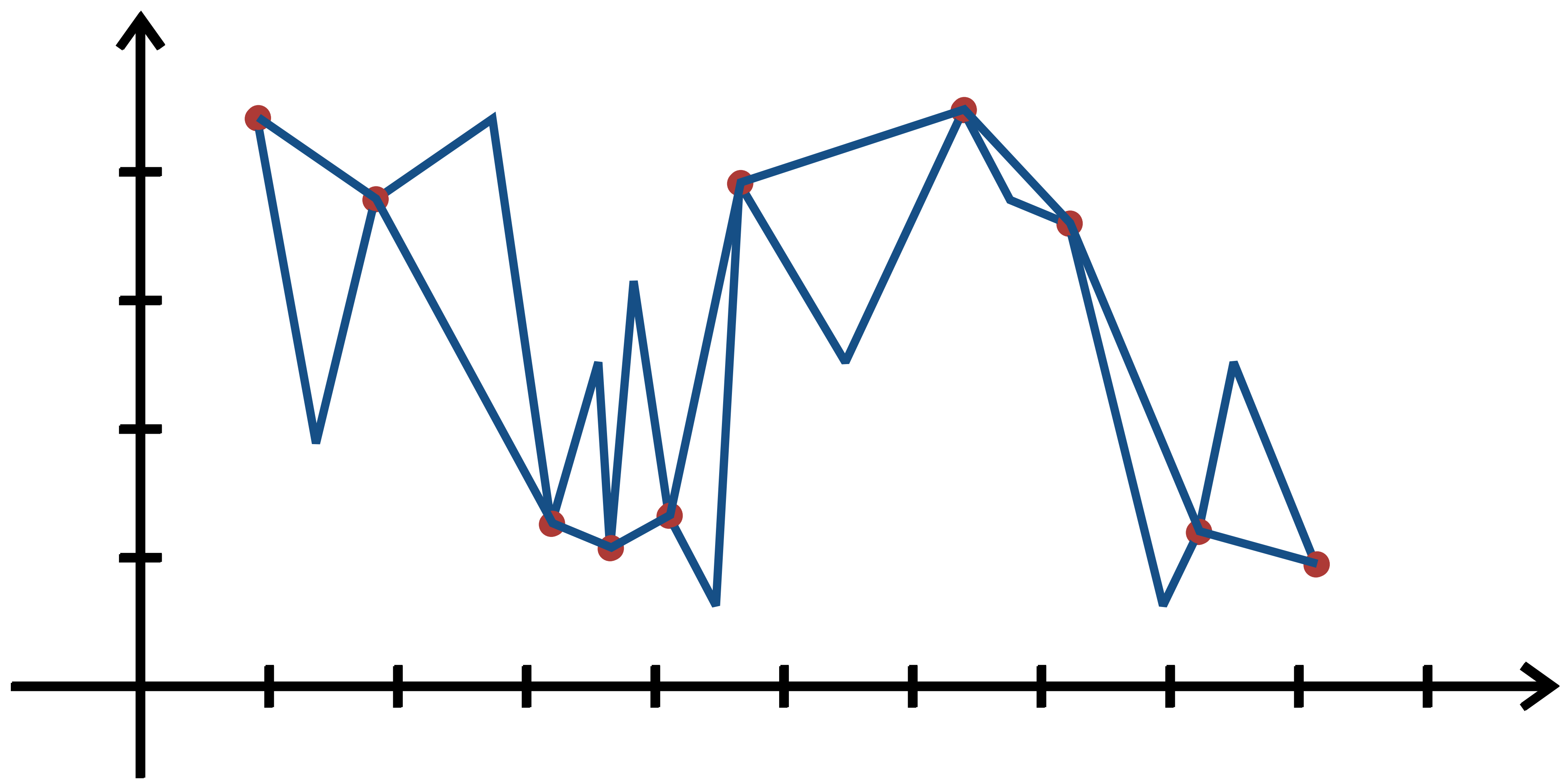
Linear Layers in ReLU Networks Promote Learning Single-/Multiple-Index Models

Suzanna Parkinson
University of Chicago

October 21, 2024
SIAM MDS 2024

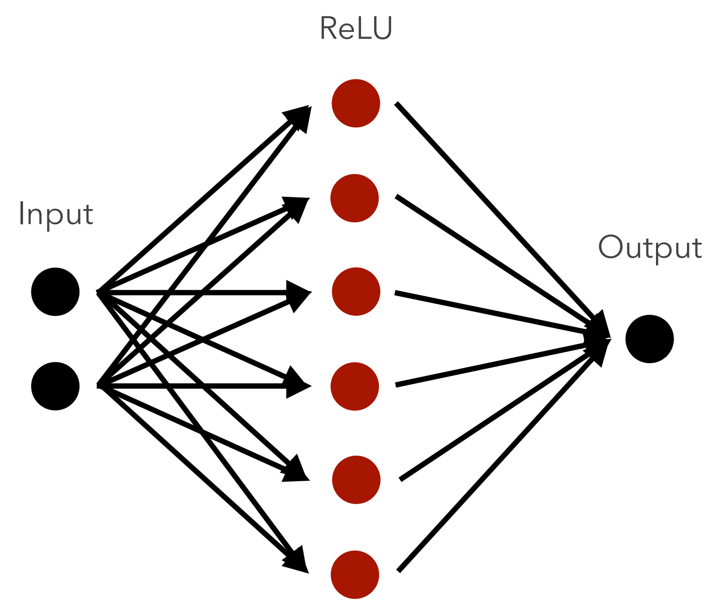
Motivation: Regularization in 1D Shallow ReLU Networks



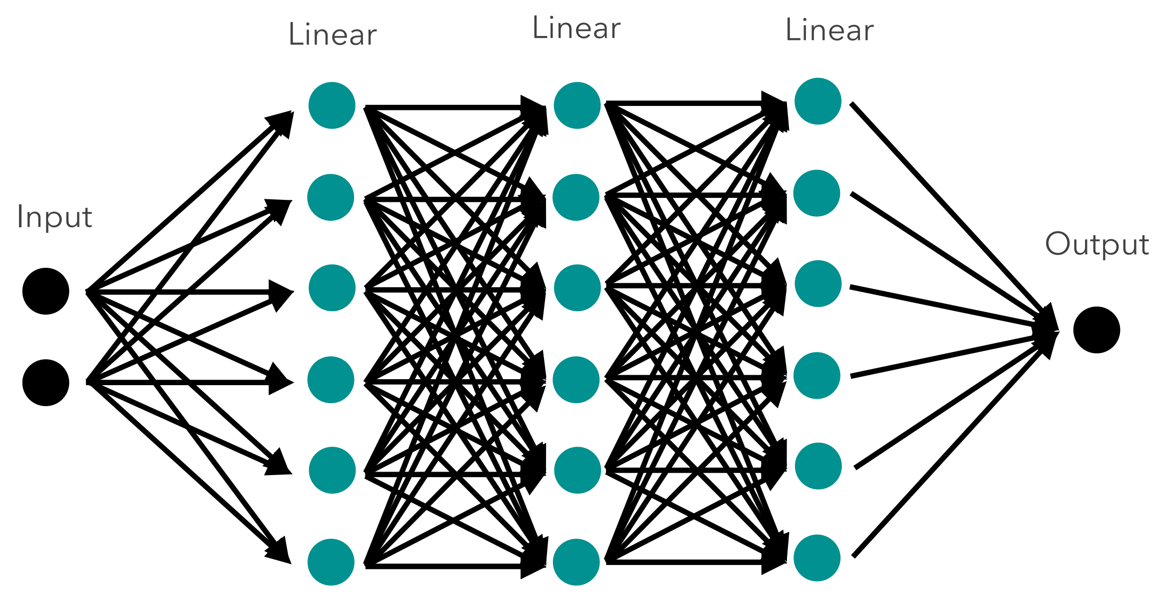
- Both functions…
  - Can be expressed as a **shallow ReLU** neural network
  - **Interpolate** the data
  - **Generalize** differently
- What functions are **preferred** by explicitly regularized neural networks?
- How do preferred functions change with network **architecture**?

