Intrinsic dimensionality and generalization properties of the $\mathscr{R}$-norm inductive bias.

Navid Ardeshir

Columbia University, Department of Statistics

Based on joint work with Clayton Sanford and Daniel Hsu

Benign Overfitting

- Large (overparameterized) deep learning models that interpolate data can generalize well. [P. Nakkiran et al. ‘19]

- Network size is not the main form of capacity control. Alternatives might be the size of weights. [B. Neyshabur et al. ’14]

- Controlling the ℓ_1-norm of the top layer weights may result in good generalization. [P. Bartlett ’98]

- These bounds do not depend on size of the network!

For valid generalization, the size of the weights is more important than the size of the network

Peter L. Bartlett
Department of Systems Engineering
Research School of Information Sciences and Engineering
Australian National University
Canberra, 0200 Australia
Peter.Bartlett@anu.edu.au

Abstract

This paper shows that if a large neural network is used for a pattern classification problem, and the learning algorithm finds a network with small weights that has small squared error on the training patterns, then the generalization performance depends on the size of the weights rather than the number of weights. More specifically, consider an ℓ-layer feed-forward network of sigmoid units, in which the sum of the magnitudes of the weights associated with each unit is bounded by A. The misclassification probability converges to an error estimate (that is closely related to squared error on the training set) at rate $O((cA)^{\ell(\ell+1)/2}\sqrt{(\log n)/m})$ ignoring log factors, where m is the number of training patterns, n is the input dimension, and c is a constant. This may explain the generalization performance of neural networks, particularly when the number of training examples is considerably smaller than the number of weights. It also supports heuristics (such as weight decay and early stopping) that attempt to keep the weights small during training.

1 Introduction

Results from statistical learning theory give bounds on the number of training examples that are necessary for satisfactory generalization performance in classification problems, in terms of the Vapnik-Chervonenkis dimension of the class of functions used by the learning system (see, for example, [13, 5]). Baum and Haussler [4] used these results to give sample size bounds for multi-layer threshold networks

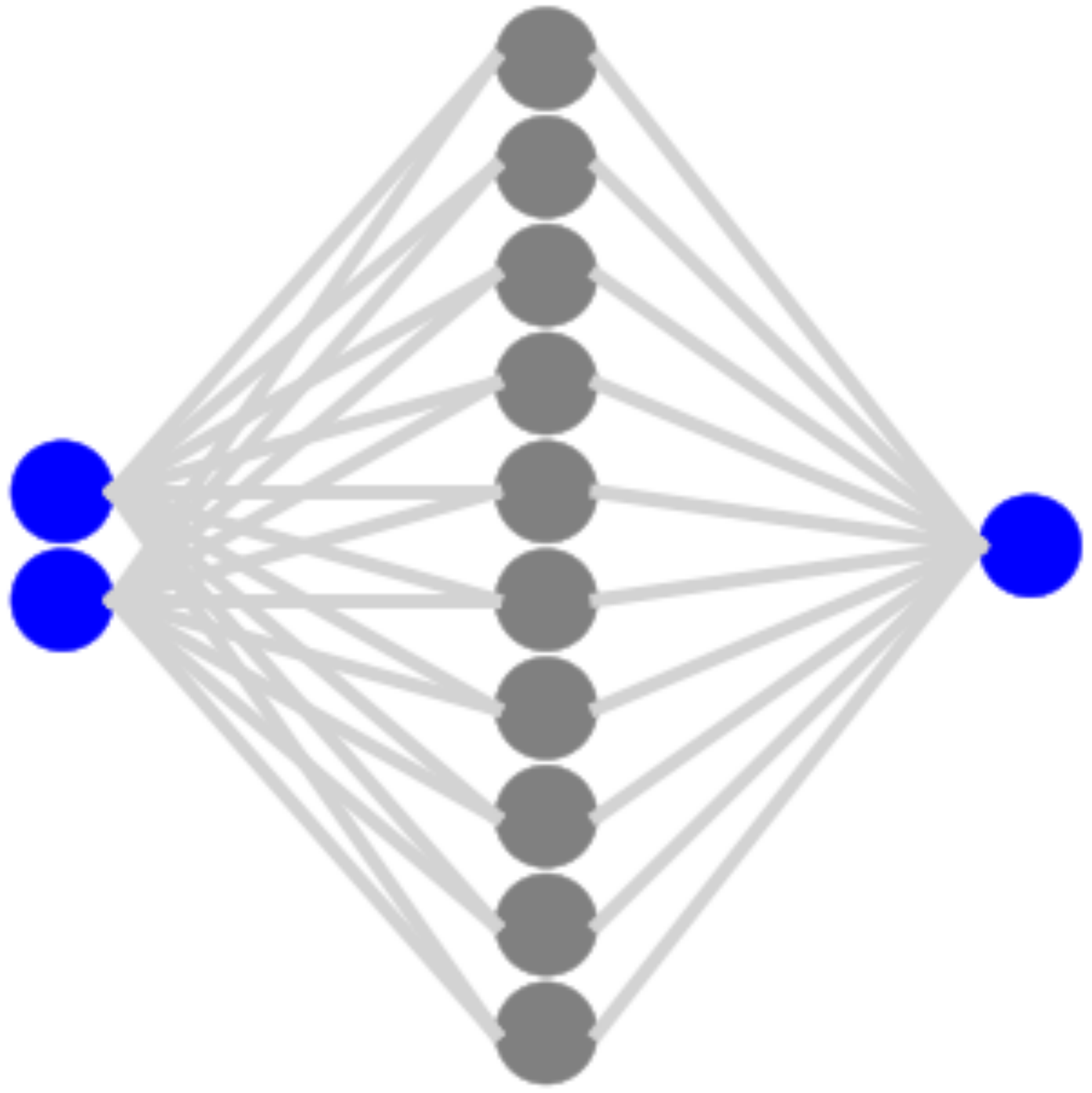
Learning with Restriction on Weight Norms

- We can think of networks with infinite size but bounded in some norm of their weights (known as “effective” capacity control).

- Almost all functions can be represented/ approximated by such infinitely wide networks. [Barron ’93][Bach ’17]

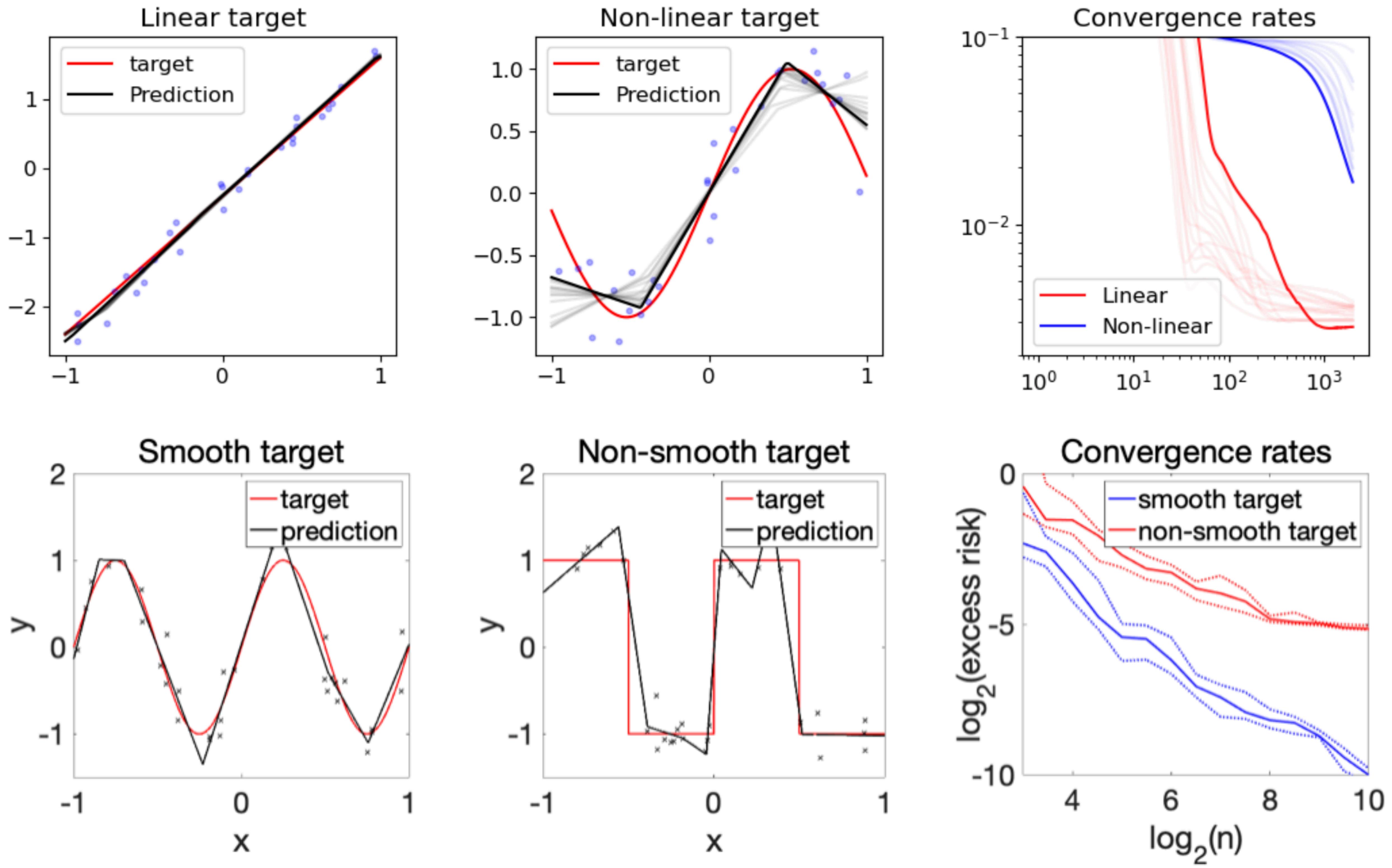
- **Learning** can be translated as finding a function (in the entire function space) that fits the data but with small “effective” capacity. [Savarese et al. ’19]

