Approximation and Non-parametric Estimation of ResNet-type Convolutional Neural Networks

Kenta Oono\textsuperscript{1,2} Taiji Suzuki\textsuperscript{1,3}

\{kenta\_oono, taiji\}@mist.i.u-tokyo.ac.jp
1. The University of Tokyo  2. Preferred Networks, Inc.  3. RIKEN AIP

Thirty-sixth International Conference on Machine Learning (ICML 2019)
June 13\textsuperscript{th} 2019, Long Beach, CA, U.S.A.
Key Takeaway

Q. Why ResNet-type CNNs work well?
Key Takeaway

Q. Why ResNet-type CNNs work well?

A. Hidden *sparse structure* promotes good performance.
Problem Setting

We consider a non-parametric regression problem:

\[ Y = f^\circ(X) + \xi \]

\( f^\circ \): True function (e.g., Hölder, Barron, Besov class), \( \xi \): Gaussian noise
Problem Setting

We consider a non-parametric regression problem:

\[ Y = f^\circ (X) + \xi \]

\( f^\circ \): True function (e.g., Hölder, Barron, Besov class), \( \xi \): Gaussian noise

Given \( N \) i.i.d. samples, we pick an estimator \( \hat{f} \) from the hypothesis class \( \mathcal{F} \), which is a set of functions realized by CNNs with a specified architecture.
Problem Setting

We consider a non-parametric regression problem:

\[
Y = f^\circ(X) + \xi
\]

\(f^\circ\): True function (e.g., Hölder, Barron, Besov class), \(\xi\): Gaussian noise

Given \(N\) i.i.d. samples, we pick an estimator \(\hat{f}\) from the hypothesis class \(\mathcal{F}\), which is a set of functions realized by CNNs with a specified architecture.

Goal: Evaluate the estimation error

\[
\mathcal{R}(\hat{f}) := \mathbb{E}_X |\hat{f}(X) - f^\circ(X)|^2
\]
Prior Work

\[
\mathcal{R}(\hat{f}) \lesssim \inf_{f \in \mathcal{F}} \| f - f^\circ \|_{\infty}^2 + \tilde{O}(M_{\mathcal{F}}/N)
\]

Approximation Error    Model Complexity

\(N\): Sample size
\(\mathcal{F}\): Set of functions realizable by CNNs with a specified architecture
\(f^\circ\): True function (e.g., Hölder, Barron, Besov etc.)
\(\tilde{O}(\cdot)\): \(O\)-notation ignoring logarithmic terms.
Prior Work

\[ R(\hat{f}) \lesssim \inf_{f \in \mathcal{F}} \| f - f^\circ \|_\infty^2 + \tilde{O}(M_\mathcal{F}/N) \]

Approximation Error \quad Model Complexity

| CNN type | Parameter Size $M_\mathcal{F}$ | Minimax Optimality | Discrete Optimization |
|----------|---------------------------------|--------------------|-----------------------|
| General  | # of all weights                | Sub-optimal 😞      | -                     |

$N$: Sample size  
$\mathcal{F}$: Set of functions realizable by CNNs with a specified architecture  
$f^\circ$: True function (e.g., Hölder, Barron, Besov etc.)  
$\tilde{O}(\cdot)$: $O$-notation ignoring logarithmic terms.
**Approximation Error** \( \mathcal{R}(\hat{f}) \lesssim \inf_{f \in \mathcal{F}} \| f - f^\circ \|_\infty^2 + O(\mathcal{M}_\mathcal{F}/N) \)  

| CNN type   | Parameter Size  | Minimax Optimality | Discrete Optimization |
|------------|-----------------|---------------------|-----------------------|
| General    | # of all weights| Sub-optimal 😞       | -                     |
| Sparse*    | # of non-zero weights | Optimal 😊          | Needed 😞             |

* e.g., Hölder case: [Yarotsuky, 17; Schmidt-Hieber, 17; Petersen & Voigtlaender, 18]

\( N \): Sample size  
\( \mathcal{F} \): Set of functions realizable by CNNs with a specified architecture  
\( f^\circ \): True function (e.g., Hölder, Barron, Besov etc.)  
\( \tilde{O}(\cdot) \): \( O \)-notation ignoring logarithmic terms.
Prior Work

$$\mathcal{R}(\hat{f}) \lesssim \inf_{f \in \mathcal{F}} \| f - f^{\circ} \|_\infty^2 + \tilde{O}(M_{\mathcal{F}} / N)$$

**Approximation Error**  **Model Complexity**

| CNN type  | Parameter Size $M_{\mathcal{F}}$ | Minimax Optimality | Discrete Optimization |
|-----------|-----------------------------------|--------------------|-----------------------|
| General   | # of all weights                  | Sub-optimal 😞      | -                     |
| Sparse*   | # of non-zero weights             | Optimal 😃           | Needed 😞              |
| ResNet    | # of all weights                  | Optimal 😃           | Not Needed 😃           |

* e.g., Hölder case: [Yarotsuky, 17; Schmidt-Hieber, 17; Petersen & Voigtlaender, 18]

$N$: Sample size

$\mathcal{F}$: Set of functions realizable by CNNs with a specified architecture

$f^{\circ}$: True function (e.g., Hölder, Barron, Besov etc.)