Effect of weight decay regularization in neural networks

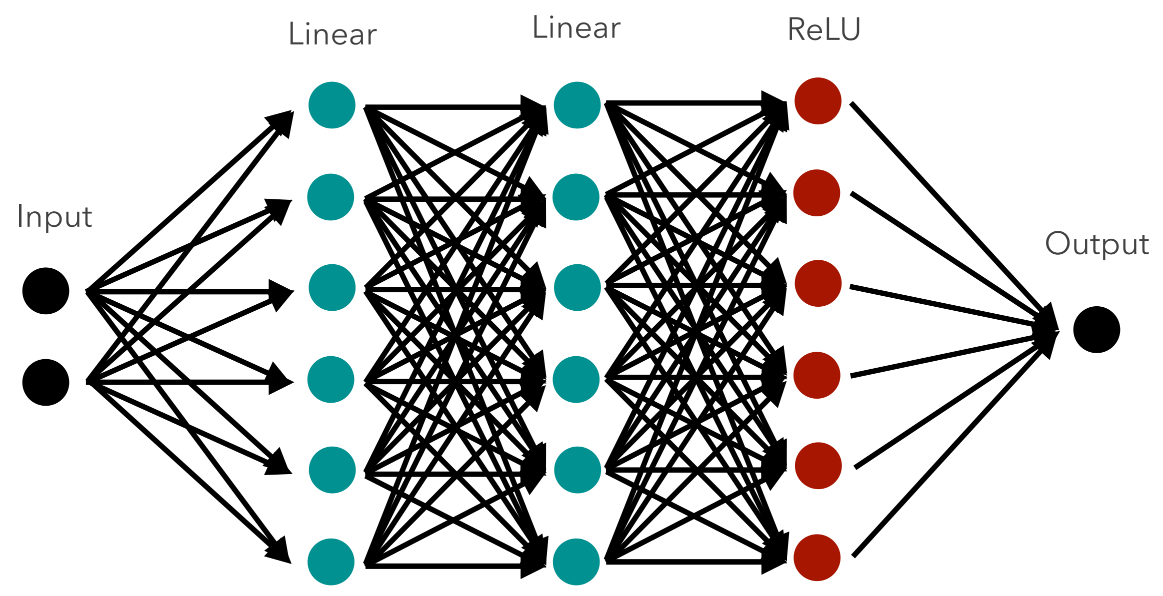
- 2-layer **ReLU** networks Bach (2017)
    - For $x \in \mathbb{R}$, prefer functions for which $\int |f''(x)|\,dx$ is small *Savarese et al. (2019), Joshi, Vadi, & Srebro (2023), Boursier & Flammarion (2023)*
    - For $x \in \mathbb{R}^d$, prefer functions for which $\|\mathscr{R}(-\Delta)^{(d+1)/2} f\|_{\mathrm{TV}}$ is small *Ongie et al. (2019)*
    - Banach space representer theorems & minimax rates *Parhi & Nowak (2021), Bartolucci et al. (2023), Unser (2023)*



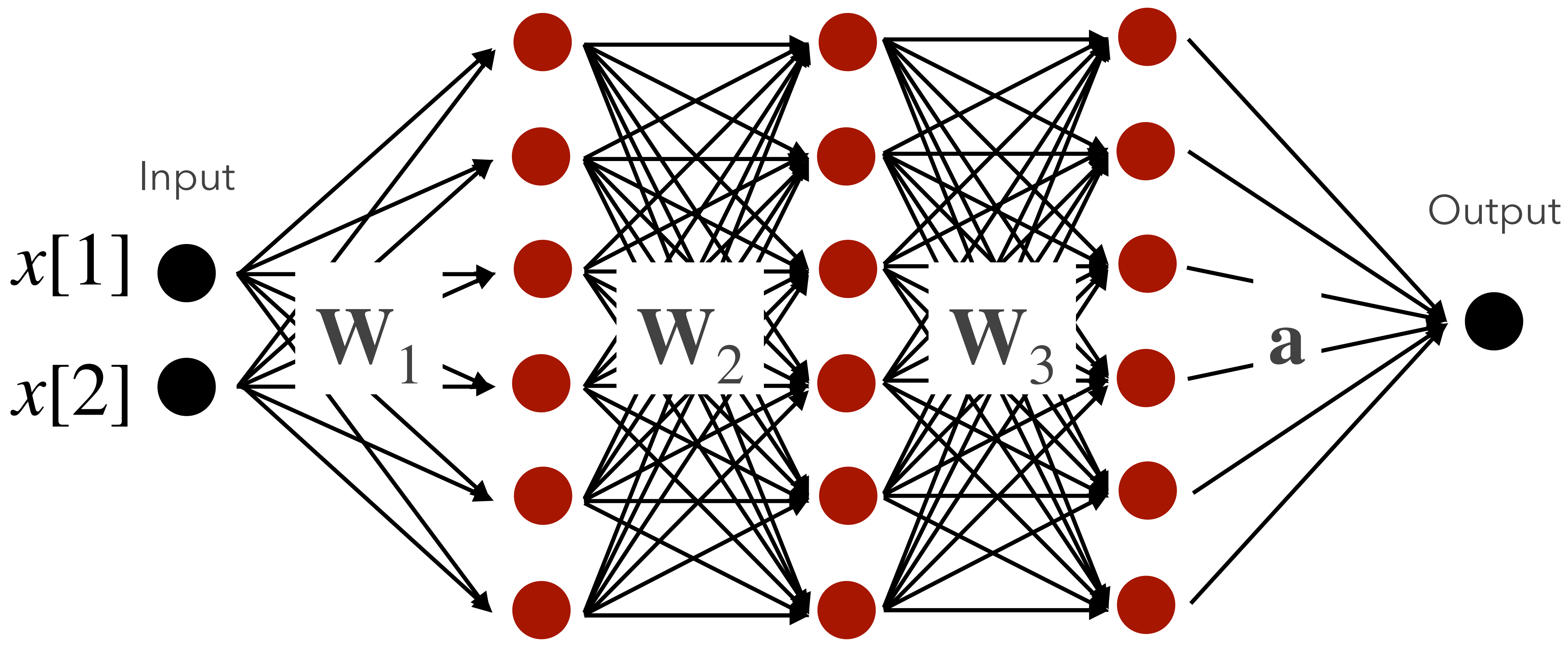
- **Multi-layer** linear networks
    - Gradient descent “aligns” layers *Ji & Telgarsky (2018)*
    - Depth induces ℓ^q and group norms depending on architecture *Dai, Karzand, & Srebro (2021)*
    - Depth promotes sparsity and low rank *Chou, Many, Rauhut (2011-2023)*



- **Multi-layer nonlinear** networks
    - Insights for rank-1 or orthonormal training data *Ergen & Pilanci (2021)*
    - Insights for low-rank vector-valued networks *Jacot (2023, 2024)*
    - **This work**



Neural Networks



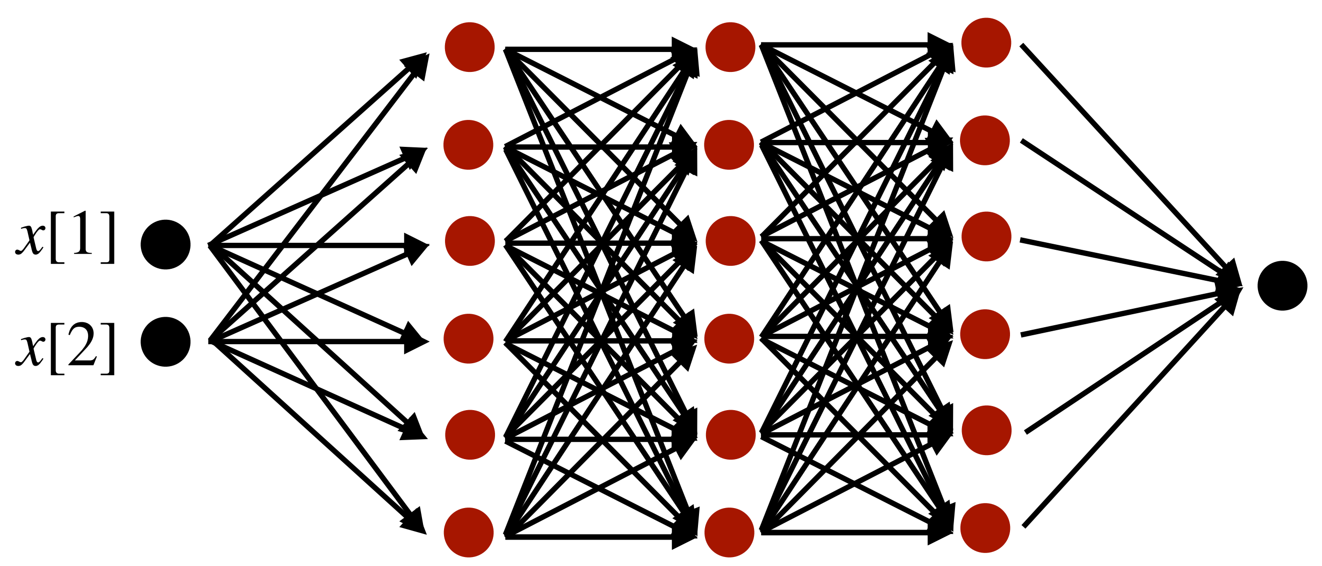
$$\theta = \left(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_{L-1}, \mathbf{a}\right)$$

$$f_\theta(\mathbf{x}) = \mathbf{a}^\top \sigma\left(\mathbf{W}_{L-1} \cdot \sigma\left(\cdots \sigma\left(\mathbf{W}_2 \sigma\left(\mathbf{W}_1 \mathbf{x}\right)\right)\right)\right)$$

Function Space Perspective

Parameter Space Cost

$$\hat{\theta}_S \in \arg\min_{\theta} \mathscr{L}_S(f_\theta) + \lambda C_L(\theta) \text{ where } C_L(\theta) = \frac{1}{L}\left(\sum_{\ell=1}^{L-1} \|\mathbf{W}_\ell\|_F^2 + \|\mathbf{a}\|_2^2\right)$$



$$\hat{f}_S \in \arg\min_{g \in \mathcal{N}_L} \mathscr{L}_S(g) + \lambda R_L(g) \text{ where } \underbrace{R_L(g) = \inf_{\theta} C_L(\theta) \text{ s.t. } f_\theta = g}_{\textbf{Representation Cost}}$$

What kinds of functions have **small representation cost**?