- We are interested in **statistical** properties of such learned functions under specific data distributions.



Learning with Wide Neural Nets

- Without any assumption on the data we are doomed to use exponentially large number of samples in data dimension.
- This is known as **curse of dimensionality**.
- Wide neural nets can **beat** the curse of dimensionality for regression. [Bach ’17]
- Adaptivity to smoothness and **low dimensional structure** (data lies on a low dimensional manifold).
- How do NNs achieve this?



[taken from Bach’s blog]

Learning with Wide Neural Nets

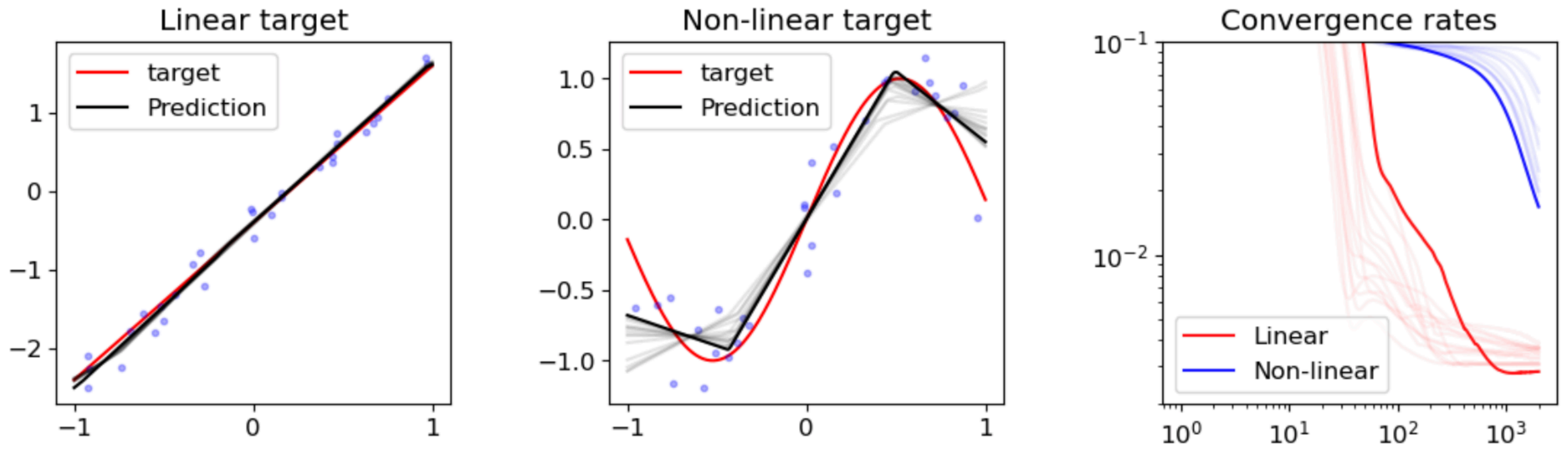
- Without any assumption on the data we are doomed to use exponentially large number of samples in data dimension.

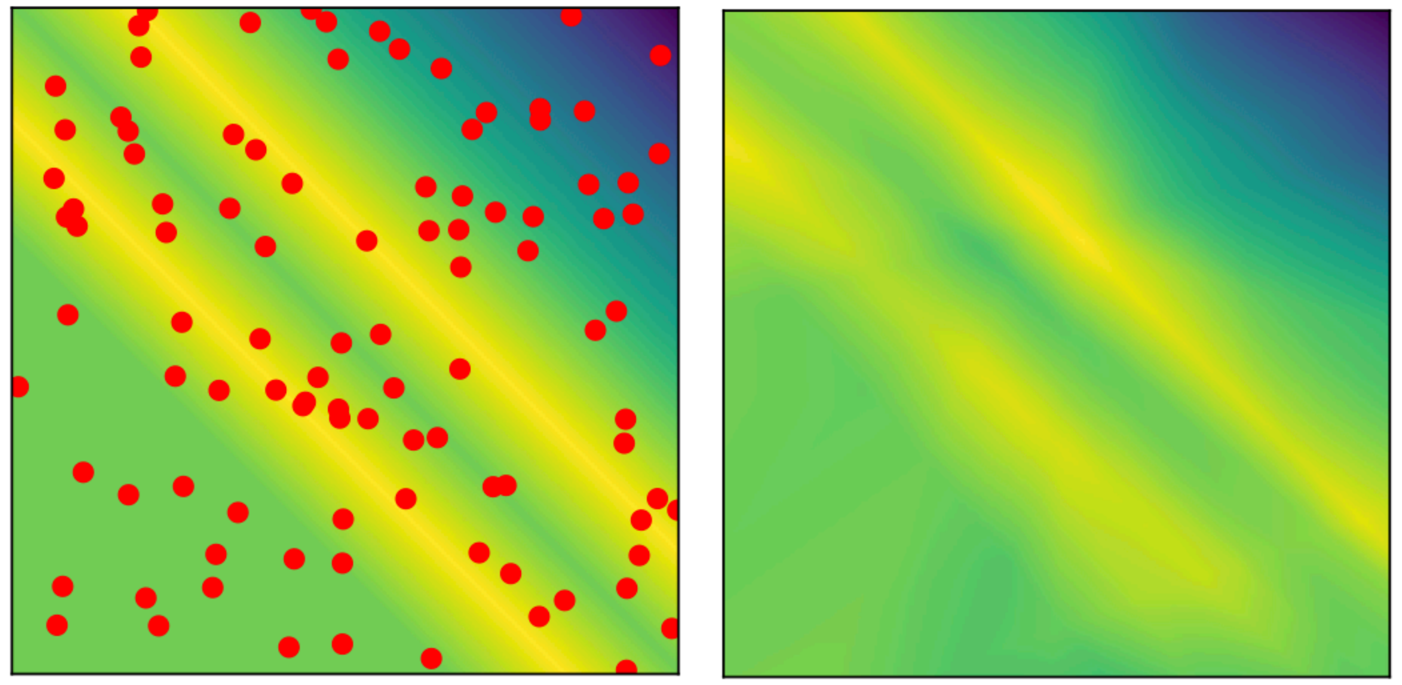
- This is known as **curse of dimensionality**.

- Wide neural nets can **beat** the curse of dimensionality for regression. [Bach ’17]

- Adaptivity to smoothness and **low dimensional structure** (data lies on a low dimensional manifold).