$\tilde{O}(\cdot)$: $O$-notation ignoring logarithmic terms.
## Contribution

ResNet-type CNNs can achieve minimax-optimal rates without unrealistic constraints.

| CNN type | Parameter Size $M_F$ | Minimax Optimality | Discrete Optimization |
|----------|----------------------|---------------------|-----------------------|
| General  | # of all weights     | Sub-optimal 😞       | -                     |
| Sparse*  | # of non-zero weights| Optimal 😊           | Needed 😞              |
| ResNet   | # of all weights     | Optimal 😊           | Not Needed 😊          |

* e.g., Hölder case: [Yarotsuky, 17; Schmidt-Hieber, 17; Petersen & Voigtlaender, 18]
Contribution

ResNet-type CNNs can achieve minimax-optimal rates without unrealistic constraints.

| CNN type  | Parameter Size $M_F$ | Minimax Optimality | Discrete Optimization |
|-----------|----------------------|--------------------|-----------------------|
| General   | # of all weights     | Sub-optimal 😞       | -                     |
| Sparse*   | # of non-zero weights | Optimal 😊          | Needed 😞             |
| ResNet    | # of all weights     | Optimal 😊          | Not Needed 😊         |

* e.g., Hölder case: [Yarotsuky, 17; Schmidt-Hieber, 17; Petersen & Voigtlaender, 18]

Key Observation

Known optimal FNNs have block-sparse structures
Block-sparse FNN

\[
FNN := \sum_{m=1}^{M} w_m^T FC_m(\cdot) - b
\]
Block-sparse FNN

\[ \text{FNN} := \sum_{m=1}^{M} w_m^T \text{FC}_m(\cdot) - b \]

Known best approximating FNNs are **block-sparse** when the true function is ---

- Barron [Klusowski & Barron, 18]
- Hölder [Yarotsky, 17; Schmidt-Hieber, 17]
- Besov [Suzuki, 19].
Block-sparse FNN to ResNet-type CNN

\[ \text{FNN} := \sum_{m=1}^{M} w_m^T \text{FC}_m(\cdot) - b \]

\[ \text{CNN} := \text{FC} \circ (\text{Conv}_M + \text{id}) \circ \cdots \circ (\text{Conv}_1 + \text{id}) \circ P \]

Known best approximating FNNs are block-sparse when the true function is ---

\[
\begin{aligned}
\text{Barron} & \quad \text{[Klusowski \& Barron, 18]} \\
\text{Hölder} & \quad \text{[Yarotsky, 17; Schmidt-Hieber, 17]} \\
\text{Besov} & \quad \text{[Suzuki, 19].}
\end{aligned}
\]
Block-sparse FNN to ResNet-type CNN

\[ \text{CNN: } = \text{FC} \circ (\text{Conv}_M + \text{id}) \circ \ldots \circ (\text{Conv}_1 + \text{id}) \circ P \]

Known best approximating FNNs are block-sparse when the true function is ---

- Barron [Klusowski & Barron, 18]
- Hölder [Yarotsky, 17; Schmidt-Hieber, 17]
- Besov [Suzuki, 19].
Block-sparse FNN to ResNet-type CNN

Known best approximating FNNs are block-sparse when the true function is ---

\[
\begin{align*}
\text{Barron} & \quad [\text{Klusowski & Barron, 18}] \\
\text{Hölder} & \quad [\text{Yarotsky, 17; Schmidt-Hieber, 17}] \\
\text{Besov} & \quad [\text{Suzuki, 19}].
\end{align*}
\]
For any block-sparse FNN with $M$ blocks, there exists a ResNet-type CNN with $M$ residual blocks which has $O(M)$ more parameters and which is identical (as a function) to the FNN.
Optimality of ResNet-type CNNs

**Theorem (e.g., Hölder Case)**

Suppose the true function $f^*$ is $\beta$-Hölder. There exists a set of ResNet-type CNNs $\mathcal{F}$ such that:
Suppose the true function $f^\circ$ is $\beta$-Hölder. There exists a set of ResNet-type CNNs $\mathcal{F}$ such that:

- $\mathcal{F}$ does **NOT** have sparse constraints
- the estimator $\hat{f}$ of $\mathcal{F}$ achieves the *minimax-optimal* estimation error rate (up to log factors).

**Theorem (e.g., Hölder Case)**
Theorem (e.g., Hölder Case)

Suppose the true function $f^*$ is $\beta$-Hölder. There exists a set of ResNet-type CNNs $\mathcal{F}$ such that:

- $\mathcal{F}$ does **NOT** have sparse constraints
- the estimator $\hat{f}$ of $\mathcal{F}$ achieves the minimax-optimal estimation error rate (up to log factors).

😊 Minimax optimal ! 😞 No discrete optimization !
Optimality of ResNet-type CNNs

Theorem (e.g., Hölder Case)

Suppose the true function $f^\circ$ is $\beta$-Hölder. There exists a set of ResNet-type CNNs $\mathcal{F}$ such that:

- $\mathcal{F}$ does NOT have sparse constraints
- the estimator $\hat{f}$ of $\mathcal{F}$ achieves the minimax-optimal estimation error rate (up to log factors).

😊 Minimax optimal! 😞 No discrete optimization!

Note

- Using the same strategy, we can prove that ResNet-type CNNs can achieve the same rate as FNNs for the Barron class etc.
- We remove unrealistic constraints on channels size, too (see the paper).
Conclusion

ResNet-type CNNs can achieve minimax-optimal rates in several function classes without implausible constraints.

| CNN type | Parameter Size $M_F$ | Minimax Optimality | Discrete Optimization |
|----------|----------------------|--------------------|-----------------------|
| General  | # of all weights     | Sub-optimal 😞      | -                     |
| Sparse*  | # of non-zero weights| Optimal 😋          | Needed 😞              |
| ResNet   | # of all weights     | Optimal 😋          | Not Needed 😞          |

↑ Minimax Optimal

↑ Minimax Optimal, too!

Poster: 13th June, Pacific Ballroom #77