How does the representation cost depend on network architecture, including **depth**?

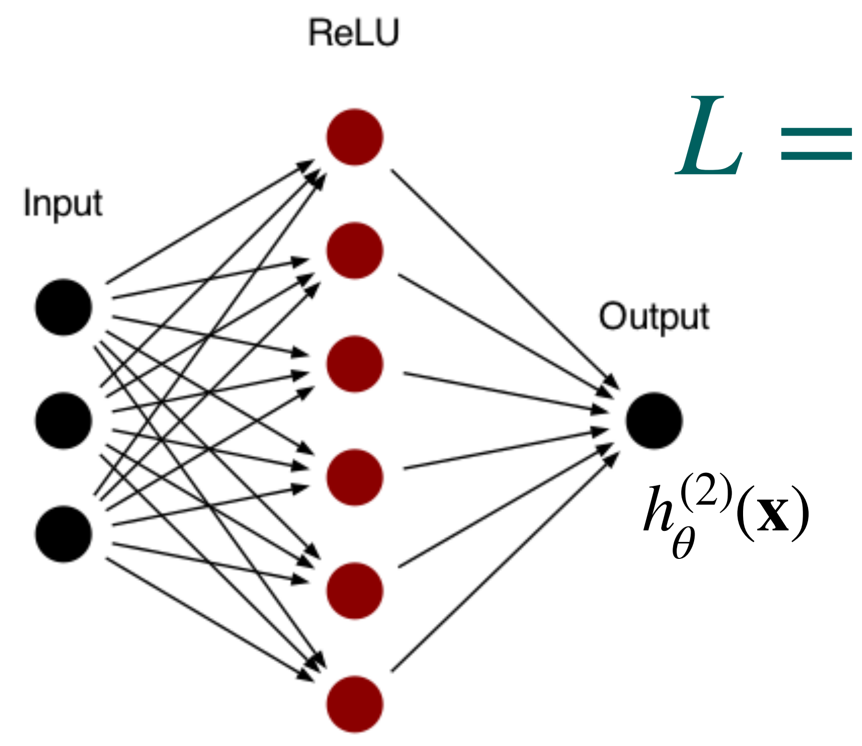
Linear layers in ReLU NNs
promotes learning
single-/multi–index models

Linear layers in **ReLU** networks

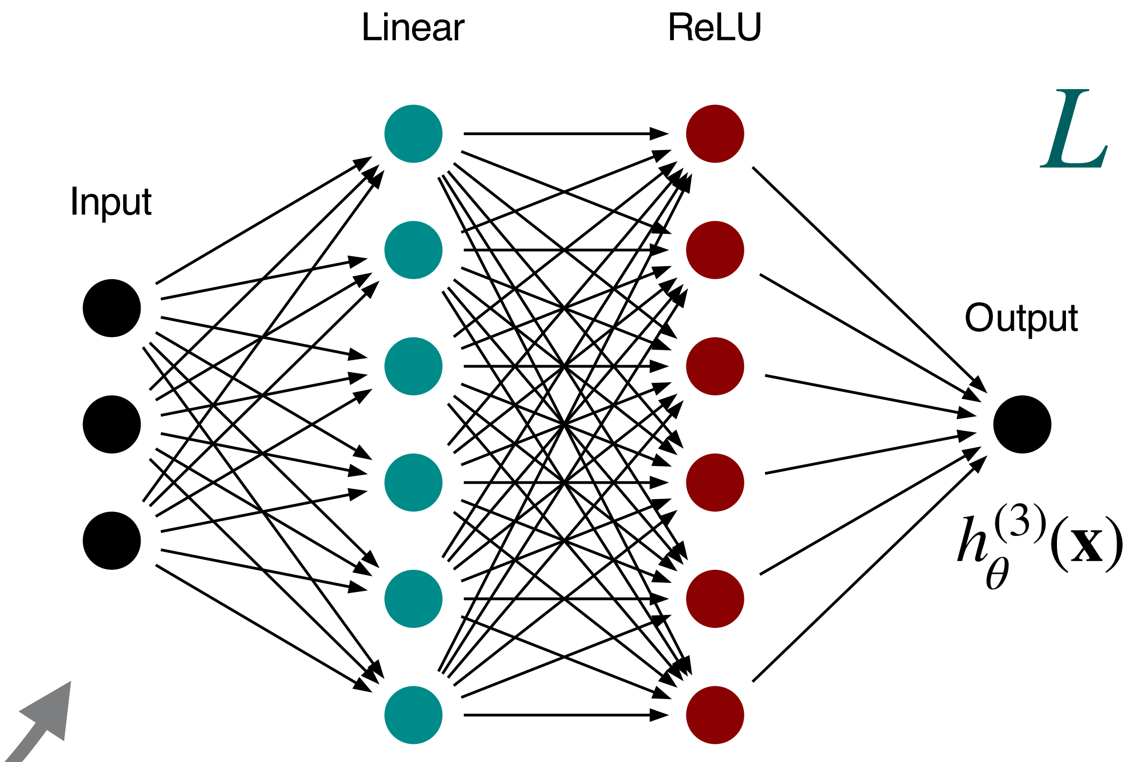
2-layer ReLU network:

$$h_\theta^{(2)}(\mathbf{x}) = \sum_{k=1}^{K} a_k[\mathbf{w}_k^\top \mathbf{x} + b_k]_+ + c$$
$$= \mathbf{a}^\top[\mathbf{W}\mathbf{x} + \mathbf{b}]_+ + c$$

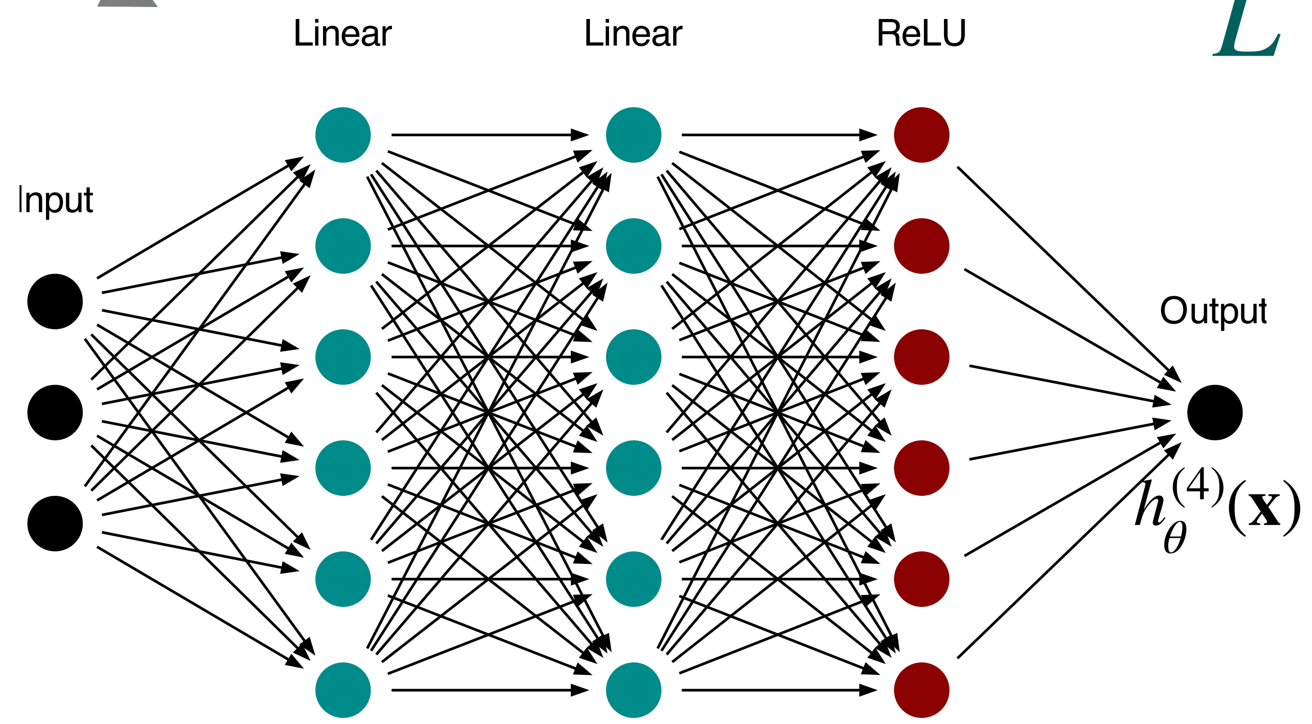
where $\theta = (\mathbf{W}, \mathbf{a}, \mathbf{b}, c)$



