Approximation with CNNs in Sobolev Space: with Applications to Classification

Guohao Shen*, Yuling Jiao*, Yuanyuan Lin† and Jian Huang†

NeurIPS 2022

Equal contribution*   Corresponding authors†

CNN Approximation

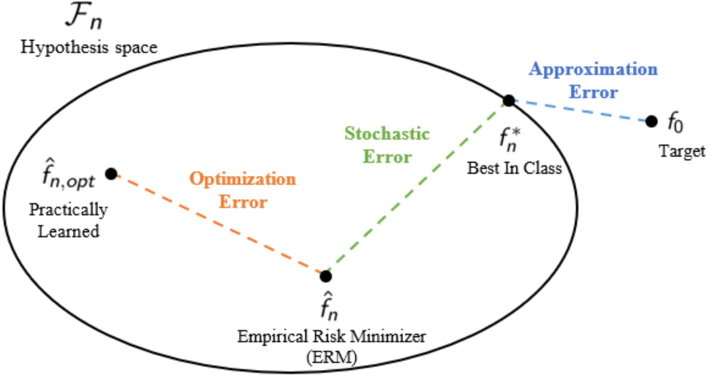


Figure: Approximation, Estimation and Optimization.

What's New

Table: A comparison of some recent CNN approximation results.

	Network	Target function	Flexible filter length	Explicit prefactor	Low-dimensional Result
[7]	CNN	FNN	✗	✗	✗
[6]	ConvResNet	FNN	✓	✗	✗
[9]	CNN	Sobolev	✓	✗	✗
[4]	CNN	Hölder	✗	✗	✓
[5]	ConvResNet	Besov	✓	✗	✓
This paper	CNN	Sobolev and Hölder	✓	✓	✓

Convolutional neural networks

- A CNN $f_{\text{CNN}} : \mathcal{X} \to \mathbb{R}$ with L hidden layers:

$$f_{\text{CNN}}(x) = A_{L+1} \circ A_L \circ \cdots \circ A_2 \circ A_1(x)$$

- Convolutional layers:
  1. $A_i(x) = \sigma(W_i^c x + b_i^c)$ with ReLU activation σ.
  2. Sparse Toeplitz weight matrix $W_i^c \in \mathbb{R}^{d_i \times d_{i-1}}$ induced by
  3. Convolutional filters $\{w_j^{(i)}\}_{j=0}^{s^{(i)}}$ with filter length $s^{(i)} \in \mathbb{N}^+$.
  4. Bias vector $b_i^c \in \mathbb{R}^{d_i}$.
- Downsampling layers
  1. $A_i(x) = D_i(x) = (x_{jm_i})_{j=1}^{\lfloor d_{i-1}/m_i \rfloor}$ for any $x \in \mathbb{R}^{d_{i-1}}$.
  2. Max Pooling, Average Pooling.
  3. Scaling parameter $m_i \leq d_{i-1}$.
- Class of CNNs $\mathcal{F}_{\text{CNN}} = \{f_{\text{CNN}} \text{ over all possible choice of } \{A_i\}_{i=1}^{L+1}\}$.
  1. Total number of parameters $\mathcal{S}$ for networks in $\mathcal{F}_{\text{CNN}}$
  2. Min. and max. filter length $s_{\min}$ and $s_{\max}$ over convolutional layers

Theorem 1 (Approximation on functions in Sobolev Space)

Assume $f \in W^{\beta,p}(\mathcal{X})$ *with* $1 \leq \beta \in \mathbb{N}_0$, $1 \leq p \leq \infty$ *and* $\|f\|_{W^{\beta,p}(\mathcal{X})} \leq B_0$. *For any* $M, N \in \mathbb{N}^+$, *and for* $m = 0, 1$, *there exists a function* $f_{CNN} \in \mathcal{F}_{CNN}$ *with*

$$L \leq 42(\lfloor\beta\rfloor + 1)^2 M\lceil\log_2(8M)\rceil\lceil\frac{\mathcal{W}-1}{s_{\min}-1}\rceil, \quad 2 \leq s_{\min} \leq s_{\max} \leq \mathcal{W}, \quad \mathcal{S} \leq 8\mathcal{W}L,$$

$$\mathcal{W} = 38^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+2} N^2\lceil\log_2(8N)\rceil^2,$$

such that $$\|f - f_{CNN}\|_{W^{m,p}(\mathcal{X})} \leq C_0(d,\beta,p)(NM)^{-2(\beta-m)/d}$$

where $C_0(d,\beta,p) = 37 \cdot 2^{2\beta+2d/p} B_0^2(\beta+1)^3 \times \{\pi^{-d/2}\Gamma(d/2+1)\}^{2/p+1}(1+2\sqrt{d})^d d^{4\beta}$.