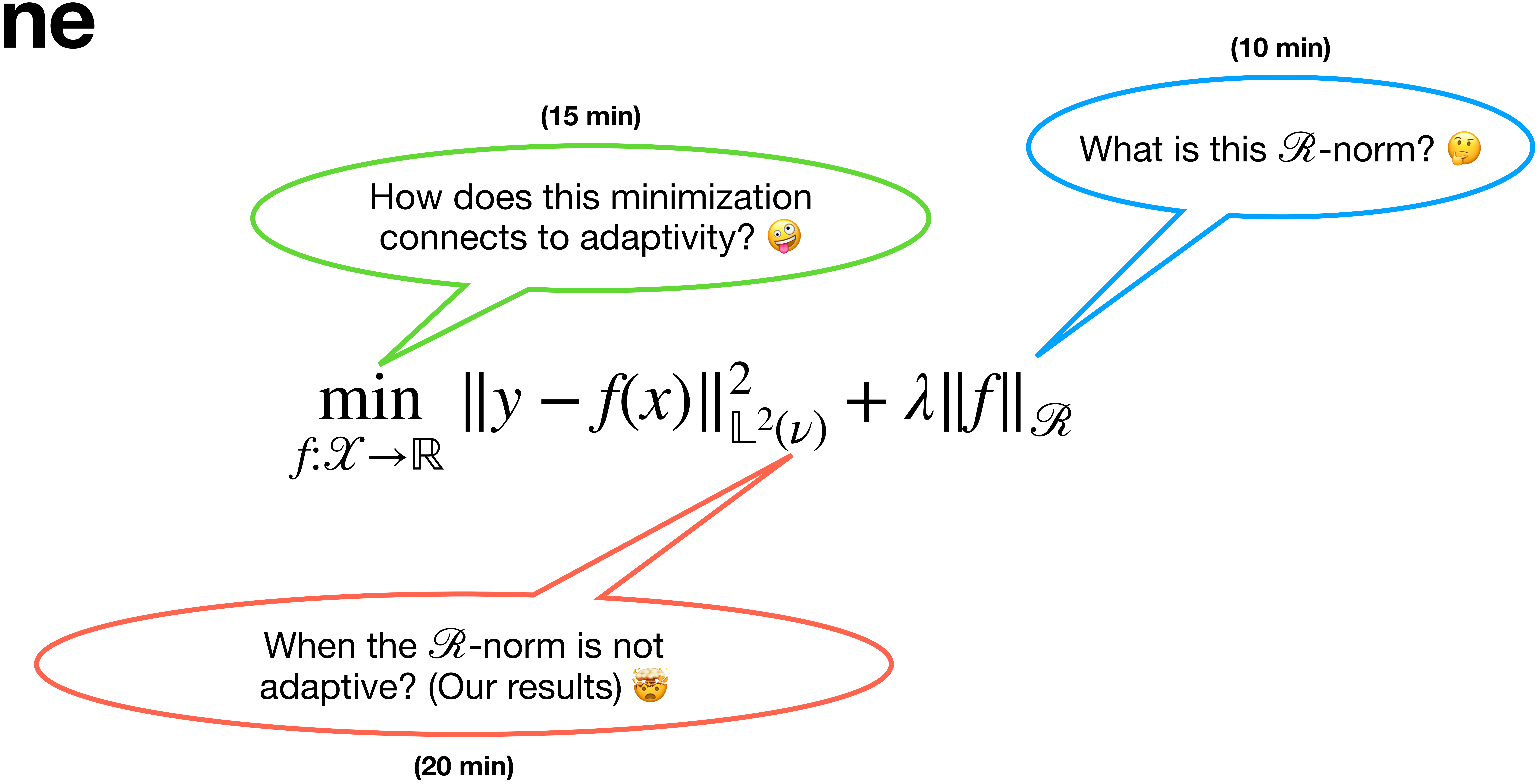
- How do NNs achieve this?





[Parhi et al. ’22]

Outline



(15 min)

How does this minimization connects to adaptivity? 🤪

(10 min)

What is this $\mathscr{R}$-norm? 🤔

$$\min_{f:\mathscr{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu)} + \lambda\|f\|_{\mathscr{R}}$$

When the $\mathscr{R}$-norm is not adaptive? (Our results) 🤯

(20 min)

How does this minimization connects to adaptivity? 🤪

(10 min)

What is this $\mathscr{R}$-norm? 🤔

$$\min_{f:\mathscr{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu)} + \lambda \|f\|_{\mathscr{R}}$$

When the $\mathscr{R}$-norm is not adaptive? (Our results) 🤯

Supervised Learning Setting

- Let $\mathscr{X} \subseteq \mathbb{R}^d$ be compact. Given samples $(x_i, y_i)_{i \le n} \sim \nu \in \mathscr{P}(\mathscr{X} \times \mathscr{Y})$ denote the emp. measure as ν_n.
- Two layer neural network with m neurons:
  - Let $\theta = (a^{(i)}, b^{(i)}, c^{(i)})_{i \le m} \in (\mathbb{R} \times \mathbb{R}^d \times \mathbb{R})^m$ and,

$$f_\theta : \mathscr{X} \to \mathbb{R} : x \mapsto \sum_{i=1}^{m} a^{(i)} \sigma(b^{(i)\top} x + c^{(i)}) .$$

- Given samples the **Goal** is to find $\hat{\theta}, \hat{m}$ that minimize the out of sample loss $\|y - f_{\hat{\theta}}(x)\|_{\mathbb{L}_2(\nu)}$.
  - We consider the following ERM regularized with a capacity function C,

$$(\hat{\theta}, \hat{m}) \in \min_{m \in \mathbb{N}} \min_{\theta} \|y - f_\theta(x)\|^2_{\mathbb{L}^2(\nu_n)} + \lambda C(\theta) .$$

Capacity Control

[Neyshabur et al. ’14]

- As argued in Neyshabur et. al. taking the size of the network is not an informative capacity control.

$$\inf_{m\in\mathbb{N}} \inf_{\theta} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + \lambda m$$

- A natural regularization used in practice is weight decay (without regularizing bias terms)

$$\inf_{m\in\mathbb{N}} \inf_{\theta} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + \lambda \sum_{i=1}^{m} |a^{(i)}|^2 + \|b^{(i)}\|^2$$

- For ReLU networks this is **equivalent** to:

$$\boxed{\inf_{m\in\mathbb{N}} \inf_{\theta\in\Theta^m} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + 2\lambda \sum_{i=1}^{m} |a^{(i)}|}$$

  - The scale of bottom layer weights can be absorbed into top layer weights.

$$\Theta = \{(a, b, c) \in \mathbb{R} \times \mathbb{S}^{d-1} \times [-c_0, c_0]\}$$

  - Transformation $(a, b, c) \mapsto (at, b/t, c/t)$ does not change the output of the network.

$$\sum_{i=1}^{m} |a^{(i)}|^2 + \|b^{(i)}\|^2 \geq \sum_{i=1}^{m} 2|a^{(i)}|\|b^{(i)}\|$$

Capacity Control

[Neyshabur et al. ’14]

- As argued in Neyshabur et. al. taking the size of the network is not an informative capacity control.

$$\inf_{m\in\mathbb{N}} \inf_{\theta} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + \lambda m$$

- A natural regularization used in practice is weight decay (without regularizing bias terms)

$$\inf_{m\in\mathbb{N}} \inf_{\theta} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + \lambda \sum_{i=1}^{m} |a^{(i)}|^2 + \|b^{(i)}\|^2$$

- For ReLU networks this is **equivalent** to:
  - The scale of bottom layer weights can be absorbed into top layer weights.
  - Transformation $(a, b, c) \mapsto (at, b/t, c/t)$ does not change the output of the network.

$$\boxed{\inf_{m\in\mathbb{N}} \inf_{\theta\in\Theta^m} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + 2\lambda \sum_{i=1}^{m} |a^{(i)}|}$$

$$\Theta = \{(a, b, c) \in \mathbb{R} \times \mathbb{S}^{d-1} \times [-c_0, c_0]\}$$

$$\sum_{i=1}^{m} |a^{(i)}|^2 + \|b^{(i)}\|^2 \geq \sum_{i=1}^{m} 2|a^{(i)}|\|b^{(i)}\|$$

Convex Optimization Problem

Proposition. [Savarese et al. ’19]
Let $\mathscr{M}$ denote the space of signed measures equipped with total variation norm $|\cdot|$. Then the two optimization problems are equivalent and the minimum is attained by an even measure:

$$\inf_{m\in\mathbb{N}} \inf_{\theta\in\Theta^m} \|y - \sum_{i=1}^{m} a^{(i)}\sigma(b^{(i)\top}x + c^{(i)})\|^2_{\mathbb{L}^2(\nu_n)} + 2\lambda \sum_{i=1}^{m} |a^{(i)}| = \min_{\rho\in\mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0])} \|y - \int \sigma(b^\top x + c)\rho(db, dc)\|^2_{\mathbb{L}^2(\nu_n)} + 2\lambda|\rho|$$

- One can always have discrete measure $\rho_\theta = \sum_{i=1}^{m} a^{(i)}\delta(\cdot - (b^{(i)}, c^{(i)}))$ with total variation $|\rho_\theta| = \sum_{i=1}^{m} |a^{(i)}|$.

