"Optimal" Nonparametric Regression on Low Dimensional Manifolds using Deep ReLU Neural Networks

Minshuo Chen, Haoming Jiang, Wenjing Liao, Tuo Zhao

Georgia Tech

Nonparametric Regression

Given $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $\boldsymbol{x}_i \in \mathbb{R}^D$ i.i.d. sampled from a distribution $\mathcal{D}$, we observe

$$y_i = f^*(\boldsymbol{x}_i) + \epsilon,$$

- $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. from $N(0, 1)$;
- $f^* \in \mathcal{F}$ with $\mathcal{F}$ being a class of smooth functions, e.g., Hölder, Sobolev and Besov spaces.

Information-Theoretic Lower Bound:

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\boldsymbol{X}|\mathcal{S}} (\widehat{f}(\boldsymbol{X}) - f^*(\boldsymbol{X}))^2 \asymp n^{-\frac{2(s+\alpha)}{2(s+\alpha)+D}},$$

- $\boldsymbol{X} \sim \mathcal{D}$ and $\mathcal{F}$ is Hölder class $\mathcal{H}^{s,\alpha}$.

Theory VS Practice

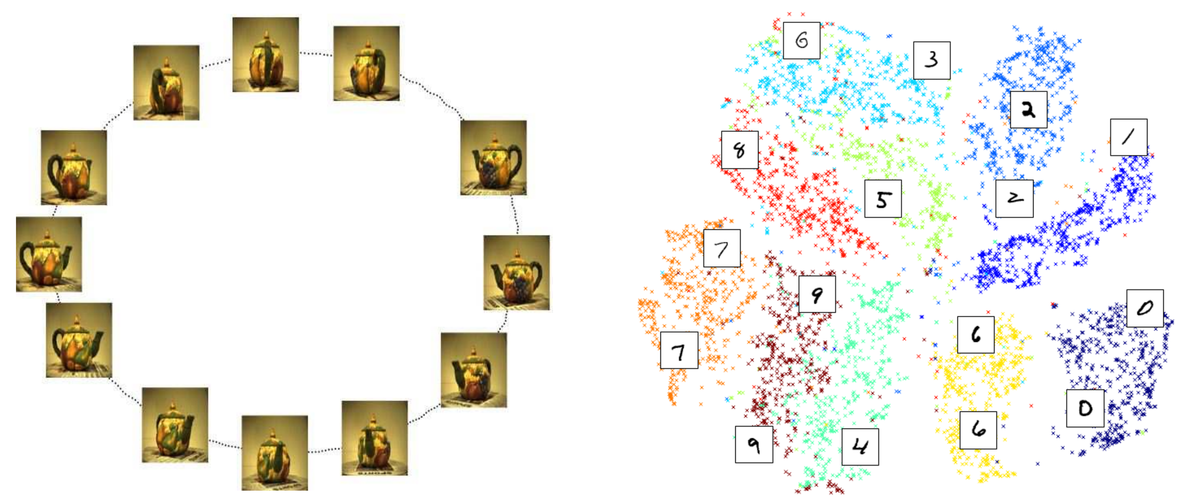
Object Recognition on ImageNet:

- ImageNet: 1000000 images in 1000 categories;
- Image Resolution: 384×384;
- Empirical Performance: Top 1 Accuracy 86.4% and Top 5 Accuracy 98.0%;
- Theoretical Bound: $n_{\text{theory}} \gtrsim \epsilon^{-D/(s+\alpha)}$;
- For moderate s and α, $n_{\text{theory}} \gg 1,000,000$.

Question:
- Why does there exist such a **huge gap** between theory and practice?

Low Dimensional Structures

▷ **Practical Motivation**: Images and acoustic signals exhibit low dimensional structures.