- Sobolev class of functions

$$W^{\beta,p}(\mathcal{X}) = \{f \in L^p(\mathcal{X}) : D^\alpha f \in L^p(\mathcal{X}) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } \|\alpha\|_1 \leq \beta\}.$$

1. For $1 \leq p < \infty$ define $\|f\|_{W^{m,p}(\mathcal{X})} := \left(\sum_{0\leq\|\alpha\|_1\leq m} \|D^\alpha f\|^p_{L^p(\mathcal{X})}\right)^{1/p}$.
2. Define $\|f\|_{W^{m,\infty}(\mathcal{X})} := \max_{0\leq\|\alpha\|_1\leq m} \|D^\alpha f\|_{L^\infty(\mathcal{X})}$.

Approximation with a lower-dimensional support

Assumption 1 (Approximate Manifolds)

The distribution of X is supported on $\mathcal{M}_\rho$, a ρ-neighborhood of $\mathcal{M} \subset \mathcal{X}$, where $\mathcal{M}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian submanifold and $\mathcal{M}_\rho = \{x \in \mathcal{X} : \inf\{\|x - y\|_2 : y \in \mathcal{M}\} \leq \rho\}$ for $\rho \in (0, 1)$.

Theorem 2 (Improved CNN Approximation)

Suppose Assumption 1 holds, $f \in W^{\beta,\infty}(\mathcal{X})$ and the distribution of X is absolutely continuous w.r.t the Lebesgue measure. For $\varepsilon \in (0, 1)$, let

$$d_\varepsilon = O(d_{\mathcal{M}} \varepsilon^{-2} \log(d/\varepsilon)), \quad \rho_\varepsilon = C_2 \frac{(NM)^{-2\beta/d_\varepsilon}(\beta + 1)^2 \sqrt{d} d_\varepsilon^{3\beta/2}}{[\sqrt{d/d_\varepsilon} + 1 - \varepsilon](1 - \varepsilon)^{\beta - 1}}.$$

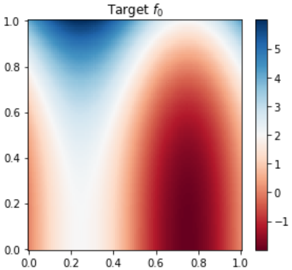
Then, for any $M, N \in \mathbb{N}^+$, there exists a CNN $f_{CNN} \in \mathcal{F}_{CNN}$ with L, $\mathcal{S}$ specified in Theorem 1 with $\mathcal{W} = 38^2(\lfloor\beta\rfloor + 1)^4 {d_\varepsilon}^{2\lfloor\beta\rfloor+2} N^2 \lceil\log_2(8N)\rceil^2$ such that

$$\mathbb{E}|f(X) - f_{CNN}(X)| \leq C(d, \beta)(NM)^{-2\beta/d_\varepsilon},$$

for $\rho \leq \rho_\varepsilon$ where $C(d, \beta) = (18 + C_2)B_0(1 - \varepsilon)^{-\beta}(\beta + 1)^2 d^{1/2} d_\varepsilon^{3\beta/2}$.

A Toy Example

- Target function: $f_0(x) = 2\sin(2\pi x_1) + 4(x_2)^3, \quad x \in [0,1]^2$.



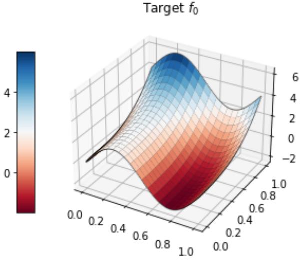


Table: Approximation errors by CNNs with different filter lengths and depths.