- Every integral can be approximated arbitrarily well with finite sums.

- Minimum is attained since the space of signed measures with bounded variation is compact as a consequence of Prokhorov’s Thm.

Neural Network Function Space

- What functions can be implemented by an infinite neural network?

$$\min_{\rho\in\mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0])} \|y-\int\sigma(b^\top x+c)\rho(db,dc)\|^2_{\mathbb{L}^2(\nu_n)}+2\lambda|\rho|$$

  - Effectively all continuous functions. [Barron ’97][Leshno et al. ’93]

$$\forall f\in\mathscr{C}(\mathscr{X})\Rightarrow\exists\rho, f(x)=f_\rho(x)=\int\sigma(b^\top x+c)\rho(db,dc)$$

- It is natural define the functional norm:

$$\boxed{\|f\|_{\mathscr{F}}=\inf\{\,|\rho|\,:\rho\in\mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0]), f_\rho(x)=f(x),\forall x\in\mathscr{X}\}}$$

  - The function space that neural networks can represent is denoted by $\mathscr{F}$

$$\mathscr{F}=\{f:\mathscr{X}\to\mathbb{R}:\|f\|_{\mathscr{F}}<\infty\}$$

- The optimization problem can be reformulated in terms of this norm.

$$\inf_{f\in\mathscr{F}}\|y-f(x)\|^2_{\mathbb{L}^2(\nu_n)}+\lambda\|f\|_{\mathscr{F}}$$

Neural Network Function Space

- What functions can be implemented by an infinite neural network?

$$\min_{\rho\in\mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0])} \|y - \int \sigma(b^\top x + c)\rho(db,dc)\|^2_{\mathbb{L}^2(\nu_n)} + 2\lambda\,|\,\rho\,|$$

  - Effectively all continuous functions. [Barron ’97][Leshno et al. ’93]

$$\forall f \in \mathscr{C}(\mathscr{X}) \Rightarrow \exists \rho, f(x) = f_\rho(x) = \int \sigma(b^\top x + c)\rho(db,dc)$$

- It is natural define the functional norm:

$$\boxed{\|f\|_{\mathscr{F}} = \inf\{\,|\,\rho\,|\, : \rho \in \mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0]), f_\rho(x) = f(x), \forall x \in \mathscr{X}\,\}}$$

  - The function space that neural networks can represent is denoted by $\mathscr{F}$

$$\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R} : \|f\|_{\mathscr{F}} < \infty\}$$

- The optimization problem can be reformulated in terms of this norm.

$$\inf_{f\in\mathscr{F}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu_n)} + \lambda\|f\|_{\mathscr{F}}$$

What is $\mathscr{R}$-Norm?

[Parhi et al. ’22]

- Can we actually characterize $\|\cdot\|_{\mathscr{F}}$ norm explicitly?

$$\|f\|_{\mathscr{F}} = \inf\{\,|\rho|\,:\,\rho \in \mathscr{M}(\mathbb{S}^{d-1}\times[-c_0,c_0]), f_\rho(x) = f(x), \forall x \in \mathscr{X}\}$$

- Yes! It’s related to Radon transform given that the function is smooth enough. [Ongie et al. ’20]

$$\|f\|_{\mathscr{F}} = \|\mathscr{R}(\Delta^{\frac{d+1}{2}}f)\|_{\mathbb{L}^1(\mathbb{S}^{d-1}\times[-c_0,c_0])} + |f(0)| + \sum_{k=1}^{d}|f(e_k) - f(0)|$$

- Linear functions are the null space for the first term in the norm.

- For a cleaner formulation Ongie et al. considered a semi-norm insensitive to linear functions $\mathscr{P}_1$.

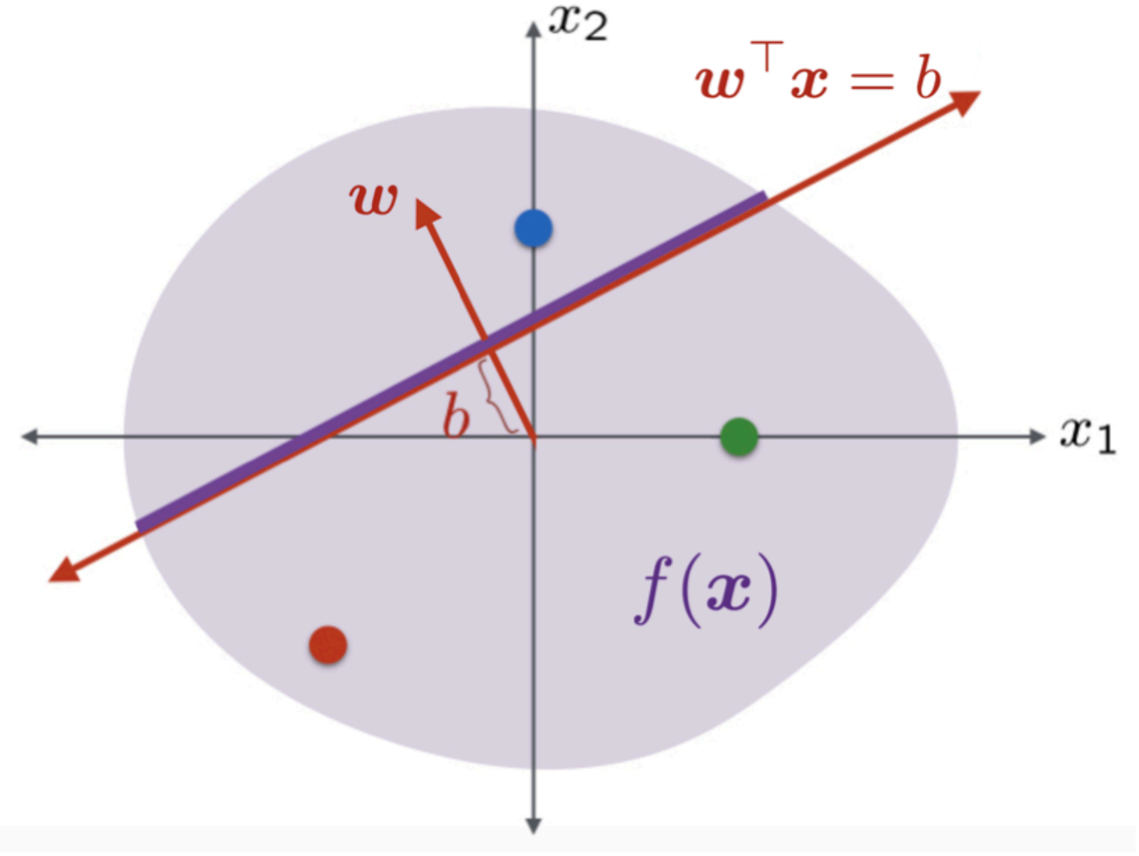
$$\boxed{\|f\|_{\mathscr{R}} = \min_{p\in\mathscr{P}_1}\|f+p\|_{\mathscr{F}} = \|\mathscr{R}(\Delta^{\frac{d+1}{2}}f)\|_{\mathbb{L}^1(\mathbb{S}^{d-1}\times[-c_0,c_0])}}$$

