A Theoretical Framework of Convolutional Kernels on Image Datasets

Song Mei

Department of Statistics, UC Berkeley

Deep Learning Theory Workshop, Simons Institute

Joint work with Theodor Misiakiewicz and Andrea Montanari

ImageNet Competition: Supervised learning

IMAGENET

Training:
1,200,000 samples
1,000 categories

Train a classifier

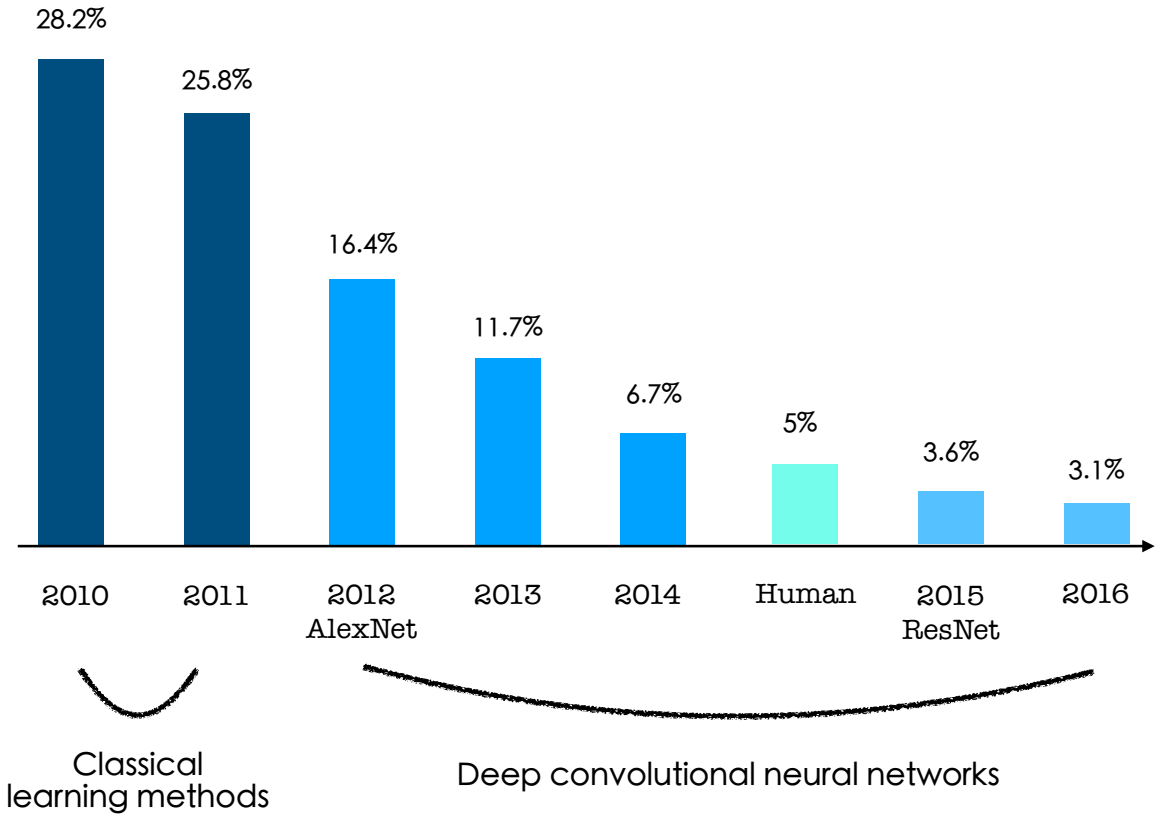
Test:
100,000 samples

Prediction

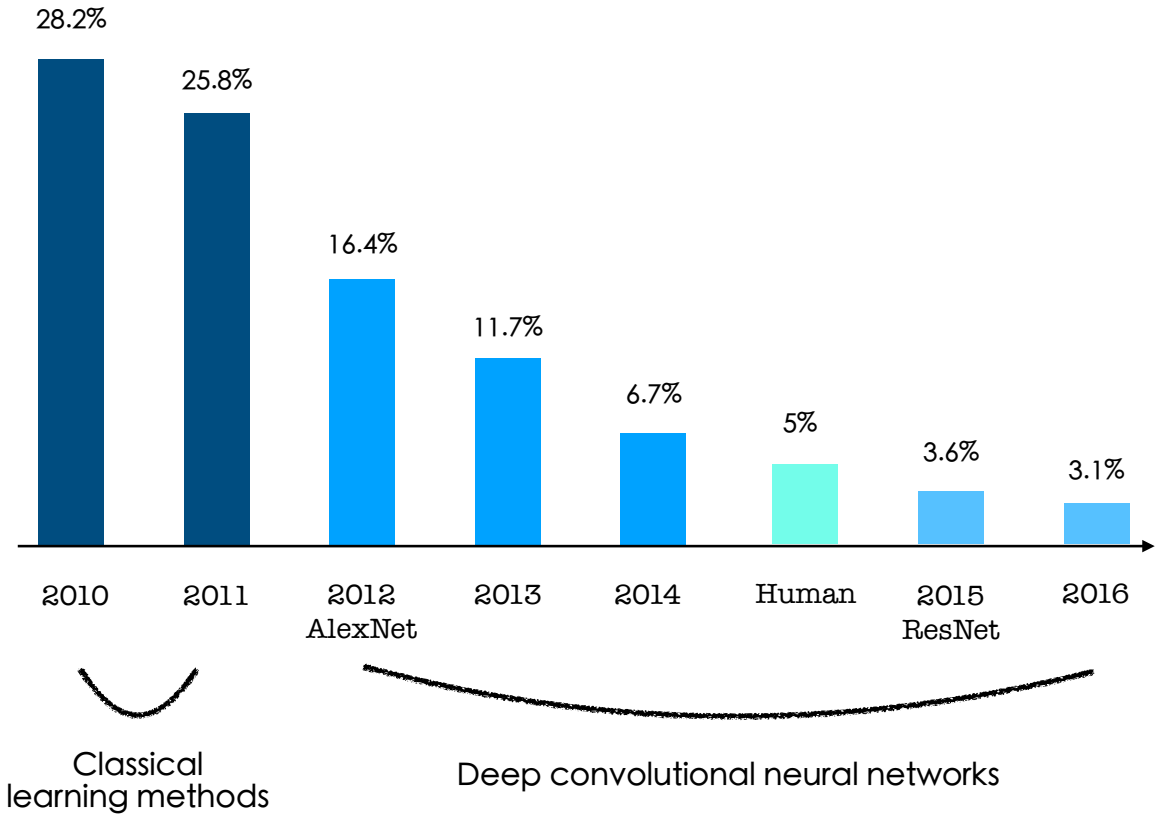
Cat, Mushroom, Brick???

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009

ImageNet Competition: Top 5 classification error



ImageNet Competition: Top 5 classification error



A folklore for the success

Convolution structure plays well with image dataset.

