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

Clayton Sanford* Navid Ardeshir* Daniel Hsu

Columbia University *Equal contribution

Scan Me!

Abstract

Our Problem: We study **statistical** and **approximation** properties of **interpolating** two layer ReLU networks with small **variational norm** ($\mathcal{R}$-norm).

- This norm captures the functional effect of **controlling the size of network weights**.
- This allows the network width to be **unbounded**.
- Practically motivated:
  - Correspond to **weight decay** regularization in neural network training.
  - It has connections to **implicit bias of GD** in the feature learning regime.
- It is known that neural networks trained with **optimal weight decay regulartization** can be **adaptive to low dimesnional structure.**

Our Findings: For **certain target distributions**, minimum $\mathcal{R}$-norm interpolants are:

1. **Intrinsically multivariate functions** (vary in many directions), even when there are ridge functions (vary in only one direction) that fit the data.
2. **Statistically sub-optimal** in terms of generalization.

Bounded Norm Neural Networks

Model: Suppose the data consist of n samples $(\mathbf{x}_i, \mathbf{y}_i)_{i \le n} \sim \nu \in \mathcal{P}(\Omega \times \mathbb{R})$, where $\Omega \subseteq \mathbb{R}^d$ is a spherically symmetric bounded domain. Let $\boldsymbol{\nu}_n$ denote the empirical data distribution.

Euclidean Formulation: Consider two layer **ReLU** neural networks, with width m, a skip connection, and parameters $\theta = (a_i, b_i, c_i)_{i \le m} \in (\mathbb{R} \times \mathbb{R}^d \times \mathbb{R})^m$,

$$f_\theta : \Omega \to \mathbb{R} : x \mapsto \sum_{i=1}^{m} a_i \left(b_i^\top x + c_i\right)_+ + a_0 \left(b_0^\top x + c_0\right).$$

The $\mathcal{R}$-norm of a function $f : \Omega \to \mathbb{R}$ is the **minimum cost** of approximating it arbitrary well by two layer ReLU networks,

$$\|f\|_{\mathcal{R}} := \lim_{\epsilon \to 0} \inf_{m,\, \theta} C(\theta) := \frac{1}{2} \sum_{i=1}^{m} |a_i|^2 + \|b_i\|_2^2 \quad \text{s.t.} \quad \|f - f_\theta\|_{\mathbb{L}^\infty(\Omega)} \le \epsilon$$

Note that the infimum is over both width, and network parameters.

Problem: What are properties of $\mathcal{R}$-norm inductive bias for certain target distributions?

$$\inf_{f:\Omega \to \mathbb{R}} \|f\|_{\mathcal{R}} \quad \text{s.t.} \quad f(x) = y \quad \nu\text{-almost everywhere} \tag{1}$$

- **Statistical:** What is the required sample complexity (if we replace ν with $\boldsymbol{\nu}_n$)?
- **Approximation:** What do solutions to (1) look like?

Properties of $\mathcal{R}$-norm

Representer Theorem: Though $\mathcal{R}$-norm is **not a RKHS norm**, [7] showed **a minimizer** of the variational problem exists with width $m \le n$,

$$\forall \epsilon \ge 0 \quad f_{\hat{\theta}_\epsilon} \in \arg\min_{f:\Omega \to \mathbb{R}} \|f\|_{\mathcal{R}} \quad \text{s.t.} \quad \|y - f(x)\|_{\mathbb{L}^2(\boldsymbol{\nu}_n)} \le \epsilon \tag{2}$$

Characterizing the Norm and Variational Problem: Though $\mathcal{R}$-norm is a variational norm, it can be explicitly characterized in terms of the functions itself under mild assumptions:

1. **Univariate Functions:**
   - For $d = 1$, [9] showed $\|f\|_{\mathcal{R}} = \|f''\|_{\mathbb{L}^1(\Omega)} = \int_\Omega |f''(x)|\, dx$.
   - [4, **?**] characterized all the solutions to the variational problem (1).
2. **Multivatiate Functions:**
   - In general [6] showed that $\mathcal{R}$-norm is related to Radon Transform of **higher order derivatives** of the function.
   - Characterizing even a solution to the variational problem in general is difficult.
   - Recent work [5] do so for rank-one datasets using convex duality.
3. Ridge Functions:
   - For functions that only vary in one direction, it reduces to the univariate case,

$$\exists\, w \in \mathbb{S}^{d-1} \ \forall\, x \in \Omega \quad f(x) = g(w^\top x) \Rightarrow \|f\|_{\mathcal{R}} = \|g\|_{\mathcal{R}}\,.$$