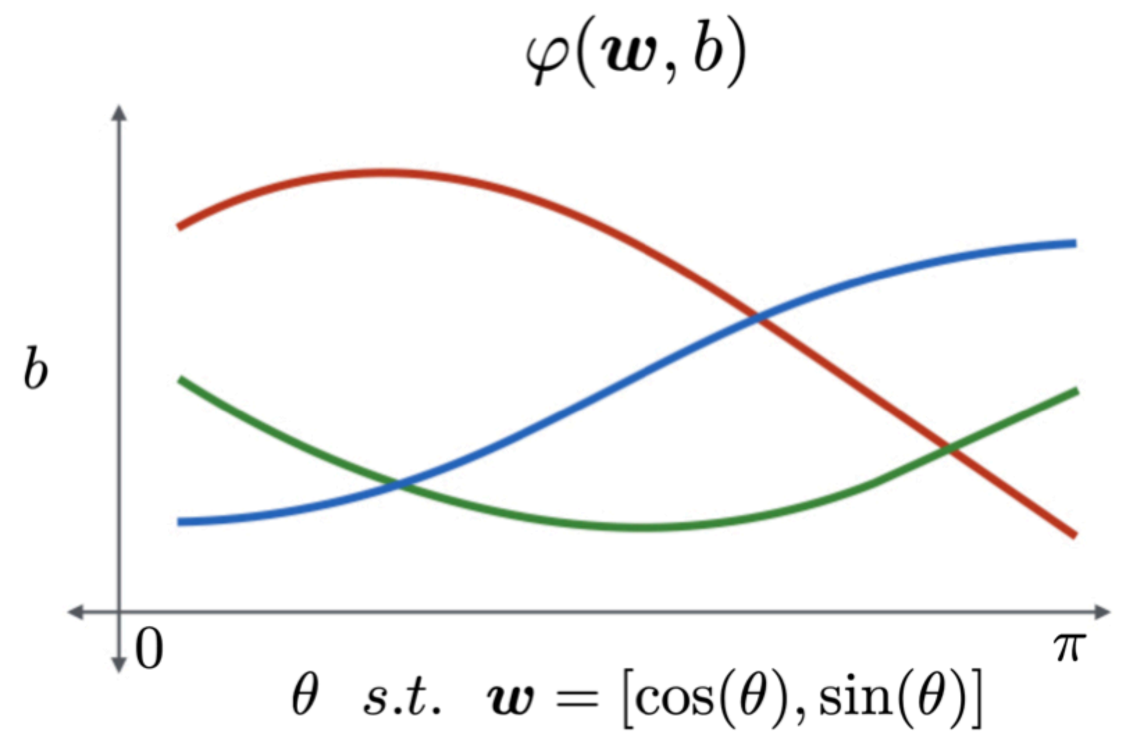
- The space induced by $\mathscr{R}$-norm remains the same as $\mathscr{F}$.

$$\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R} : \|f\|_{\mathscr{R}} < \infty\} = \{f : \mathscr{X} \to \mathbb{R} : \|f\|_{\mathscr{F}} < \infty\}$$

What is $\mathscr{R}$-Norm?

[Parhi et al. ’22]





- Can we actually characterize $\|\cdot\|_{\mathscr{F}}$ norm explicitly?

$$\|f\|_{\mathscr{F}} = \inf\{\,|\rho| : \rho \in \mathscr{M}(\mathbb{S}^{d-1} \times [-c_0, c_0]), f_\rho(x) = f(x), \forall x \in \mathscr{X}\}$$

- Yes! It’s related to Radon transform given that the function is smooth enough. [Ongie et al. ’20]

$$\|f\|_{\mathscr{F}} = \|\mathscr{R}(\Delta^{\frac{d+1}{2}} f)\|_{\mathbb{L}^1(\mathbb{S}^{d-1}\times[-c_0,c_0])} + |f(0)| + \sum_{k=1}^{d} |f(e_k) - f(0)|$$

- Linear functions are the null space for the first term in the norm.

- For a cleaner formulation Ongie et al. considered a semi-norm insensitive to linear functions $\mathscr{P}_1$.

$$\boxed{\|f\|_{\mathscr{R}} = \min_{p \in \mathscr{P}_1} \|f + p\|_{\mathscr{F}} = \|\mathscr{R}(\Delta^{\frac{d+1}{2}} f)\|_{\mathbb{L}^1(\mathbb{S}^{d-1}\times[-c_0,c_0])}}$$

- The space induced by $\mathscr{R}$-norm remains the same as $\mathscr{F}$.

$$\mathscr{F} = \{f : \mathscr{X} \to \mathbb{R} : \|f\|_{\mathscr{R}} < \infty\} = \{f : \mathscr{X} \to \mathbb{R} : \|f\|_{\mathscr{F}} < \infty\}$$

Properties of $\mathscr{R}$-Norm

Lemma. [Parhi et al. '19]
For a discrete neural network with distinct weights $(b^{(i)}, c^{(i)}) \neq \pm\,(b^{(j)}, c^{(j)})$ the $\mathscr{R}$-norm corresponds to ℓ_1-norm of the top layer weights.

$$\left\| \sum_{i=1}^{m} a^{(i)} \sigma(b^{(i)\top} x + c) \right\|_{\mathscr{R}} = \sum_{i=1}^{m} |\, a^{(i)} |$$

Theorem. [Parhi et al. '19]
For a Lipschitz univariate function $f \in \mathrm{Lip}([-c_0, c_0])$ the $\mathscr{R}$-norm corresponds to the total variation of its weak first derivative.

$$\|f\|_{\mathscr{R}} = \|f'\|_{\mathrm{TV}} = \sup\{ \sum_{i=1}^{r} |\, f'(t_i) - f'(t_{i-1}) \,| : -c_0 \le t_0 < t_1 < \ldots < t_r \le c_0 \}$$

- When the function has second derivative then $\|f\|_{\mathscr{R}} = \|\mathscr{R}(\Delta^{\frac{d+1}{2}} f)\|_{\mathbb{L}^1(\mathbb{S}^{d-1} \times [-c_0, c_0])} = \|f''\|_{\mathbb{L}^1(\mathscr{X})} = \|f'\|_{\mathrm{TV}}$.

- Draws connection to spline theory.

(15 min)

How does this minimization connects to adaptivity? 🤪

What is this $\mathscr{R}$-norm? 🤔

$$\min_{f:\mathscr{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu)} + \lambda \|f\|_{\mathscr{R}}$$

When the $\mathscr{R}$-norm is not adaptive? (Our results) 🤯

Properties of the Function Space

(Representation Theorem)

- How does the solutions to the infinite dimensional optimization problem look like?

$$\min_{f:\mathcal{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu_n)} + \lambda\|f\|_{\mathcal{R}}$$

$$\|f\|_{\mathcal{R}} = \inf\{\, |\rho| : f(x) = \int \sigma(b^\top x + c)\rho(db, dc) + b_0^\top x + c_0, \forall x \in \mathcal{X}\,\}$$

- The functions space $(\mathcal{F}, \|\cdot\|_{\mathcal{R}})$ is a (non-Hilbertian) **Banach** space. [Siegel et al. ’22][Parhi et al. ’21]

Theorem. [Rosset et al. ’07][Parhi et al. ’21]
For any regularization parameter $\lambda \in [0,\infty)$ there exists a finite network with $m \leq n + 1 - d$ neurons and parameters $\hat{\theta}_\lambda \in \Theta^m$ which attains the minimum.

$$f_{\hat{\theta}_\lambda} \in \arg\min_{f:\mathcal{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu_n)} + \lambda\|f\|_{\mathcal{R}}$$

Adaptivity to Low Dimensional Structure

[Bach ’17][Parhi et al. ’22]

- Consider the non-parametric regression setting $y = f^*(x) + \epsilon$ and denote $(x, y) \sim \nu$.