How about convolutional kernels?

Neural networks $\approx$ Kernels (in certain scaling regime)
[Jacot, Gabriel, Hongler, 2018]

How about convolutional kernels?

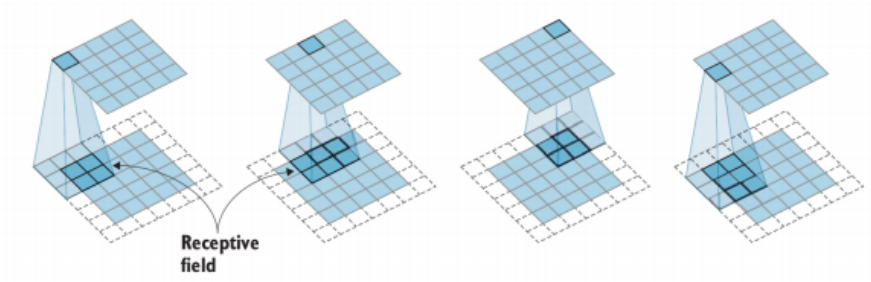
Neural networks $\approx$ Kernels (in certain scaling regime)
[Jacot, Gabriel, Hongler, 2018]

CIFAR10 accuracy of convolutional kernels:

Paper	method	Error
-	Gaussian kernel	43%
[Arora, et.al., 2019a]	CNTK (data independent)	23%
[Li, et.al., 2019]	CRFK (data dependent preprocessing)	11%
[Shankar, et.al., 2020]	Myrtle10 CK	10%
[Bietti, 2022]	3-layers CK	12%

Two important properties of images

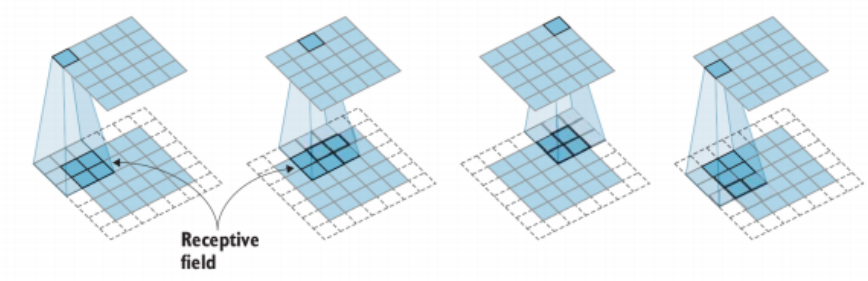
- Locality of features in images.



