Minimax optimality of convolutional neural networks for infinite dimensional input output problems and separation from kernel methods

Yuto Nishimura[1] and Taiji Suzuki[1,2]   [1]The University of Tokyo, [2]AIP-RIKEN

Background

・**Why NNs perform well with infinite-dimensional input and output?**

Existing work:
- Dealing with long sequence lengths in input and output for text, images, and audio: [Brown et al., 2020], [Rombach et al., 2022], [Radford et al., 2022]
- Research on infinite-dimensional inputs and outputs as linear operators: [Oliva et al., 2013;2014], [Fischer & Steinwart, 2020], [Talwai et al., 2022], [Jin et al., 2022]
- Via the low dimensional structure or smoothness argument: [Chen et al., 2019; 2022], [Nakada & Imaizumi , 2020], [Suzuki & Nitanda , 2021]
- DeepONet can approximate nonlinear operators: [Lu et al., 2021], [Lanthaler et al., 2022]
- CNNs achieve estimation errors depends on smoothness [Okumoto & Suzuki 2021]:
  - Using γ-*smooth space*, it showed the error rate with infinite dimension of input (not output).

We show that
- **CNNs achieve estimation error depends on smoothness and output's decay rate with infinite input & output dimension settings**
- **its lower bounds on the minimax optimal rate**
- **CNNs outperforms linear estimators**

Problem Settings

Target function:
$f^\circ = (f_i^\circ)_{i=1}^\infty \in (U(\mathcal{F}_{p,q}^\gamma))^\infty$

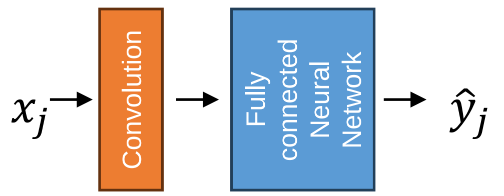
Observation:
$y_j = f^\circ(x_j) + \xi_j \quad (j = 1, \dots, n)$

Condition:
$$\|f_i^\circ\|_2 \le B_2 i^{-\frac{1}{r}},$$
$$\|f_i^\circ\|_\infty \le B_\infty \ (\forall i \in \mathbb{N}),$$
where $B_2 > 0$, $B_\infty > 0$, $0 < r < 1$ are constants.
$$\xi = (\xi_i)_{i=1}^\infty, \quad \xi_i \sim \mathcal{N}(0, \sigma_i^2),$$
where σ_i is uniformly bounded by $\bar{\sigma} < \infty$.

γ-smooth space:
$$\mathcal{F}_{p,q}^\gamma := \{f \in L^2 : \|f\|_{\mathcal{F}_{p,q}^\gamma} < \infty\},$$
where $\|f\|_{\mathcal{F}_{p,q}^\gamma} := \left(\sum_{s\in\mathbb{N}_0^\infty}\left(2^{\gamma(s)}\|\delta_s(f)\|_p\right)^q\right)^{\frac{1}{q}}$

Notations:
$$\mathbb{N}_0^\infty := \{l \in (\mathbb{N}\cup\{0\})^\infty : \mathrm{supp}(l) < \infty\}$$
$$\psi_{l_i}(x_i) := \begin{cases} \sqrt{2}\cos(2\pi|l_i|x_i) & (l_i < 0) \\ \sqrt{2}\sin(2\pi|l_i|x_i) & (l_i > 0) \\ 1 & (l_i = 0) \end{cases}$$
$$\psi_l(x) := \prod_{i=1}^\infty \psi_{l_i}(x_i)$$
$$\delta_s(f) := \sum_{l\in\mathbb{Z}_0^\infty : \lfloor 2^{s_i-1}\rfloor \le |l_i| < 2^{s_i}} \langle f, \psi_l\rangle \psi_l$$
γ: monotonically increasing function

Architecture:



Estimation Error:
$$\mathbb{E}_{D_n}\|\hat{f} - f^\circ\|_{L_2(P_X)}^2$$

Main result I: Lower bound for γ-smooth space

Theorem The minimax optimal rate for estimating a function in $(\mathcal{F}_{p,q}^\gamma)^\infty$ is:
$$\inf_{\hat{f}} \sup_{f^\star \in (U(\mathcal{F}_{p,q}^\gamma))^\infty \cap B_r} \mathbb{E}_{D_n}\left[\|\hat{f} - f^\star\|_{L_2(P_X)}^2\right] \gtrsim n^{-(2-r)a_1/(2a_1+1)}$$
Here, $\gamma(s) = \langle a, s\rangle$, $a = (a_i)_{i=1}^\infty > 0$ is a monotonically increasing sequence, and $a_i = \Omega(i^\eta)$, $\eta > 0$.

- We call this γ as **mixed smoothness:** smoothness varies by direction.
  → The reason it can generalize even in infinite dimensions.
- The output's decay rate r is included, marks a difference from standard settings.

Main result II: Estimation error of nonlinear operators by CNNs

Empirical risk minimization (ERM) estimator $\hat{f}$:
$$\hat{f} \in \mathrm{argmin}_{f\in\bar{\mathcal{P}}} \frac{1}{n}\sum_{j=1}^n \|f(x_j) - y_j\|_{\ell^2}^2,$$
where $\bar{\mathcal{P}}$ is the set of **CNNs**

- CNNs can perform appropriate variable selection from infinite-dimensional inputs.