	Functional form	Generalization bounds	Nonparametric rates
	$\mathscr{G}$	$\inf_{\lambda} \sup_{f^* \in \mathscr{G}} \|f^* - f_{\hat{\theta}_\lambda}\|_{\mathbb{L}^2(\nu)}$	$\inf_{\hat{f}} \sup_{f^* \in \mathscr{G}} \|f^* - \hat{f}\|_{\mathbb{L}^2(\nu)}$
No assumption		$\tilde{O}(n^{-\frac{1}{d+3}})$	$\tilde{\Theta}(n^{-\frac{1}{d}})$
Projection Pursuit	$\sum_{j=1}^{k} g_j(w_j^\top x), w_j \in \mathbb{R}^d$	$\tilde{O}(k\sqrt{d}n^{-\frac{1}{4}})$	$\tilde{\Theta}(n^{-\frac{2}{3}})$
Dependence on subspace	$g(W^\top x), W \in \mathbb{R}^{k \times d}$	$\tilde{O}(\sqrt{d}n^{-\frac{1}{k+3}})$	$\tilde{\Theta}(n^{-\frac{2}{k+2}})$
Bounded norm	$\|g\|_{\mathscr{R}} \leq B$	$\tilde{O}(n^{-\frac{d+3}{2d+3}})$	$\tilde{\Theta}(n^{-\frac{d+3}{2d+3}})$

Adaptivity (Ctd.)

- How do neural networks achieve such adaptivity?

$$\min_{f:\mathscr{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu_n)} + \lambda \|f\|_{\mathscr{R}}$$

- The $\mathscr{R}$-norm is adaptive to low dimensional structure:
  - For a symmetric domain $\mathscr{X}$ and projection matrix $W \in \mathbb{R}^{k\times d}, W^\top W = I_{k\times k}$

$$\boxed{\forall x \in \mathscr{X} \quad f(x) = g(Wx) \Rightarrow \|f\|_{\mathscr{R}} = \|g\|_{\mathscr{R}}}$$

  - If the neural net could find the low dimensional subspace then it easily can achieve optimal nonparametric rates.
  - One might speculate that **such minimizers** with dependence on a **low dimensional subspace** might exists when such structure is present in the true regression function.

Ridge Functions

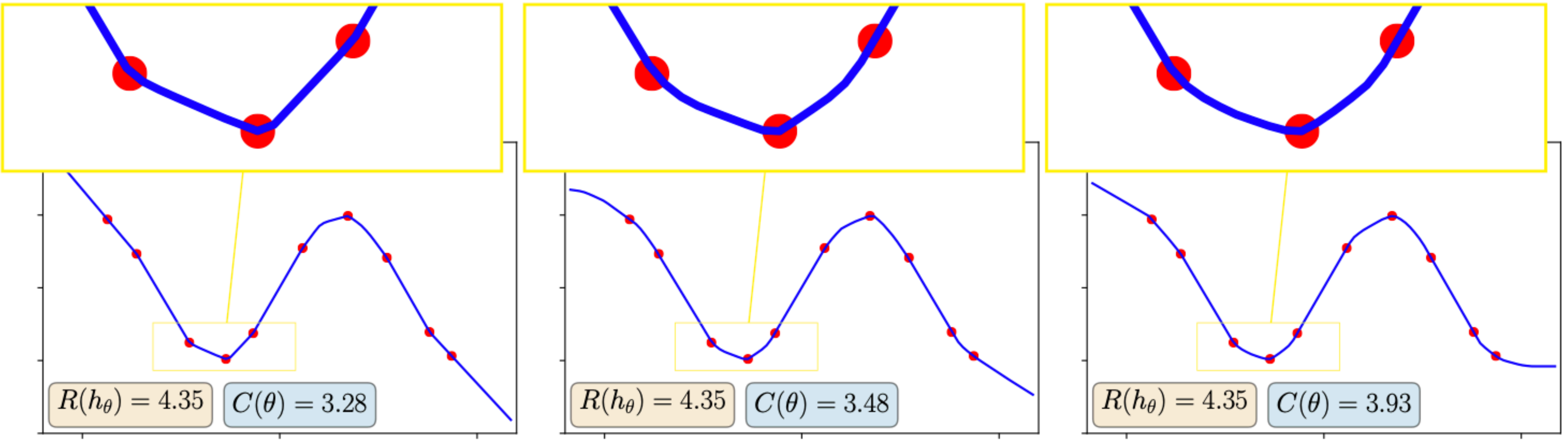
- How does fitting data while minimizing $\mathscr{R}$-norm look like?

- For multivariate functions finding even **one** solution is difficult in general, but ridge functions can be reduced to univariate case.

- For univariate functions:
  - Linear spline is always a solution [Savarese et al. '19]
  - Hanin '21 characterized all the possible solutions.

$$\arg\min\{\|f\|_{\mathscr{R}} : f(x_i) = y_i, i \leq n\}$$

$$\arg\min\{\|f'\|_{\mathrm{TV}} : f(x_i) = y_i, i \leq n\}$$



3 layers, 100 units

4 layers, 100 units

5 layers, 100 units

[Savarese '19]

Ridge Functions

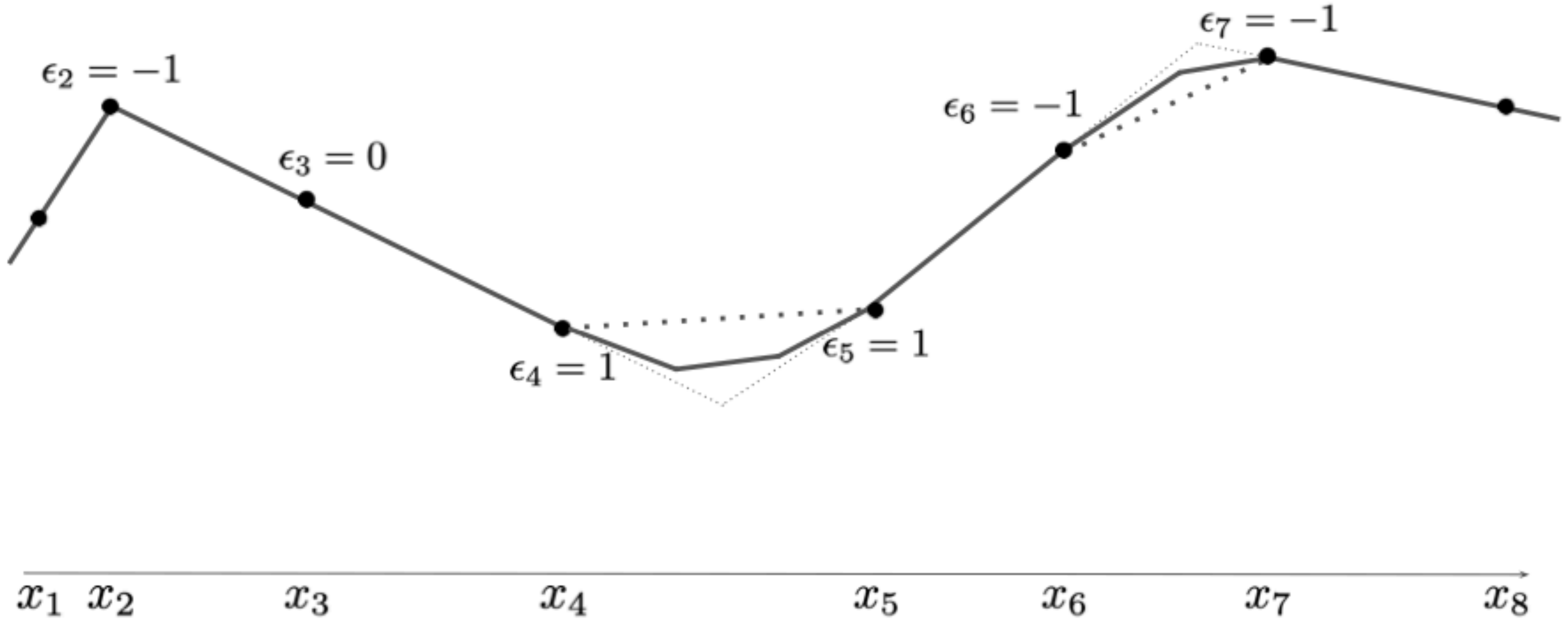
- How does fitting data while minimizing $\mathscr{R}$-norm look like?

- For multivariate functions finding even **one** solution is difficult in general, but ridge functions can be reduced to univariate case.