$L = 2$



$L = 3$



$L = 4$

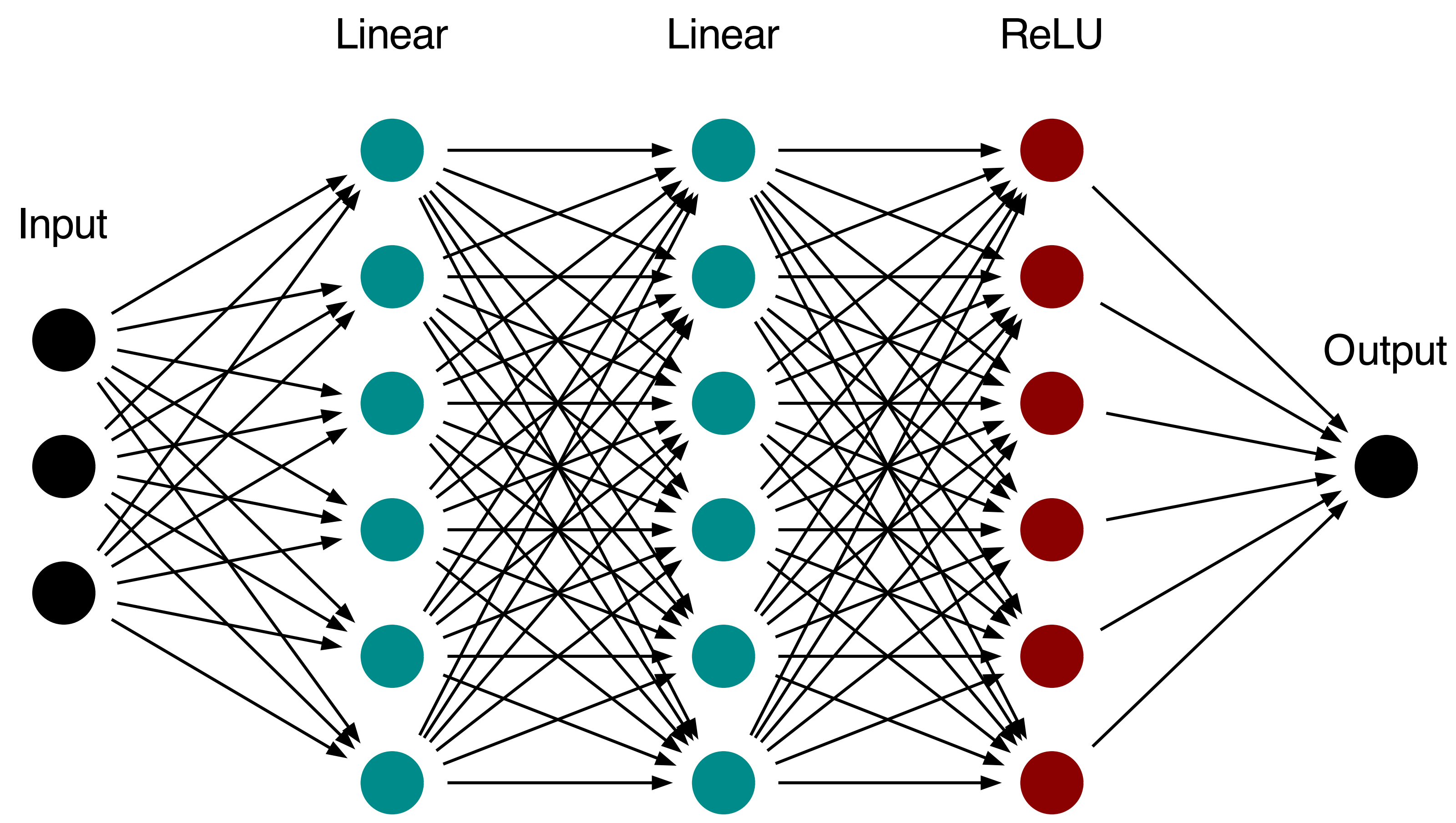
Our focus: networks with L layers in which $L-1$ layers have linear activations followed by a ReLU activation:

$$h_\theta^{(L)}(\mathbf{x}) = \mathbf{a}^\top[\mathbf{W}_{L-1}\cdots\mathbf{W}_2\mathbf{W}_1\ \mathbf{x} + \mathbf{b}]_+ + c$$

where $\theta = (\mathbf{W}_{L-1}, \ldots, \mathbf{W}_2, \mathbf{W}_1, \mathbf{a}, \mathbf{b}, c)$

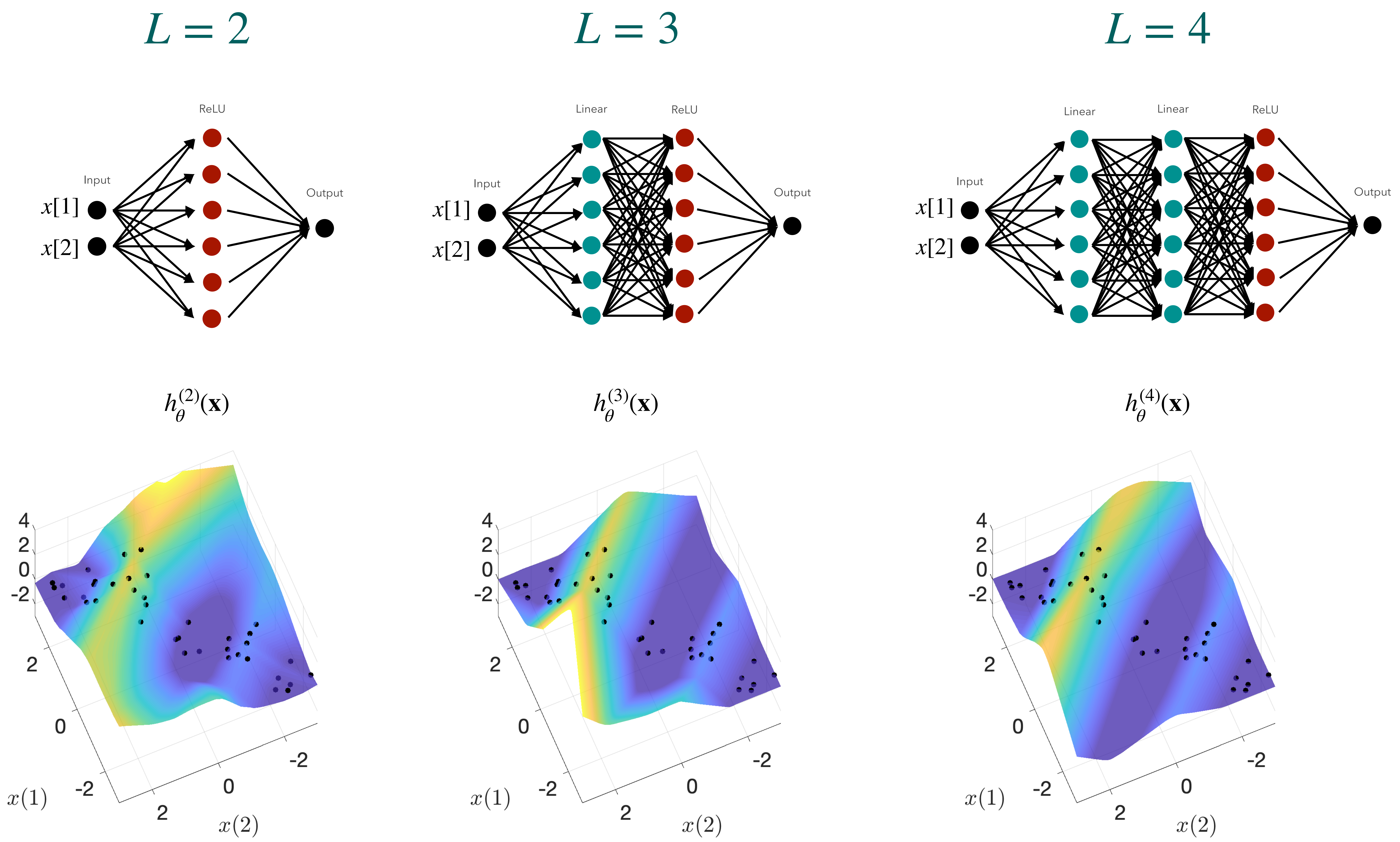
Why care about linear layers?

- The *capacity or expressivity* of the network is the same regardless of L – that is, **different behaviors for different depths solely are independent of capacity.** That is, $h_\theta^{(L)}(\mathbf{x}) = \mathbf{a}^\top[\mathbf{W}\mathbf{x} - \mathbf{b}]_+ + c$ for some $(\mathbf{W}, \mathbf{a})$ for each L.