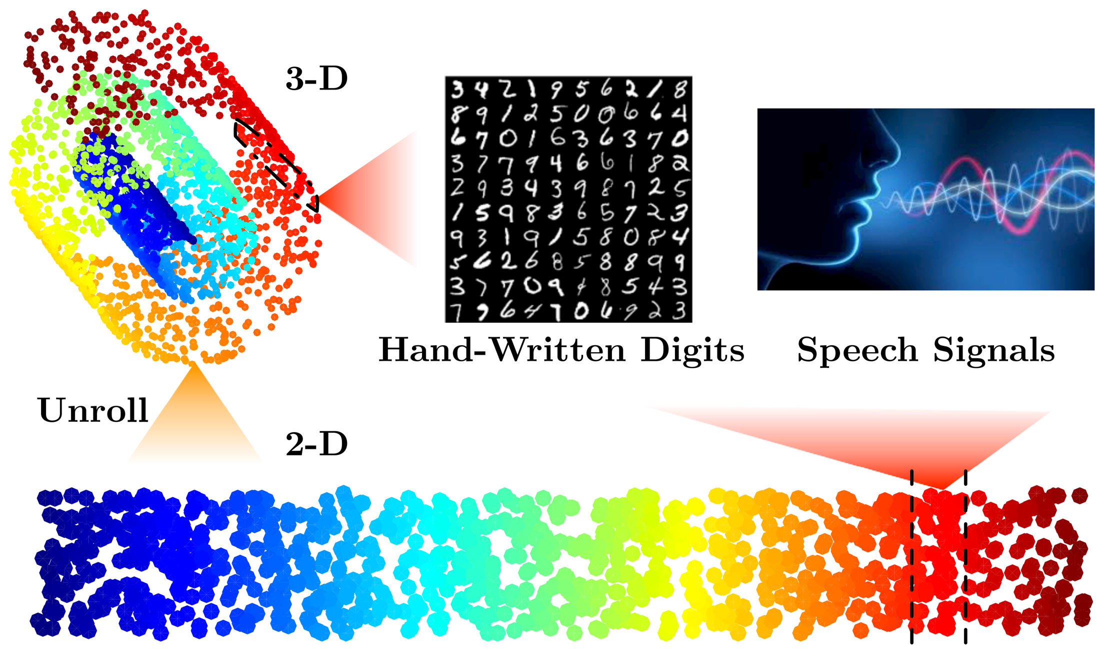


▷ **Key Observation**: Data intrinsic dimension is much smaller than the ambient dimension — making statistical estimation manageable.

Model data using a low dimensional manifold.

Regression on Low Dimensional Manifolds

▷ **Low Dimensional Smooth Manifolds**

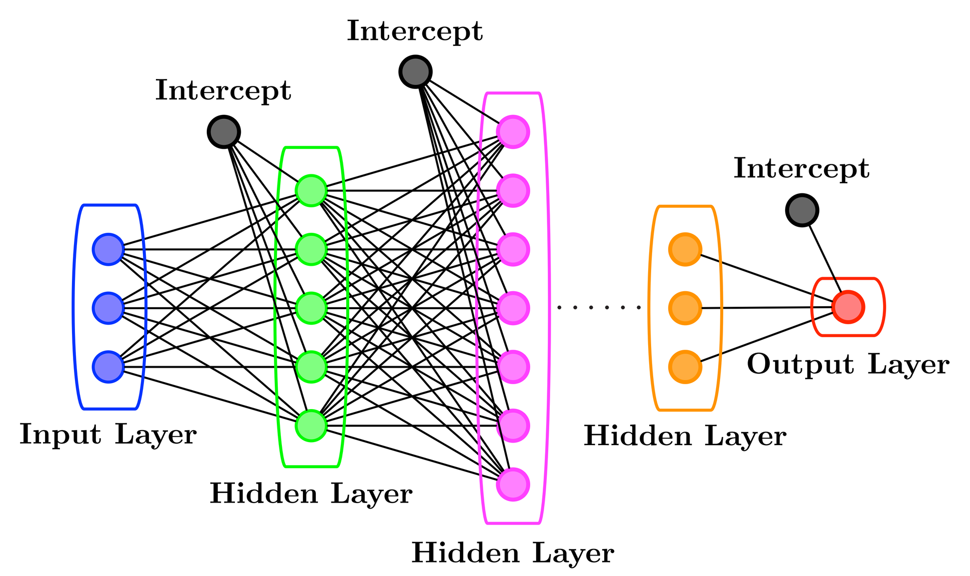


▷ **Regression on Low Dimensional Manifolds**

- $\boldsymbol{x}_i$ sampled from a d-dimensional manifold $\mathcal{M}$ ($d \ll D$);
- $\mathcal{F}$ is Hölder on $\mathcal{M}$ (distinguish from Hölder on $\mathbb{R}^D$).
- We learn $\widehat{f}$ by minimizing the empirical ℓ_2 risk, i.e.,