- For univariate functions:
  - Linear spline is always a solution [Savarese et al. ’19]
  - Hanin ’21 characterized all the possible solutions.

$$\arg\min\{\|f\|_{\mathscr{R}} : f(x_i) = y_i, i \leq n\}$$

$$\arg\min\{\|f'\|_{\text{TV}} : f(x_i) = y_i, i \leq n\}$$



[Hanin ’21]

How does this minimization connects to adaptivity? 🤪

What is this $\mathscr{R}$-norm? 🤔

$$\min_{f:\mathcal{X}\to\mathbb{R}} \|y - f(x)\|^2_{\mathbb{L}^2(\nu)} + \lambda \|f\|_{\mathscr{R}}$$

When the $\mathscr{R}$-norm is not adaptive? (Our results) 🤯

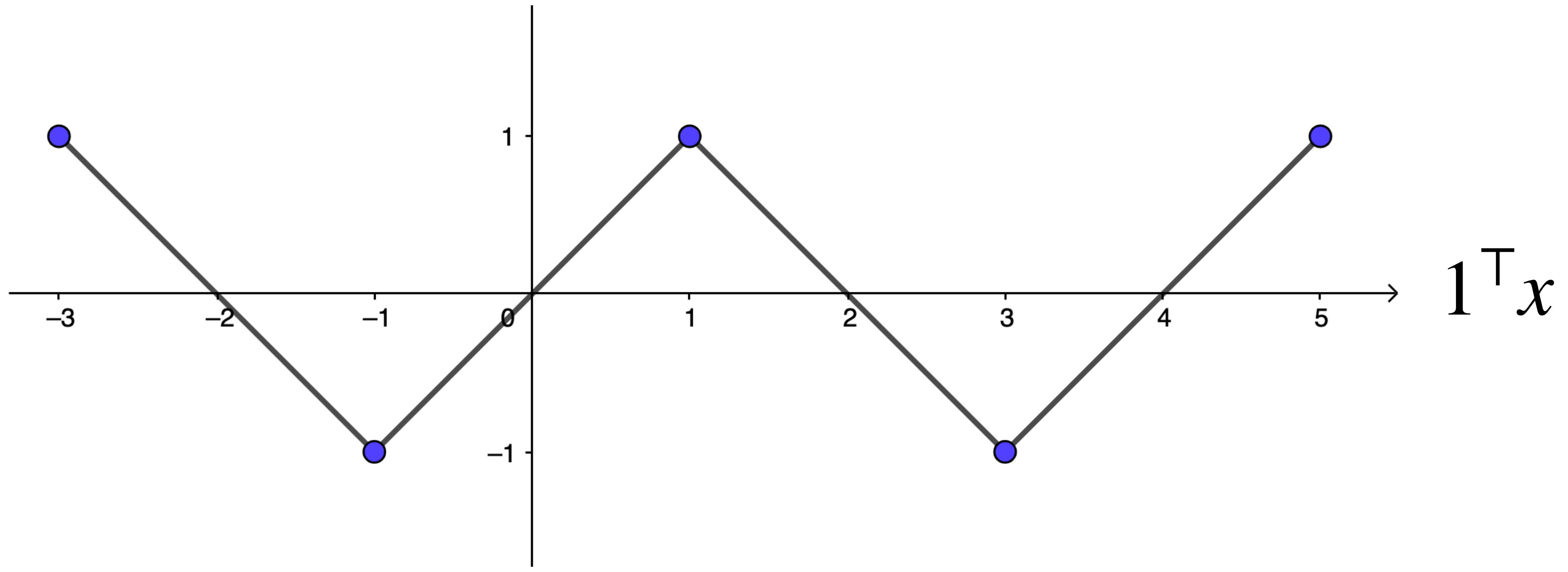
(20 min)

Parity Dataset

- Consider the distribution $(x, y) \sim \nu$ where $x \sim \text{Unif}\{\pm 1\}^d$ is sampled uniformly from the hypercube and labeled $y = \chi(x) = \Pi_{j=1}^d x_j$ where χ is defined over the bounded domain $\mathscr{X} = \sqrt{d}\mathbb{B}_{\ell_2}^d(1)$.

- Parity data can be represented exactly by ridge functions.

$$\forall x \in \{\pm 1\}^d \quad \chi(x) = g(1^\top x) = \sum_{j=1}^{d} x_j \text{ mod 2.}$$



Theorem. [Our work]

Though the parity function χ can be represented by ridge functions but $\mathscr{R}$-norm minimizers which fits the parity data are not ridge functions.

$$\Theta(d^{\frac{3}{2}}) = \inf\left\{\|f\|_{\mathscr{R}} : f \in \text{Ridge}_d,\ \|\chi - f\|_{\mathbb{L}^\infty(\nu)} \le \frac{1}{2}\right\} \gg_d \inf\left\{\|f\|_{\mathscr{R}} : \|\chi - f\|_{\mathbb{L}^\infty(\nu)} = 0\right\} = \Theta(d)$$

Proof Ideas

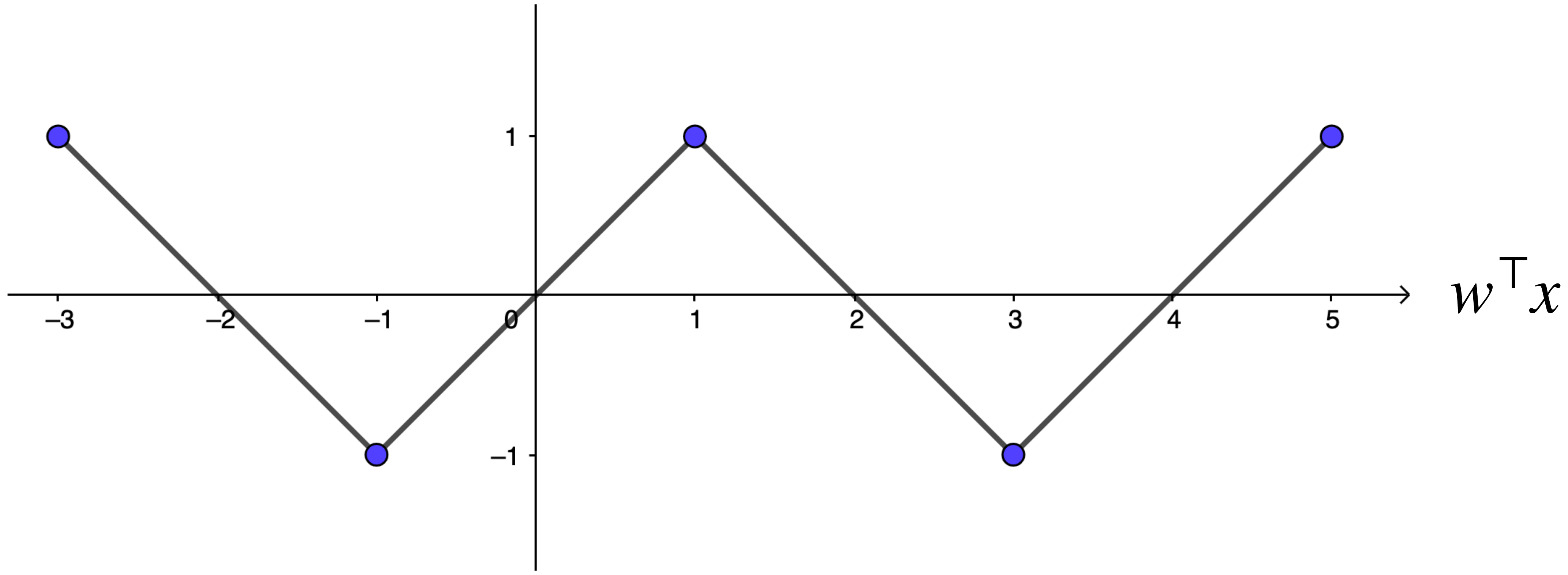
(Ridge Functions)

Theorem.
For ridge functions $f : \mathscr{X} \to \mathbb{R} : x \mapsto = g(w^\top x)$ the best achievable rate is $d^{\frac{3}{2}}$.

$$\inf \left\{ \|f\|_{\mathscr{R}} : f \in \text{Ridge}_d,\ \|\chi - f\|_{\mathbb{L}^\infty(\nu)} \le \frac{1}{2} \right\} = \Theta(d^{\frac{3}{2}})$$