Theorem The ERM estimator $\hat{f}$ achieves the following estimation error:
$$\mathbb{E}_{D_n}\left[\|\hat{f} - f^\star\|_{L_2(P_X)}^2\right] \lesssim n^{-(2-r)(a_1-v)/(2(a_1-v)+1)}$$
Here, $v := \max\{1/p - 1/2, 0\}$, and we ignore poly-log order.
Therefore, when $p \ge 2$, **CNNs achieve minimax optimal rate**.

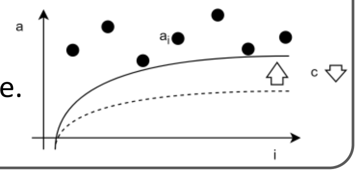
- CNNs, by using convolution, can **adaptively select important features** from training data (corresponding here to a_1).

→ This is a type of feature learning that **linear estimators cannot achieve**.

Main result III: Comparison with linear estimators

An estimator is called **linear** if it is written by:
$$\hat{f}(x) = \sum_{j=1}^n y_j \phi_j(x), \text{ where } \phi \text{ is an any function.}$$
ex., **Kernel ridge regression**:
$$\hat{f}(x) = [k(x, x_1), k(x, x_2), \cdots, k(x, x_n)](K + \lambda I)^{-1}Y$$

We define **the union of γ-smooth space $(\mathrm{F}_{p,q}(\Gamma))^\infty$** as follows:
$$(\mathrm{F}_{p,q}(\Gamma))^\infty := \bigcup_{a\in\Gamma} (\mathcal{F}_{p,q}^{\gamma_a})^\infty,$$
where Γ is a set of a, determined by the parameter c.
c changes the range of existence of a, as shown in the right figure.
As c decreases, the range of a also narrows.



Theorem The linear estimators achieve the following estimation error:
$$\inf_{\hat{f}:\text{linear}} \sup_{f^\circ \in U((\mathrm{F}_{2,2}(\Gamma))^\infty)\cap B_r} \mathbb{E}_{D_n}\left[\|\hat{f} - f^\circ\|_{L_2(P_X)}^2\right] \gtrsim n^{-2\underline{a}/(2\underline{a}+1+c)}$$
Here, $\underline{a}$ is a min value of a in Γ.

- When the following condition is satisfied, **CNNs outperform linear estimators:**
$$c > \frac{(2\underline{a}+1)r}{2-r}.$$
- When c is large, the set of a, Γ is larger, meaning the corresponding union set is bigger, making it more challenging for linear estimators.

Proof Strategy

For main result I:

Using Fano's inequality [Yang & Barron, 1999], [Raskutti et al., 2012], [Suzuki, 2019], what we need is **a lower bound for covering number of γ-smooth space.**
$$\inf_{\hat{f}} \sup_{f^\star \in (U(\mathcal{F}_{p,q}^\gamma))^\infty} \mathbb{E}_{D_n}\left[\|\hat{f} - f^\star\|_{L_2(P_X)}^2\right] \gtrsim \frac{\delta_n^2}{2}\left(1 - \frac{\log N + \frac{nd}{2}\epsilon_n^2 + \log 2}{\log M}\right)$$

$$(\mathcal{F}_{p,q}^\gamma)^\infty \supset (\mathcal{F}_{p,q}^\gamma)^d \ \Rightarrow$$
1. Select δ-packing of $\mathcal{F}_{p,q}^\gamma$: $\{0, f^1, \dots, f^k, \dots, f^N\}$
2. The packing of $(\mathcal{F}_{p,q}^\gamma)^d$ contains the combination of f^k [Raskutti et al., 2012]

A lower bound of the covering number of $\mathcal{F}_{p,q}^\gamma$ can be obtained by evaluating that of the following subset:
$$\{\delta_s(f) \mid \|\delta_s(f)\|_{\mathcal{F}_{p,q}^\gamma} \le 1\}.$$
This ball is equivalent to a Euclidean ball, so we get the covering number.

For main result III:

For linear estimators [Hayakawa & Suzuki, 2020]:
"The minimax rate of a space and the space of its convex hull are **equal**."

On the other hand…

For CNNs (see main result II):
The minimax rate depends **only on a_1** → **Feature extraction abilities of CNNs.**

Therefore, when we consider the union of the space, it makes the difference.

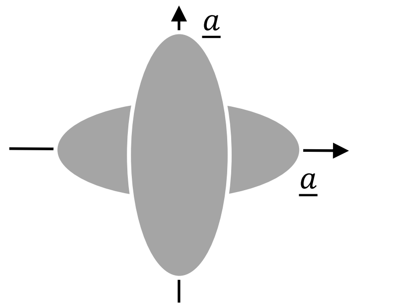


Illustration of the union of γ-smooth space.

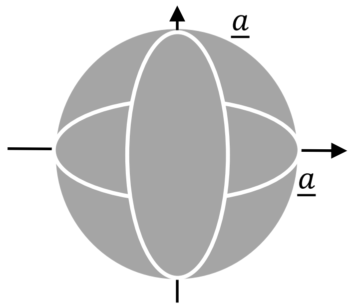


Illustration of the convex hull of the union of γ-smooth space.

Summary

We analyzed the estimation error of CNNs in infinite input & output settings.
Our contribution can be summarized as follows:

- Calculated **lower bound for γ-smooth space.**
- Showed **estimation error of CNNs & it's minimax optimal.**
- Confirmed **CNNs outperform linear estimators.**