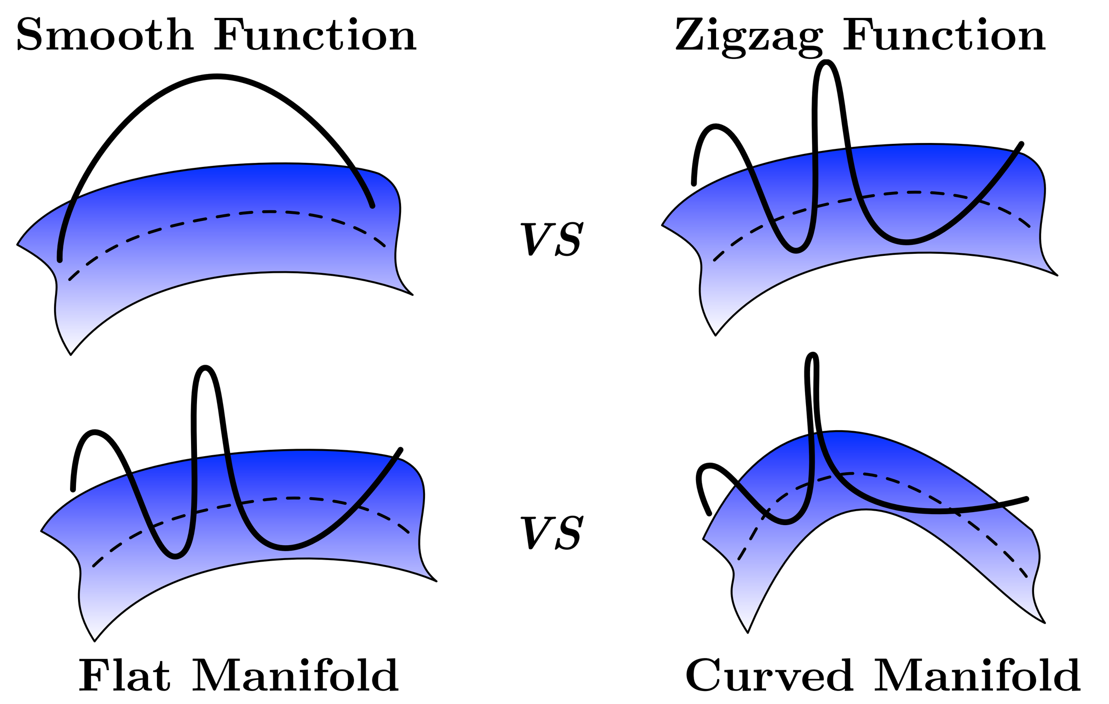
$$\widehat{f} = \operatorname*{argmin}_{f \in \mathcal{F}_{\mathsf{NN}}} \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - y_i)^2, \qquad (1)$$

where $\mathcal{F}_{\mathsf{NN}}$ denotes a class of ReLU neural networks.



$f(\boldsymbol{x}) = \boldsymbol{W}_L \sigma(\cdots \sigma(\boldsymbol{W}_2 \sigma(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) \cdots) + \boldsymbol{b}_L,\ \sigma(\cdot) = \max\{0, \cdot\}$

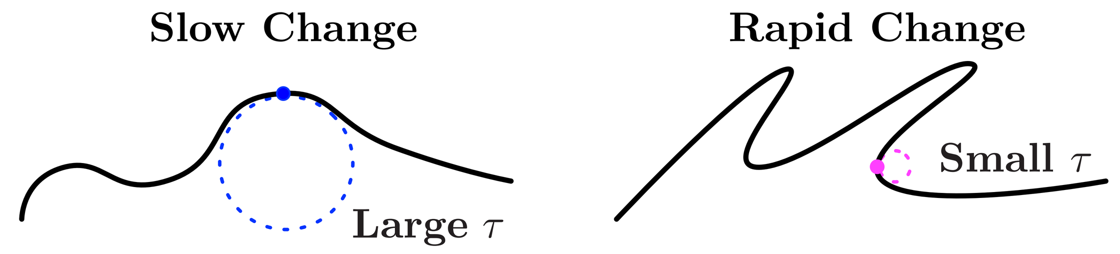
▷ **Size of ReLU Neural Networks in $\mathcal{F}_{\mathsf{NN}}$**



Highly depends on **smoothness** of f^*, **curvature** of $\mathcal{M}$.

Assumptions

- $\mathcal{M}$ is compact with reach bounded by τ.