- Empirically, linear layers…
  - Help with **generalization** *Golubeva et al. (2020)*
  - Uncover **low rank structure** *Kodak et al. (2020), Zeng and Graham (2023)*
  - Improve **training speed** *Ba and Caruana (2013); Urban et al. (2016); Arora et al. (2018)*

First pass intuition

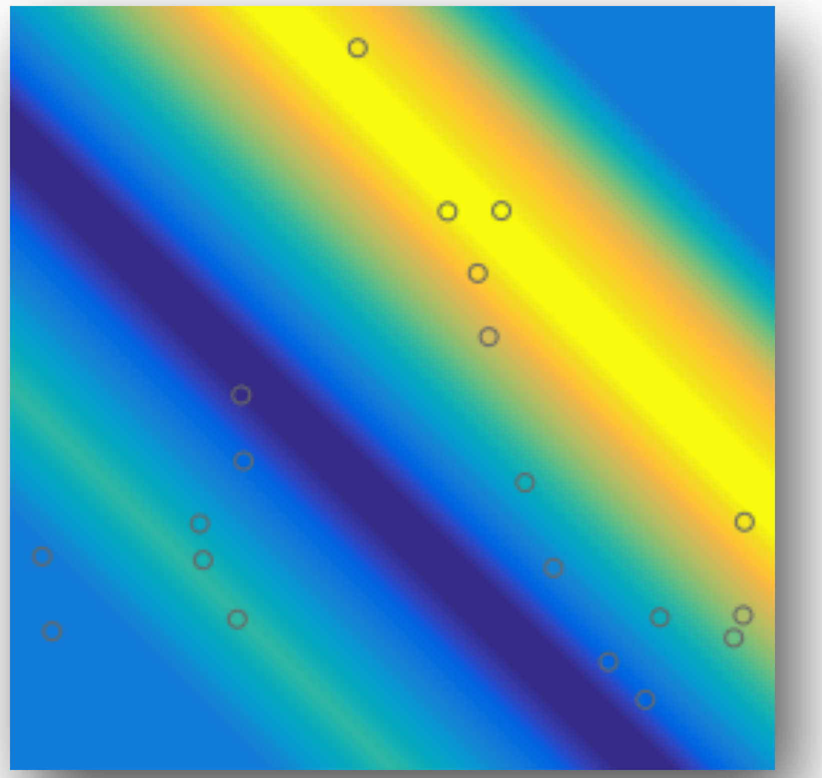


Single-Index Models

Definition: A ***single-index model*** *is a function* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *of the form*

$$f(\mathbf{x}) = g(\mathbf{v}^\top \mathbf{x}),$$

for some ***link function*** $g : \mathbb{R} \mapsto \mathbb{R}$, *where* $\mathbf{v} \in \mathbb{R}^d$ *and* $\text{range}(\mathbf{v})$ *is called the* ***central subspace.***



single-index model in $d = 2$

Zhu & Zhang (2006); Xia (2008); Yin, Li, & Cook (2008); Kakade, Kanade, Shamir, & Kalai (2011); Ganti, Balzano, & Willett (2015); Ganti, Rao, Balzano, Willett, & Nowak (2017), Bach (2017), Gollakta et al. (2024), Liu & Liao (2024)

Multi-Index Models

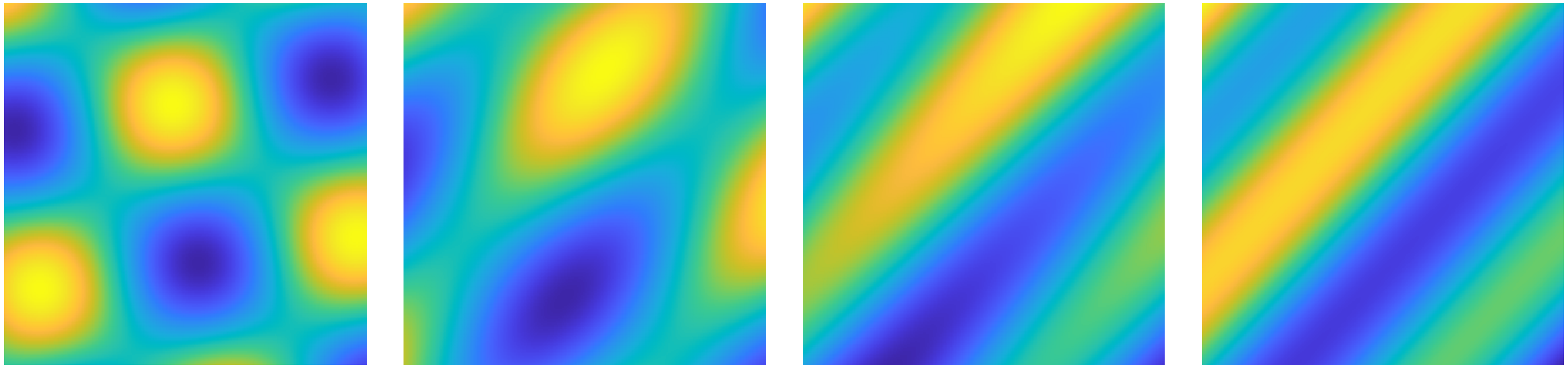
Definition: *More generally, a* ***multi-index model*** *is a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ of the form*

$$f(\mathbf{x}) = g(\mathbf{V}^\top \mathbf{x}),$$

for some ***link function*** $g : \mathbb{R}^r \mapsto \mathbb{R}$*, where* $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ *and range*$(\boldsymbol{V})$ *is called the* ***central subspace.***

Zhu & Zhang (2006); Xia (2008); Yin, Li, & Cook (2008); Kakade, Kanade, Shamir, & Kalai (2011); Ganti, Balzano, & Willett (2015); Ganti, Rao, Balzano, Willett, & Nowak (2017), Bach (2017), Gollakta et al. (2024), Liu & Liao (2024)

None of these functions are single-index models



"Far" from a single-index model

"Close" to a single-index model

Functions may be "close" to a single-index model when they vary significantly more in one direction than another

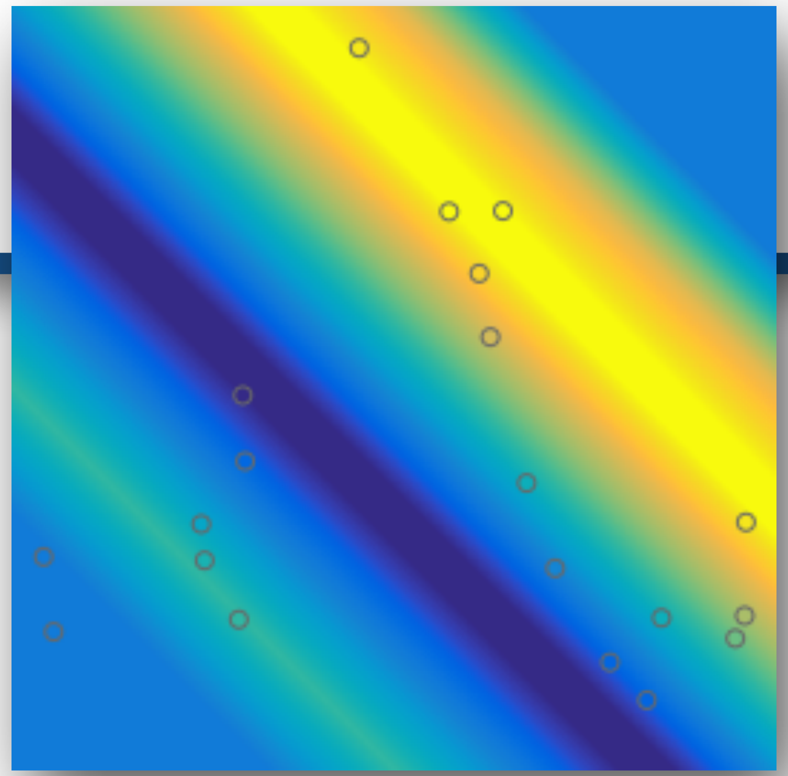
Expected Gradient Outer Product (EGOP) Matrix

Definition: *Consider the expected gradient outer product matrix of a function* $f : \mathscr{X} \mapsto \mathbb{R}$*:*

$$C_f := \mathbb{E}_X[\nabla f(X) \nabla f(X)^\top].$$

The ***principal subspace*** *of f is range(C_f). The* ***index rank*** *of f is* $\mathrm{rank}_I(f) := \mathrm{rank}(C_f)$.

$$\mathbf{v}^\top C_f \mathbf{v} = \mathbf{E}_X \left[(\mathbf{v}^\top \nabla f(X))^2 \right]$$



Index rank = 1

Samarov (1993); Hristache et al. (2001); Wu et al. (2010); Trivedi et al. (2014); Constantine, Dow, & Wang (2014); Constantine (2015); Radhakrishnan, Beaglehole, Pandit, & Belkin (2024); Radhakrishnan, Belkin, & Drusvyatskiy (2024); Radhakrishnan, Belkin, & Drusvyatskiy (2024);

Mixed variation functions and **effective** index rank

Definition: *Given a function* $f : \mathcal{X} \mapsto \mathbb{R}$ *and* $q \in (0,1]$, *the* ***mixed variation of*** f ***of order*** q *is*

$$\mathcal{MV}(f; q) := \|C_f^{1/2}\|_{\mathcal{S}^q}.$$

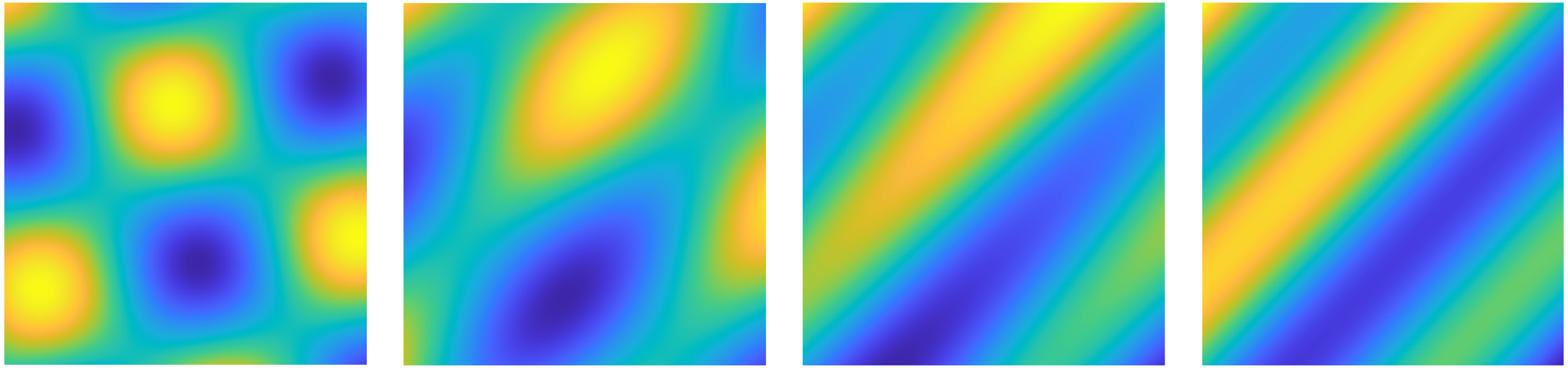
Definition: *Given a function* f *and* $\varepsilon > 0$, *define the* ***effective index rank***

$$\text{rank}_{I,\varepsilon}(f)$$

to be the number of singular values of $C_f^{1/2}$ *that are bigger than* ε.