Adaptivity

Curse of dimensionality

- Without any assumption on the data we are doomed to require $n = e^{\Omega(d)}$ number of samples in the in the worst case.
- Inductive biases based on certain variational norms, such as the $\mathcal{R}$-norm, are believed to offer a way around the curse of dimensionality **suffered by kernel methods** [1].
- For optimally chosen ϵ, solutions to (2) can be **adaptive to low dimensional structure** and have sample complexity bounds whose exponent depends on the **intrinsic dimension** [1, 8].
- But how? One may believe that $\mathcal{R}$-norm inductive bias achieves this adaptivity by **favoring functions with low dimensional structure**.
- Empirical/theoretical evidence that neural networks with weight decay regularization can **identify** the low dimensional architecture for certain learning tasks.

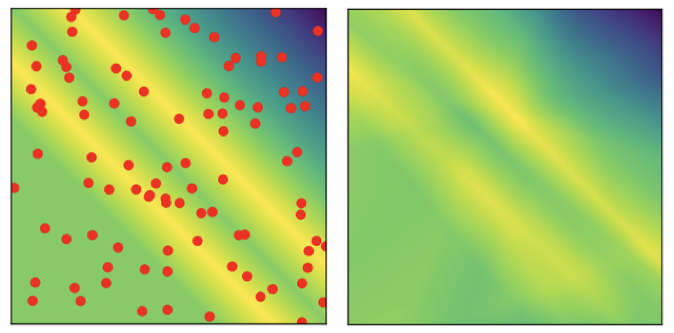


Figure 1. Image from [8]

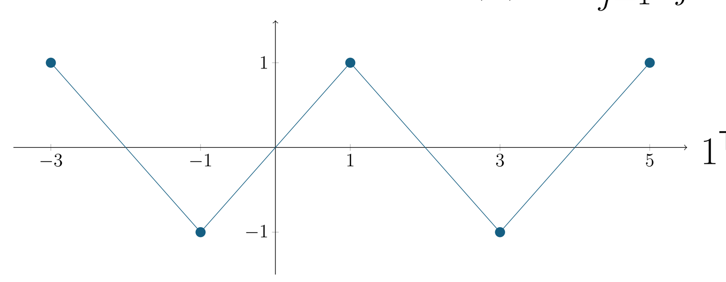
Question: Do minimum $\mathcal{R}$-norm interpolants have a low dimensional structure when such structure is present in the target distribution?

Main Results (Simplified)

Parity Distribution: Consider the target distribution $(\mathbf{x}, \mathbf{y}) \sim \nu \in \mathcal{P}(\{\pm 1\}^d \times \{\pm 1\})$ where $\mathbf{x} \sim \text{Uniform}\{\pm 1\}^d$ is uniformly sampled from **hypercube** and labeled $\mathbf{y} = \chi(\mathbf{x}) = \Pi_{j=1}^{d} \mathbf{x}_j$.

- Parity can be represented by **ridge functions**,

$$\forall x \in \{\pm 1\}^d \quad \chi(x) = g(1^\top x).$$



Approximation

Theorem: For parity distribution $\nu \in \mathcal{P}(\{\pm 1\}^d \times \{\pm 1\})$,

- **Ridge function** approximators suffer **high variational norms**,

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

- **Multidirectional functions** can interpolate more efficiently,

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

- No solution to the variational problem with low-dimensional structure is guaranteed to exist, even when the data distribution has low-dimensional structure.
- Results can be extended to distributions other than parity (see paper).