- f^* belongs to the Hölder ball $\mathcal{H}^{s,\alpha}_{\mathcal{M}}$ of radius 1 on $\mathcal{M}$.

$$\left| D^{\boldsymbol{s}}(f^* \circ \phi^{-1})\big|_{\phi(\boldsymbol{x}_1)} - D^{\boldsymbol{s}}(f^* \circ \phi^{-1})\big|_{\phi(\boldsymbol{x}_2)} \right| \leq \|\phi(\boldsymbol{x}_1) - \phi(\boldsymbol{x}_2)\|_2^{\alpha}.$$

Statistical Recovery

▷ **Architecture: Sparse ReLU Network**

Theorem 1 (Bias Characterization). *For any given $\epsilon \in (0, 1)$, there exists a ReLU network architecture with*

- *no more than $\mathcal{O}(\log \frac{1}{\epsilon} + \log D)$ layers;*
- *at most $\mathcal{O}(\epsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D)$ neurons and weight parameters,*

such that for any $f^* \in \mathcal{H}^{s,\alpha}_{\mathcal{M}}$, the network with properly chosen parameters yields $\widetilde{f}$ satisfying

$$\|\widetilde{f} - f^*\|_\infty \leq \epsilon.$$

$\Big\Downarrow$ **Bias-Variance Trade-off**

▷ **Statistical Recovery Guarantee**

Theorem 2 (Estimation Error). *Suppose that $\widehat{f}_n$ minimizes the empirical risk* (1)*, where*

$\mathcal{F}_{\mathsf{NN}} = \{f \mid f$ *is an L-layer ReLU network with width bounded by p,* $\|f\|_\infty \leq R$, $\|\boldsymbol{W}_j\|_\infty \leq \kappa$, $\|\boldsymbol{b}_j\|_\infty \leq \kappa$ *and* $\sum_{j=1}^{L} \|\boldsymbol{W}_j\|_0 + \|\boldsymbol{b}_j\|_0 \leq K\}$.

Let $L = \mathcal{O}\big(\frac{(s+\alpha)\log n}{d+2s+2\alpha}\big)$, $p = \mathcal{O}\big(n^{\frac{d}{d+2s+2\alpha}}\big)$ *and* $K = \mathcal{O}(Lp)$,

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\boldsymbol{X}|\mathcal{S}} \big(\widehat{f}_n(\boldsymbol{X}) - f^*(\boldsymbol{X})\big)^2 = \mathcal{O}\Big(n^{-\frac{2(s+\alpha)}{2s+2\alpha+d}} \log^3 n\Big).$$

$\Big\Downarrow$ **Optimal?**

▷ **Minimax Lower Bound**

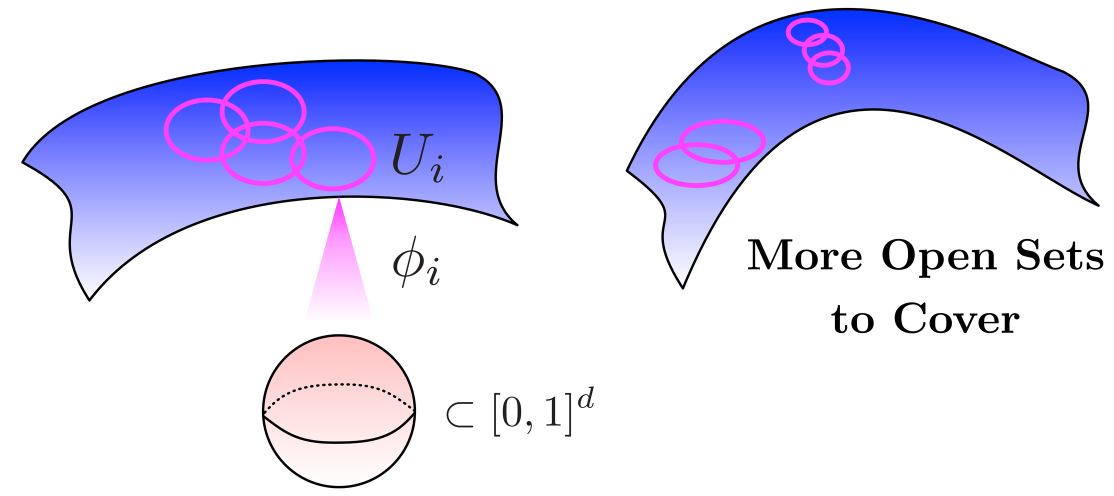
Conjecture 3 (Common Belief). *Given $\boldsymbol{X} \sim \mathcal{D}$ supported on $\mathcal{M}$, we have*

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{H}^{s,\alpha}_{\mathcal{M}}} \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\boldsymbol{X}|\mathcal{S}} (\widehat{f}(\boldsymbol{X}) - f^*(\boldsymbol{X}))^2 = \Omega\Big(n^{-\frac{2(s+\alpha)}{2(s+\alpha)+d}}\Big).$$

Long lasting **belief** for nonparametric regression on low dimensional manifolds — no rigorous proof.