Donoho (2000); Parhi & Nowak (2022)

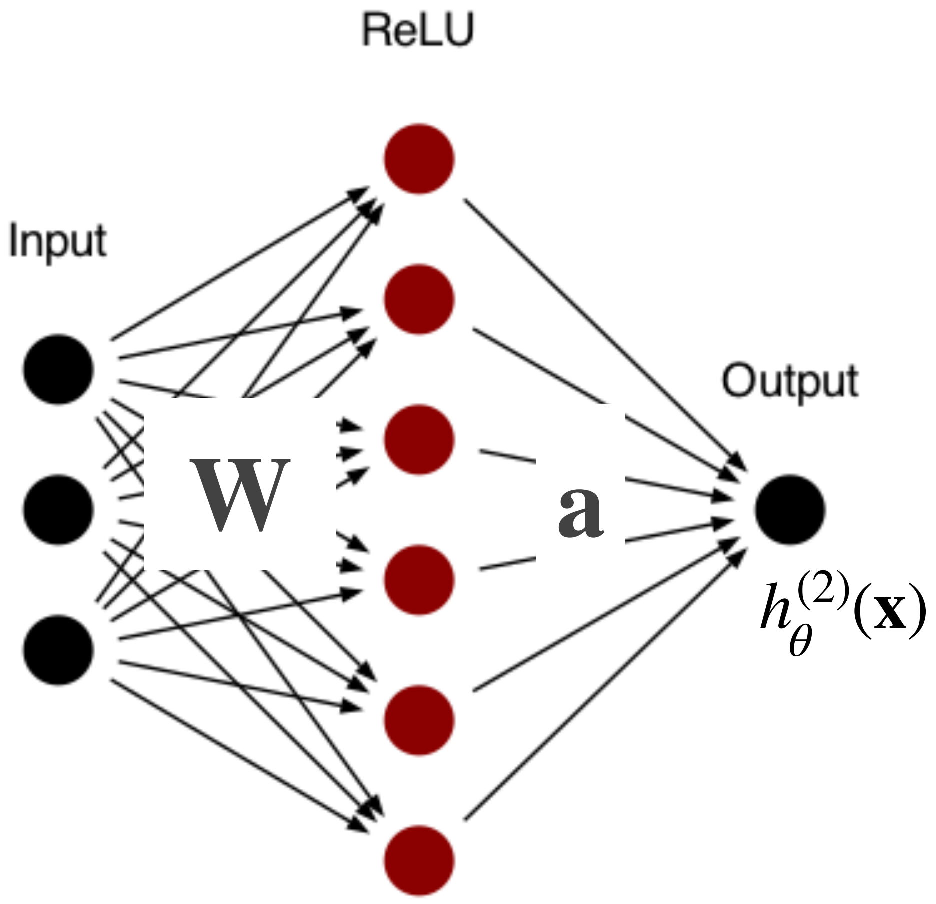
None of these functions are single-index models



Large $\mathscr{M}\mathscr{V}(f)$ → Small $\mathscr{M}\mathscr{V}(f)$

Functions with small mixed-variation are “close” to having small index rank and can vary significantly more in one direction than another

Two-Layer Network

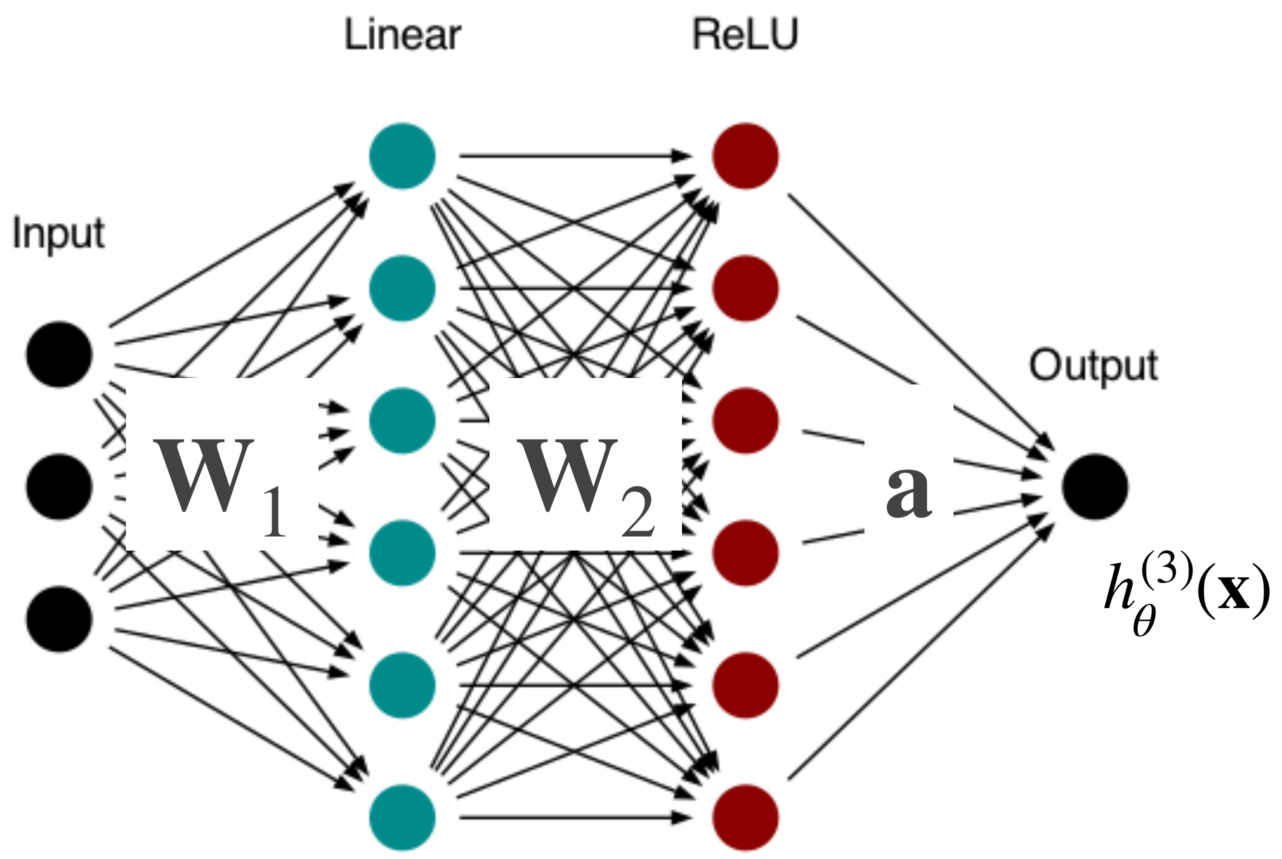


$$h_\theta^{(2)}(\mathbf{x}) = \mathbf{a}^\top[\mathbf{W}\mathbf{x} + \mathbf{b}]_+ + c$$

$$C_2(\theta) = \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{W}\|_F^2$$

$$R_2(f) = \min_\theta \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{W}\|_F^2 \quad \text{s.t.} \quad f = h_\theta^{(2)}$$

Three-Layer Network



$$h_\theta^{(3)}(\mathbf{x}) = \mathbf{a}^\top[\mathbf{W}\mathbf{x} + \mathbf{b}]_+ + c$$

where $\mathbf{W} = \mathbf{W}_2\mathbf{W}_1$

$$C_3(\theta) = \frac{1}{3}\|\mathbf{a}\|_2^2 + \frac{1}{3}\|\mathbf{W}_1\|_F^2 + \frac{1}{3}\|\mathbf{W}_2\|_F^2$$

$$R_3(f) = \min_\theta \frac{1}{3}\|\mathbf{a}\|_2^2 + \frac{1}{3}\|\mathbf{W}_1\|_F^2 + \frac{1}{3}\|\mathbf{W}_2\|_F^2 \quad \text{s.t.} \quad f = h_\theta^{(3)}$$

$$\boxed{\min_{\mathbf{W}_1\mathbf{W}_2 = \mathbf{W}} \frac{1}{2}\|\mathbf{W}_1\|_F^2 + \frac{1}{2}\|\mathbf{W}_2\|_F^2 = \|\mathbf{W}\|_*}$$