Approximation error		Filter length $\longrightarrow$			
$L_1(L_2)$		20	50	100	200
Hidden layers $\downarrow$	1	0.807(0.969)	0.450(0.539)	0.139(0.186)	0.062(0.084)
	2	0.112(0.144)	0.055(0.070)	0.047(0.064)	0.025(0.037)
	3	0.078(0.098)	0.051(0.070)	0.037(0.046)	0.032(0.045)

A Toy Example

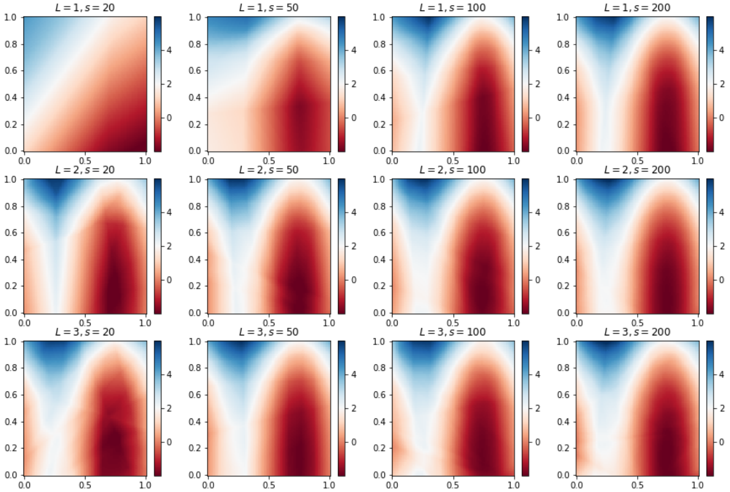


Figure: Heatmaps for the CNN approximations on the target function. The CNNs are designated with depth $L = 1, 2, 3$ and filter length $s = 20, 50, 100, 200$.

Application to binary classifications

- Sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \{1, -1\}$.
- Use Surrogate loss functions for 0-1 loss (or Misclassification loss).
- Given convex loss $\phi : \mathbb{R} \to [0, \infty)$. Risk $R(f) := \mathbb{E}\phi(Yf(X))$.
  1. Risk minimizer
  $$f_0 := \arg \min_{f \text{ measurable}} \mathbb{E}\phi(Yf(X)).$$
  2. Empirical risk minimizer
  $$\hat{f}_n \in \arg \min_{f \in \mathcal{F}_{\text{CNN}}} \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i f(X_i)). \tag{1}$$
  3. The classifier $\hat{h}_n(x) := \mathrm{sign}(\hat{f}_n(x))$ and $h_0(x) := \mathrm{sign}(f_0(x))$.

Application to binary classifications

- Self-calibration [7]:

$$\psi(\mathbb{P}(\hat{h}_n(X) \neq Y) - \mathbb{P}(h_0(X) \neq Y)) \leq R(\hat{f}_n) - R(f_0).$$

- Focus on the excess risk: $R(\hat{f}_n) - R(f_0)$.

Table: Surrogate loss, minimizer and self-calibration ψ.

	$\phi(a)$	$f_0(x)$	$\psi(\theta)$	$\psi^{-1}(\theta)$
Least squares	$(1-a)^2$	$2\eta - 1$	θ^2	$\sqrt{\theta}$
SVM	$\max\{1-a, 0\}$	$\operatorname{sign}(2\eta - 1)$	$\lvert\theta\rvert$	$\lvert\theta\rvert$
Exponential	$\exp(-a)$	$\frac{1}{2}\log(\frac{\eta}{1-\eta})$	$1 - \sqrt{1-\theta^2}$	$\sqrt{1-(1-\theta)^2}$
Logistic	$\log\{1+\exp(-a)\}$	$\log(\frac{\eta}{1-\eta})$	θ^2	$\sqrt{\theta}$
Cross entropy	$-\log\{0.5 + a\}$	$\eta - 0.5$	θ^2	$\sqrt{\theta}$

Note: $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$.