(Credit: [Elgendy, 2020])

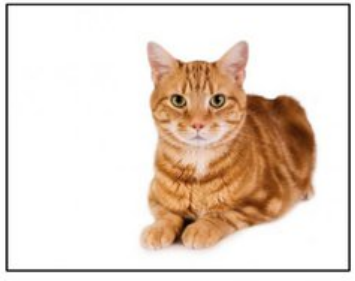
Two important properties of images

- Locality of features in images.



(Credit: [Elgendy, 2020])

- Translation invariance in images.

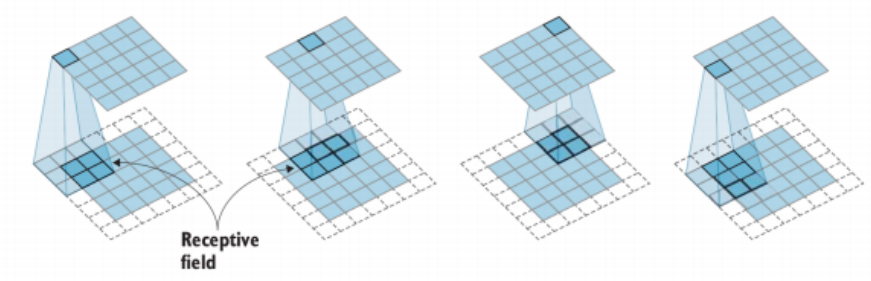


Cat

Cat

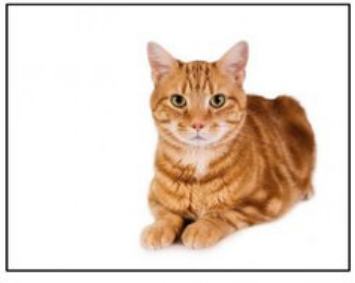
Two important properties of images

- Locality of features in images.



(Credit: [Elgendy, 2020])

- Translation invariance in images.


Cat


Cat

Question

How to mathematically quantify the advantage of convolution and pooling operations for image dataset?

Modeling invariance in image dataset?

Cat

Cat

The generative model: invariant functions

- Covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \mathbb{R}^d$.

The generative model: invariant functions

- Covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \mathbb{R}^d$.
- Consider a group $\mathcal{G}_d$ that can act on $\mathbb{R}^d$, e.g., the cyclic group $\mathcal{G}_d = \{g_0, g_1, \ldots, g_{d-1}\}$

$$g_i \cdot \boldsymbol{x} = (x_i, x_{i+1}, \ldots, x_d, x_1, \ldots, x_{i-1}).$$

  Other groups: 2D cyclic group; band-limited shift-invariant group.

The generative model: invariant functions

- Covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \mathbb{R}^d$.
- Consider a group $\mathcal{G}_d$ that can act on $\mathbb{R}^d$, e.g., the cyclic group $\mathcal{G}_d = \{g_0, g_1, \ldots, g_{d-1}\}$

$$g_i \cdot \boldsymbol{x} = (x_i, x_{i+1}, \ldots, x_d, x_1, \ldots, x_{i-1}).$$

  Other groups: 2D cyclic group; band-limited shift-invariant group.

- Label y induced by an invariant function $f_\star$

$$f_\star(\boldsymbol{x}) = f_\star(g \cdot \boldsymbol{x}), \qquad \forall g \in \mathcal{G}_d.$$

  e.g., $f_\star(\boldsymbol{x}) = \sum_{i=1}^{d} x_i x_{i+1}$: a degree 2 cyclic polynomial.