Srebro, Rennie, & Jaakkola, 2004

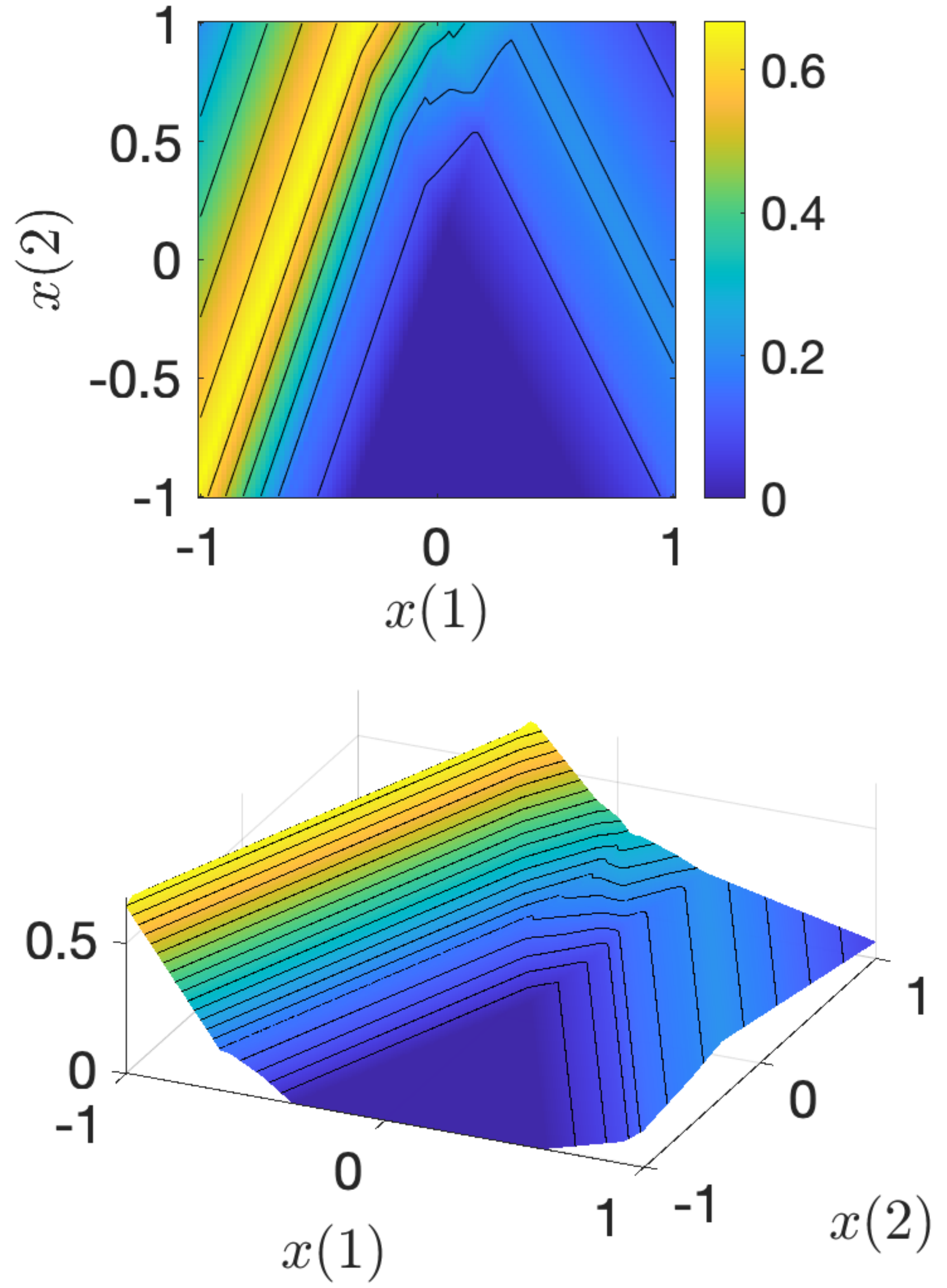
$$R_L(f) = \min_{\theta} \frac{1}{L} \|\mathbf{a}\|_2^2 + \frac{L-1}{L} \|\mathbf{W}\|_{\mathcal{S}^q}^q \quad \text{s.t.} \quad f = h_\theta^{(2)}$$

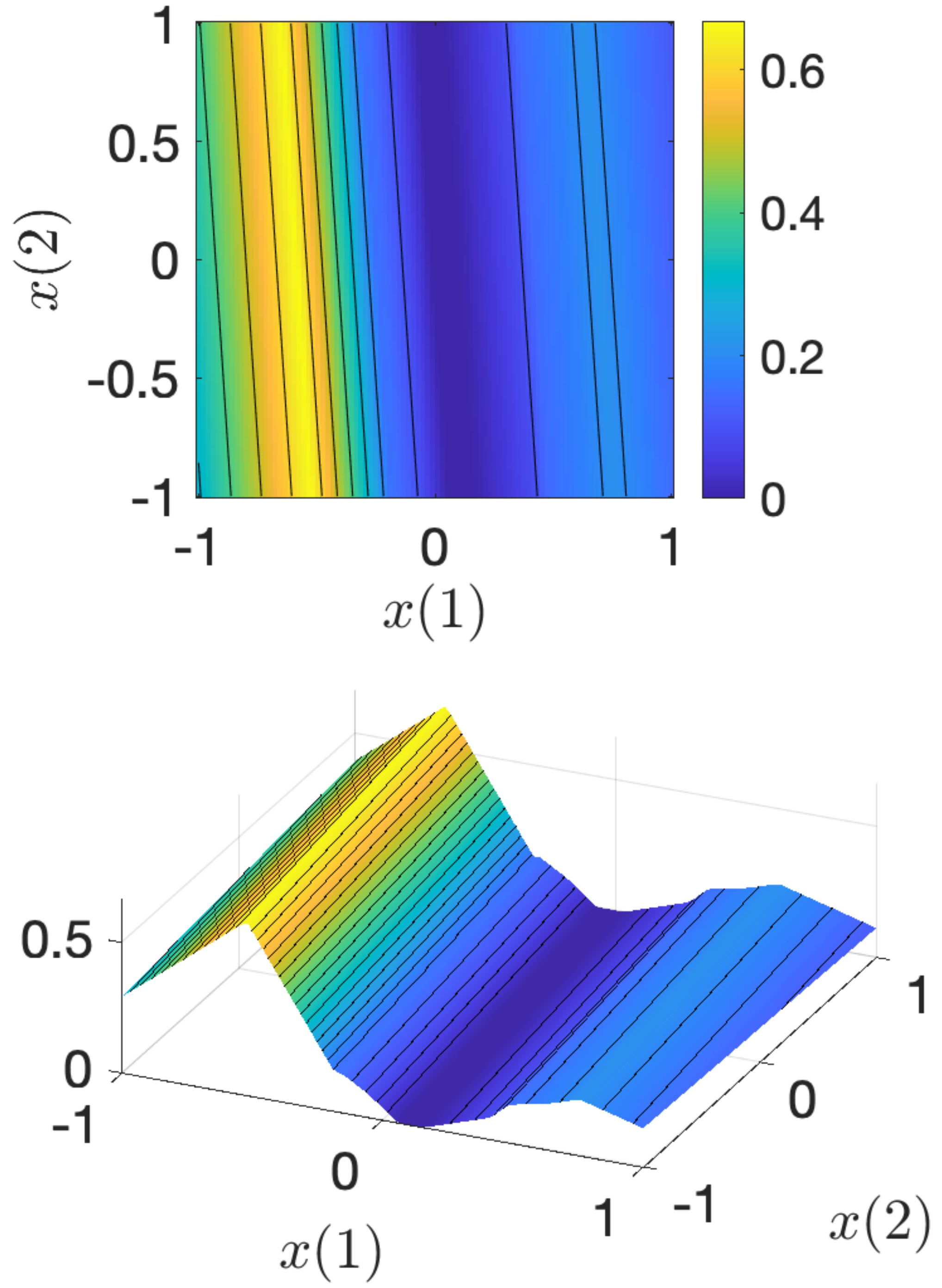
where $q = \dfrac{2}{L-1}$

$$R_L(f) = \min_{\theta} \frac{1}{L}\|\mathbf{a}\|_2^2 + \frac{L-1}{L}\|\mathbf{W}\|_{\mathcal{S}^q}^q \quad \text{s.t.} \quad f = h_\theta^{(2)}$$

where $q = \dfrac{2}{L-1}$

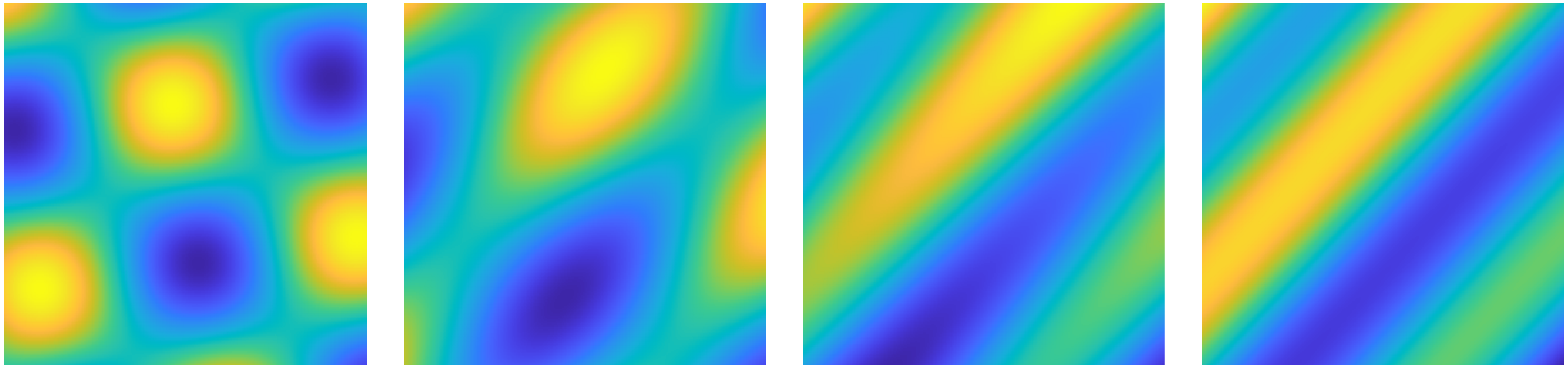
Function with **unaligned** ReLU units: zig-zig contour lines