Application to binary classifications

Theorem 3 (Non-asymptotic excess ϕ-risk bound)

Suppose $f_0 \in W^{\beta,\infty}([0,1]^d, B_0)$ For any $M, N \in \mathbb{N}^+$, let depth L and filter lengths of $\mathcal{F}_{CNN}$ specified as in Theorem 1. Under mild conditions , for any $\delta \in (0,1)$, with probability $\geq 1-\delta$, the ERM $\hat{f}_n$ defined in (1) satisfies

$$R(\hat{f}_n) - R(f_0) \leq \frac{2\phi_B}{\sqrt{n}}\left(C_0\sqrt{\mathcal{S}L\log(S)\log(n)} + \sqrt{\log(1/\delta)}\right) \tag{2}$$

$$+ C(d,\beta)(NM)^{-2\beta/d} + \Delta_\phi(T). \tag{3}$$

where $C(d,\beta) = 18B_\phi B_0(\beta+1)^2 d^{\beta+(\beta\vee 1)/2}$, $C_0 > 0$ is a universal constant, and truncation error $\Delta_\phi(T) := \inf_{|a|\leq T}\phi(a) - \inf_{a\in Ran(f_0)}\phi(a)$ where $Ran(f_0)$ is the range of f_0. Additionally, if conditions in Theorem 2 holds, the approximation error (3) is improved to be $C(d,\beta)(NM)^{-2\beta/d_\epsilon} + \Delta_\phi(T)$ where d_ϵ is defined in Theorem 2.

Application to binary classifications

Table: Excess Misclassification Error

Hypothesis space	Loss	Condition	Rate	Reference
Measurable functions	0-1 loss	θ-noise condition; α-Hölder decision boundary	$n^{-\frac{\beta(\theta+1)}{\beta(\theta+2)+(d-1)\theta}}$	Theorem 1 in [8]
DNN	Hinge		$n^{-\frac{\beta(\theta+1)}{\beta(\theta+2)+(d-1)(\theta+1)}}$	Theorem 1 in [3]
Deep CNNs	1-norm	$f_0 \in W^{\beta,p}(\mathbb{S}^{d-1})$	$n^{-\frac{\beta}{\beta(2-\tau)+2\gamma(d-1)}}$	Theorem 2 in [2]
	p-norm		$n^{-\frac{p\beta}{2p\beta(2-\tau)+2p(\gamma+1)(d-1)}}$	
	2-norm	$f_0 \in W^{\beta,p}(\mathbb{S}^{d-1})$; θ-noise condition;	$n^{-\frac{2\beta\theta}{(2+\theta)((\gamma+1)(d-1)+2\beta)}}$	Theorem 3 in [2]
	Hinge	θ-noise condition; $f_0 \in W^{\beta,p}([0,1]^d)$	$n^{-\frac{\beta(\theta+1)}{d+2\beta(\theta+1)}}$	This paper
	Logistic	$f_0 \in W^{\beta,p}([0,1]^d)$	$n^{-\frac{\beta}{2d+4\beta}}$	
	Exponential	$f_0 \in W^{\beta,p}([0,1]^d)$	$n^{-\frac{\beta}{2d+4\beta}}$	
	Least square	$f_0 \in W^{\beta,p}([0,1]^d)$	$n^{-\frac{4\beta}{3d+16\beta}}$	

The p-norm hinge loss: $\phi(u) = \max\{1-u, 0\}^p$ with $p \geq 1$ (it is hinge loss when $p = 1$).

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

[2] H. Feng, S. Huang, and D.-X. Zhou. Generalization analysis of cnns for classification on spheres. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] Y. Kim, I. Ohn, and D. Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

[4] M. Kohler and S. Langer. Statistical theory for image classification using deep convolutional neural networks with cross-entropy loss. *arXiv preprint arXiv:2011.13602*, 2020.

[5] H. Liu, M. Chen, T. Zhao, and W. Liao. Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks. In *International Conference on Machine Learning*, pages 6770–6780. PMLR, 2021.

[6] K. Oono and T. Suzuki. Approximation and non-parametric estimation of resnet-type convolutional neural networks. In *International Conference on Machine Learning*, pages 4922–4931. PMLR, 2019.

[7] P. Petersen and F. Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020.

[8] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32 (1):135–166, 2004.

[9] D.-X. Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2):787–794, 2020.

Thank you!