Generalization

Theorem: Given n samples from parity distribution $\nu \in \mathcal{P}(\{\pm 1\}^d \times \{\pm 1\})$,

$$\hat{\mathcal{F}} = \arg\min_{f:\Omega \to \mathbb{R}} \|f\|_{\mathcal{R}} \quad \text{s.t.} \quad f(\mathbf{x}_i) = \mathbf{y}_i.$$

- **(Upper Bound)** When $n = \tilde{\omega}(d^3)$ all minima **approximates parity well** with high probability.

$$\forall \hat{f} \in \hat{\mathcal{F}} \quad \left\|\chi - \text{clip} \circ \hat{f}\right\|_{\mathbb{L}^2(\nu)} = o(1)$$

- **(Lower Bound)** When $n = \tilde{o}(d^2)$ all minima are **far from parity** with high probability,

$$\forall \hat{f} \in \hat{\mathcal{F}} \quad \left\|\chi - \text{clip} \circ \hat{f}\right\|_{\mathbb{L}^2(\nu)} = 1 - o(1)$$

- Information theoretically $n = \Omega(d)$ is sufficient to learn parity (gaussian elimination).
- $\mathcal{R}$-norm inductive bias is not sufficient to achieve statistically optimal sample complexity for learning parity functions.

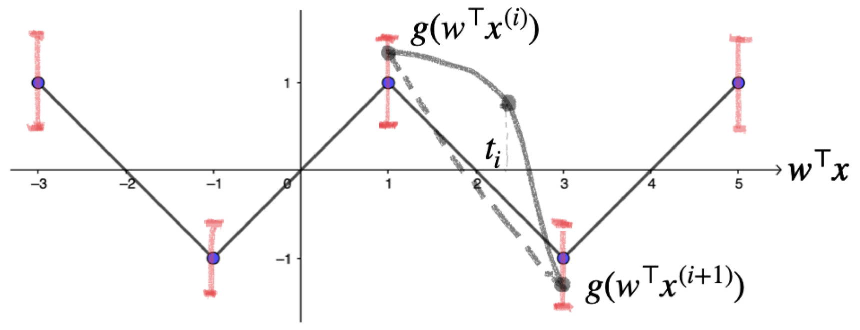
Proof Ideas (Informal)

1. **Approximation:**
   - The $\mathcal{R}$-norm is adaptive to low dimensional structure, e.g. the $\mathcal{R}$-norm of a ridge function is equivalent to its univariate function,

$$f(x) = g(w^\top x) \Rightarrow \|f\|_{\mathcal{R}} = \|w\|\, \|g\|_{\mathcal{R}} = \|w\|\, \|g''\|_{\mathbb{L}^1(\Omega)} = \|w\|\, \|g'\|_{\text{TV}}$$

   - Any ridge function that approximates parity alternates between ± 1 values at least d times.
   - Through a careful usage of the mean value theorem its tangent slope must alternate $\pm\Theta(\sqrt{d})$ at least d times,

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



   - For the upper bound we employ an **averaging strategy** that combines a collection of distinct ridge functions, each of which has **few alternations**, and perfectly **fits a fraction** of the parity dataset.

$$f(x) = \frac{1}{2^d} \sum_{w \in \{\pm 1\}^d} \chi(x) \mathbb{1}\left\{w^\top x = 0\right\}$$

2. **Generalization:**
   - For the upper bound we use standard **Rademacher complexity** bounds for **bounded $\mathcal{R}$-norm function class.**
   - For the lower bound we use a "cap construction" from [2] to produce a robust network with **small Lipschitz and $\mathcal{R}$-norm** $\tilde{O}(\frac{n}{d})$ interpolating the n samples.

References

[1] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

[2] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, 2021.

[3] Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Joan Bruna. Finding the needle in the haystack with convolutions: On the benefits of architectural bias. In *Advances in Neural Information Processing Systems 32*, 2019.

[4] Thomas Debarre, Quentin Denoyelle, Michael Unser, and Julien Fageot. Sparsest piecewise-linear regression of one-dimensional data. *Journal of Computational and Applied Mathematics*, 406:114044, 2022.

[5] Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *Journal of Machine Learning Research*, 22(212):1–63, 2021.

[6] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International Conference on Learning Representations*, 2019.

[7] Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

[8] Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *arXiv preprint arXiv:2109.08844*, 2021.

[9] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, 2019.

https://arxiv.org/pdf/2206.05317.pdf

navid.ardeshir@columbia.edu ,clayton@cs.columbia.edu ,djhsu@cs.columbia.edu