Function with **aligned** ReLU units: parallel contour lines

Minimizing the R_L-cost promotes learning functions that have small **mixed variation**, such as **single- and multi- index models**



Mixed Variation, Index Rank, and the Representation Cost

Theorem:

$$\max\left(\mathscr{MV}(f;\tfrac{2}{L-1})^{2/L}, R_2(f)^{2/L}\right) \leq R_L(f) \leq \mathit{rank}_I(f)^{\frac{L-2}{L}} R_2(f)^{2/L}$$

Minimizing the R_L-cost favors functions that vary primarily along a **low-dimensional subspace**, and are **smooth** along that subspace.

Mixed Variation, Index Rank, and the Representation Cost

Theorem:

$$\max\left(\mathscr{MV}(f;\tfrac{2}{L-1})^{2/L}, R_2(f)^{2/L}\right) \leq R_L(f) \leq rank_I(f)^{\frac{L-2}{L}} R_2(f)^{2/L}$$

Corollary:

$$\lim_{L\to\infty} R_L(f) = rank(f)$$

Corollary: *If f_ℓ, f_h are such that $rank_I(f_\ell) < rank_I(f_h)$, then for L sufficiently large,*

$$R_L(f_\ell) < R_L(f_h).$$

Minimal-norm interpolants are nearly low index rank

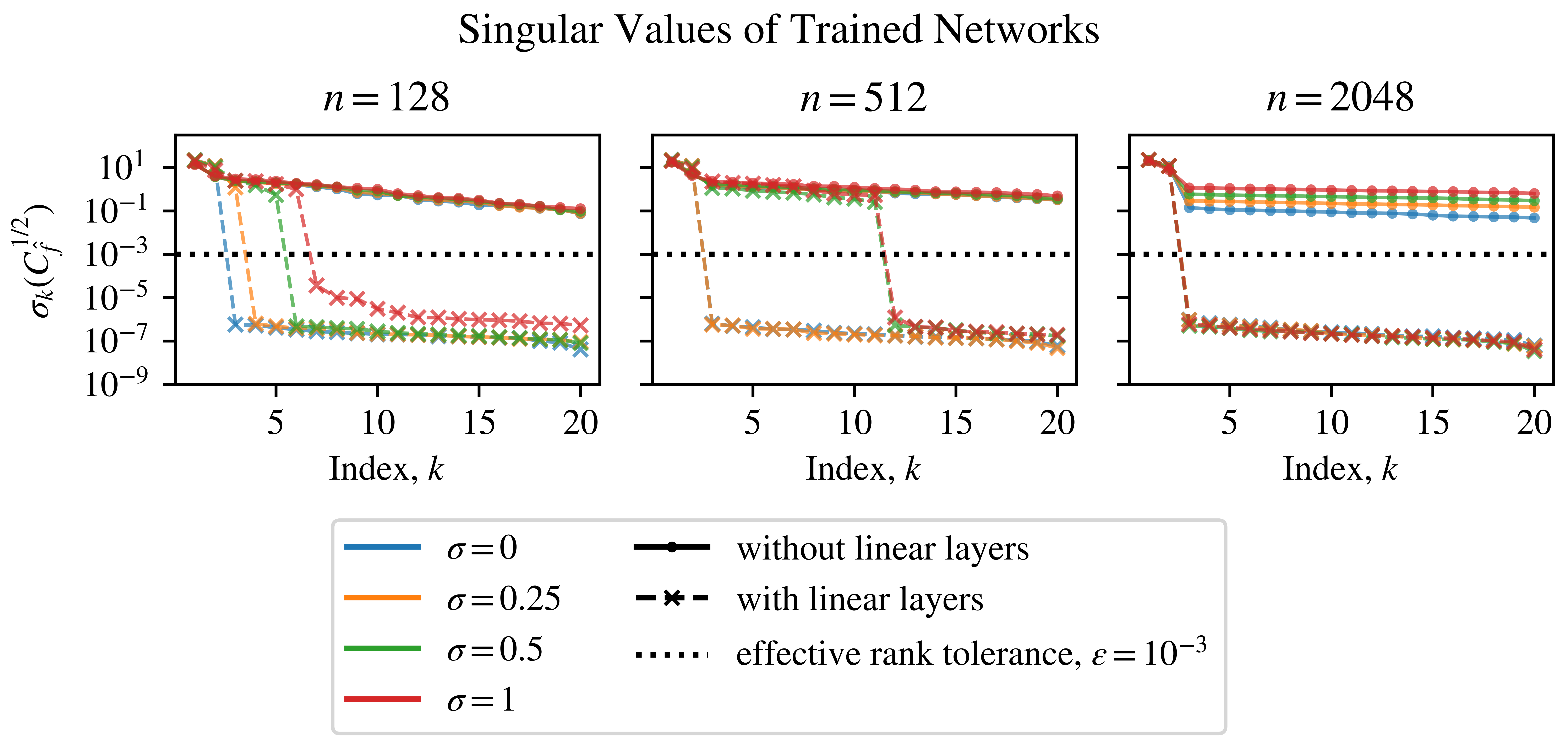
Theorem: *Any interpolant $\hat{f}_L$ of a dataset $\mathscr{D}$ that has minimal R_L cost has effective index rank bounded as*

$$\text{rank}_{I,\varepsilon}(\hat{f}_L) \leq \min_{s\in[d]} s \left(\frac{\mathscr{I}_s(\mathscr{D})}{\varepsilon\sqrt{s}} \right)^{\frac{2}{L-1}}$$

where $\mathscr{I}_s(\mathscr{D})$ denotes the R_2 cost needed to interpolate $\mathscr{D}$ with a function of index rank $\leq s$.

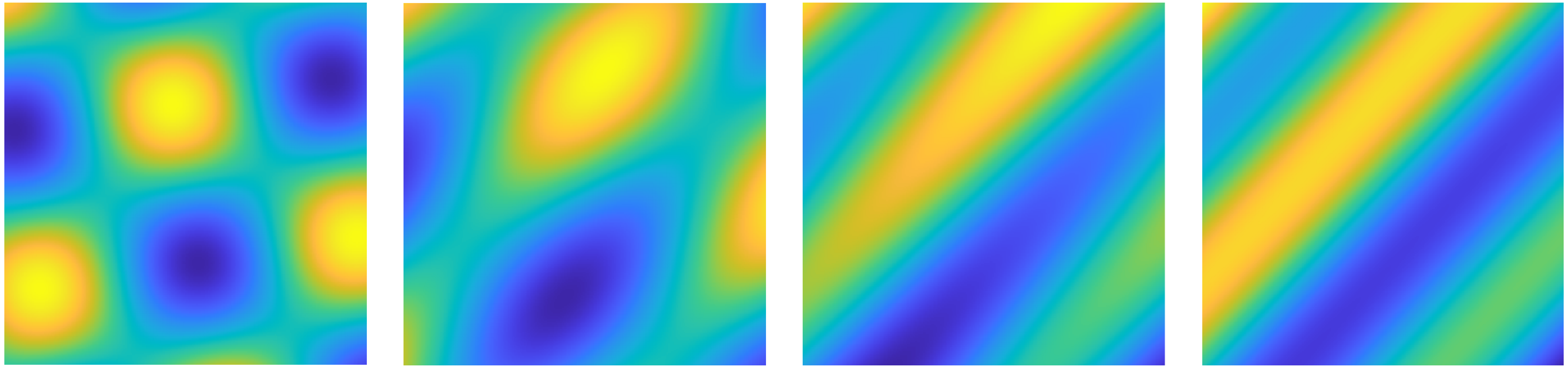
Corollary: *Suppose that a dataset $\mathscr{D}$ is generated by a function f^* with $\text{rank}_I(f^*) = r$ and bounded R_2 cost. Then If $R_2(f^*) < \varepsilon\sqrt{r}\left(1 + \frac{1}{\sqrt{r}}\right)^{\frac{L-1}{2}}$. Then $\text{rank}_{I,\varepsilon}(\hat{f}_L) \leq r$.*

Numerical Example



Adding linear layers causes trained networks to have low effective index rank.

Minimizing the R_L-cost promotes learning functions that have small **mixed variation**, such as **single- and multi- index models**



Thank you!

Greg Ongie

Rebecca Willett

https://arxiv.org/pdf/2305.15598

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2140001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.