplying circulant matrices) corresponds to polynomial multiplication



Theorem

A network architecture is filling (i.e. the neuromanifold is a vector space) if and only if there is at most 1 filter of even width.

- In the non-filling case, the neuromanifold is a semi-algebraic set whose boundary is contained in the discriminant hypersurface of polynomials.
- **Example:** If there are 2 filters of even width, the complement of the neuromanifold is a union of two convex cones.