Construct Sparse ReLU Network

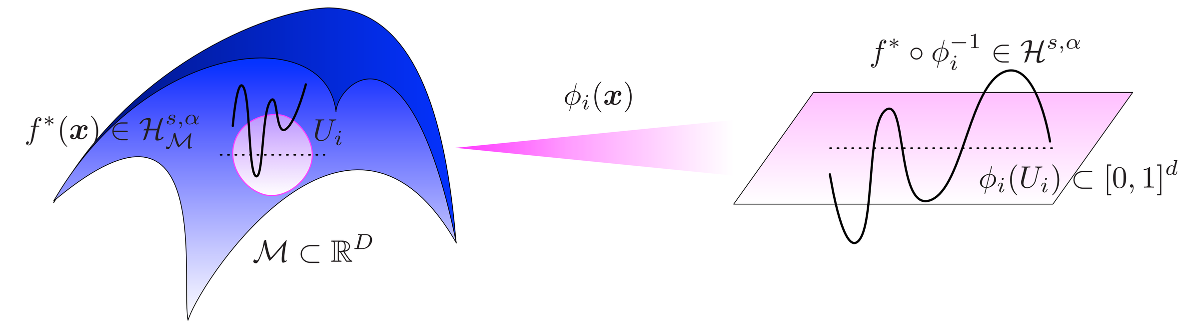
▷ **Step I: Chart Construction**



- $\mathcal{M} = \bigcup_{i=1}^{C_{\mathcal{M}}} U_i$ with $C_{\mathcal{M}} = \left\lceil \frac{\mathrm{Vol}(\mathcal{M}) \cdot d \log d}{\tau^d} \right\rceil$;
- $\phi_i : \mathbb{R}^D \mapsto \mathbb{R}^d$ data transformation.

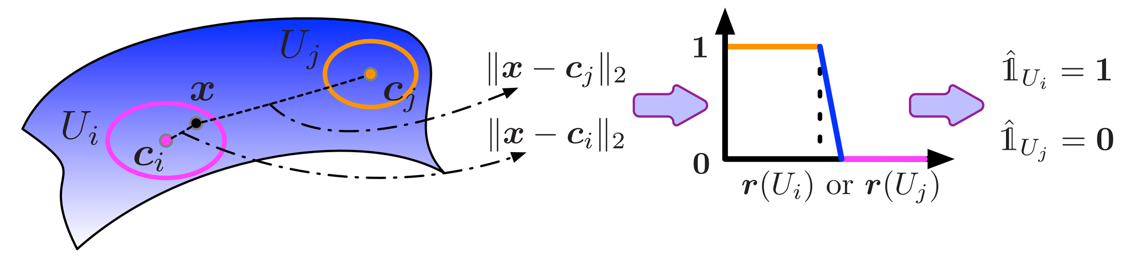
Overlapping Charts $\Longrightarrow$ Balance Contribution (Step III)

▷ **Step II: Local Taylor Approximation**



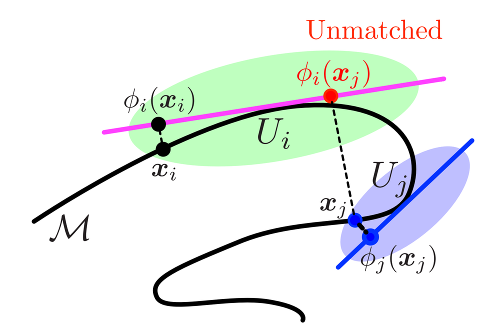
Approximation in $\mathbb{R}^d$ $\Longrightarrow$ Ambient Dimension D Free

▷ **Step III: Chart Determination and Pairing**



Weighted sum to approximate f^*:

$$\widetilde{f} = \sum_{i=1}^{C_{\mathcal{M}}} \eta_i \cdot \widehat{\mathbb{1}}_{U_i} \cdot (P_i \circ \phi_i) \approx f^*,$$



where $\sum_{i=1}^{C_{\mathcal{M}}} \eta_i(\boldsymbol{x}) = 1$ to balance the contribution of local Taylor approximations P_i's.

Discussions

▷ **ReLU activation** vs **Smooth activation**
- Popular in modern applications;
- Mitigate the gradient vanishing issue.

▷ **Feedforward Network** vs **Convolutional Network**
- Conv. filters to extract the low dimensional structures.

▷ **Computational Concerns**
- Global optimum?
- Overparameterization?