- **Upper Bound:** Parity data can be represented exactly by ridge functions:
- For $w \in \{\pm 1\}^d$ we have $\chi(x) = \sum_{j=1}^{d} x_j = \sum_{j=1}^{d} w_j x_j$ mod 2.
- We need d ReLUs to construct parity..

$$\|\chi\|_{\mathscr{R}} = O(d\sqrt{d})$$



Proof Ideas

(Ridge Functions)

Theorem.
For ridge functions $f : \mathcal{X} \to \mathbb{R} : x \mapsto = g(w^\top x)$ the best achievable rate is $d^{\frac{3}{2}}$.

$$\inf \left\{ \|f\|_{\mathcal{R}} : f \in \text{Ridge}_d,\ \|\chi - f\|_{\mathbb{L}^\infty(\nu)} \leq \frac{1}{2} \right\} = \Theta(d^{\frac{3}{2}})$$

- **Lower Bound:** For ridge functions

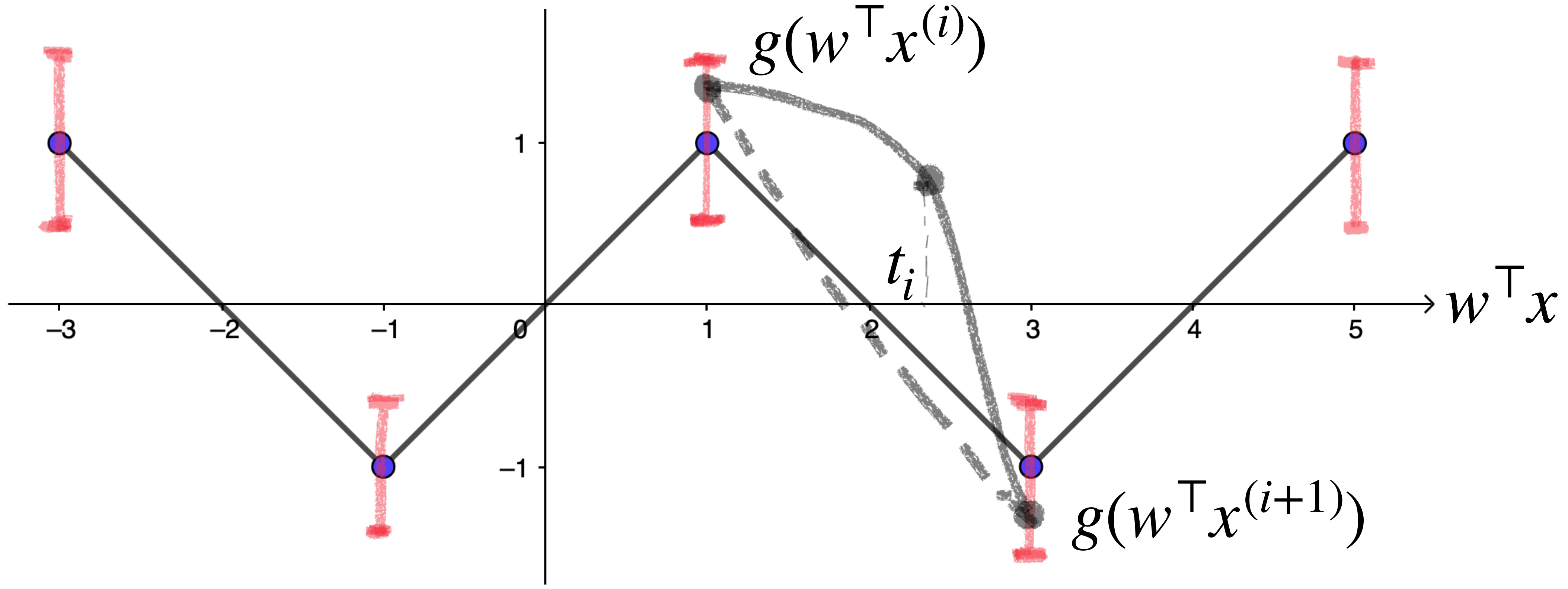
$$\|f\|_{\mathcal{R}} = \|g'\|_{\text{TV}} = \sup_{-\sqrt{d} \leq t_0 < t_1 < \ldots < t_r \leq \sqrt{d}} \sum_{i=1}^{r} |g'(t_i) - g'(t_{i-1})|$$

- Take the subset of points for which $\chi(x^{(i)}) \neq \chi(x^{(i+1)})$,

$$x^{(i)} = (\text{Sign}(w_1), \ldots, \text{Sign}(w_i), -\text{Sign}(w_{i+1}), \ldots, -\text{Sign}(w_d))$$

- Using mean value theorem choose t_i such that

$$|g'(t_i)| \geq \frac{1}{2} \left| \frac{g(w^\top x^{(i+1)}) - g(w^\top x^{(i)})}{w^\top x^{(i+1)} - w^\top x^{(i)}} \right|$$



Proof Ideas

(Multi-index Functions)

Theorem.
For function $f : \mathcal{X} \to \mathbb{R}$ that fits the parity data

$$\inf \left\{ \|f\|_{\mathscr{R}} : \|\chi - f\|_{\mathbb{L}^\infty(\nu)} = 0 \right\} = \Theta(d)$$

- **Upper Bound:** Use an averaging technique to combines a collection of distinct ridge functions.

$$s_w(x) = \chi(x)\mathbf{1}\{w^\top x = 0\}$$

- Pick $w \in \{\pm 1\}^d$ randomly and take the scaled average sawtooth function in that direction.

$$f(x) = \frac{\mathbb{E}[s_w(x)]}{\mathbf{P}[w^\top x = 0]} = \frac{1}{\binom{d}{d/2}} \sum_{w \in \{\pm 1\}^d} s_w(x) \approx \frac{\sqrt{d}}{2^d} \sum_{w \in \{\pm 1\}^d} s_w(x)$$

- This function fits the parity data.

$$f(x) = \chi(x), \forall x \in \{\pm 1\}^d$$

- Since $\mathscr{R}$-norm satisfies triangle inequality:

$$\|f\|_{\mathscr{R}} \lesssim \frac{\sqrt{d}}{2^d} \sum_{w \in \{\pm 1\}^d} \|s_w\|_{\mathscr{R}} \leq \frac{\sqrt{d}}{2^d} \sum_{w \in \{\pm 1\}^d} \|w\|_2 = d$$

Proof Ideas

(Multivariate Functions)

Theorem.

For function $f : \mathcal{X} \to \mathbb{R}$ that fits the parity data

$$\inf \left\{ \|f\|_{\mathscr{R}} : \|\chi - f\|_{\mathbb{L}^\infty(\nu)} = 0 \right\} = \Theta(d)$$

- **Lower Bound:** Utilizes the fact a fixed ReLU neuron cannot much correlate with parity in $\mathbb{L}^2(\nu)$.

- By definition any function with finite $\mathscr{R}$-norm admits an integral representation.

$$f(x) = \int \sigma(b^\top x + c)\rho(db, dc) + b_0^\top x + c_0$$

- Linear terms does not correlate with parity

$$1 = \mathbb{E}_{x\sim\nu}[f(x)\chi(x)] = \int \mathbb{E}_{x\sim\nu}[\sigma(b^\top x + c), \chi(x)]\rho(db, dc) \leq |\rho| \sup_{b,c} \mathbb{E}_{x\sim\nu}[\sigma(b^\top x + c), \chi(x)]$$

Recap

- $\mathscr{R}$-norm is a functional norm with connections to neural network **training**.
- $\mathscr{R}$-norm regularization which enjoys **adaptivity** to low dimensional structure.
- Our results demonstrate some **limitations** of that for certain data distributions.

Other Contributions

- We further study the generalization properties of $\mathscr{R}$-norm interpolators.
    - When $n = \tilde{\Omega}(d^3)$ with high probability all minima **approximates** the parity well.
    - When $n = \tilde{o}(d^2)$ with constant probability all minima are **far** from the parity function.
- This separation phenomenon between ridge and multidirectional functions remains for other distributions analogous to parity.
- Experiments indicating training with SGD prefer low variational norm functions.
    - Comparing architectures which forces low dimensional structure as opposed to fully connected nets.

Thank you!