The generative model: invariant functions

- Covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \mathbb{R}^d$.
- Consider a group $\mathcal{G}_d$ that can act on $\mathbb{R}^d$, e.g., the cyclic group $\mathcal{G}_d = \{g_0, g_1, \ldots, g_{d-1}\}$

$$g_i \cdot \boldsymbol{x} = (x_i, x_{i+1}, \ldots, x_d, x_1, \ldots, x_{i-1}).$$

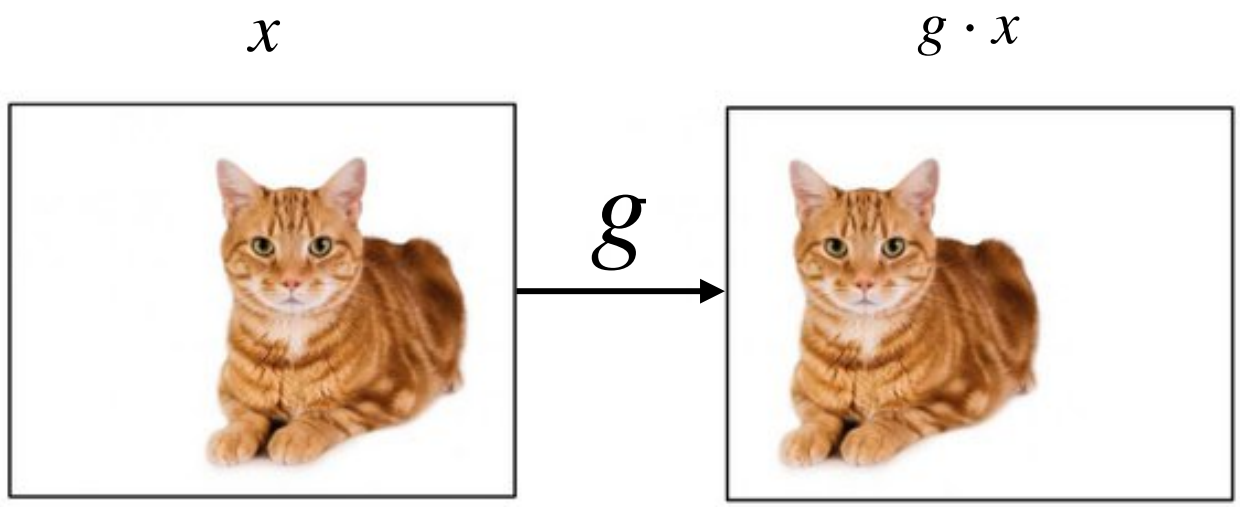
  Other groups: 2D cyclic group; band-limited shift-invariant group.

- Label y induced by an invariant function $f_\star$

$$f_\star(\boldsymbol{x}) = f_\star(g \cdot \boldsymbol{x}), \qquad \forall g \in \mathcal{G}_d.$$

  e.g., $f_\star(\boldsymbol{x}) = \sum_{i=1}^{d} x_i x_{i+1}$: a degree 2 cyclic polynomial.

Stylized model for an image label $y = f_\star(\boldsymbol{x}) + \varepsilon$ invariant by translation of image $\boldsymbol{x}$.



$y = f(x) =$ **"cat"**

$y = f(g \cdot x) =$ **"cat"**

Two-layer neural networks

- Two-layer fully-connected (FC) NN:

$$\hat{f}_{\mathsf{FC}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{N} a_i \cdot \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle).$$

Two-layer neural networks

- Two-layer fully-connected (FC) NN:

$$\hat{f}_{\mathsf{FC}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{N} a_i \cdot \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle).$$

- Two-layer invariant (IV) NN:

$$\hat{f}_{\mathsf{IV}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{N} a_i \cdot \int_{g \in \mathcal{G}_d} \sigma(\langle g \cdot \boldsymbol{x}, \boldsymbol{w}_i \rangle) \pi(\mathrm{d}g)$$

Two-layer neural networks

- Two-layer fully-connected (FC) NN:

$$\hat{f}_{\mathsf{FC}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{N} a_i \cdot \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle).$$

- Two-layer invariant (IV) NN:

$$\hat{f}_{\mathsf{IV}}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{N} a_i \cdot \int_{g \in \mathcal{G}_d} \sigma(\langle g \cdot \boldsymbol{x}, \boldsymbol{w}_i \rangle) \pi(\mathrm{d}g)$$

- When $\mathcal{G}_d$ is the cyclic group, this is two-layers convolutional NN (full window size) with global average pooling.

... and their corresponding kernels

- Fully-connected (FC) NNs induce inner-product kernels (the NTK):

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d),$$

where $h(\langle \boldsymbol{x}, \boldsymbol{y} \rangle / d) = \mathbb{E}_{\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)\sigma(\langle \boldsymbol{y}, \boldsymbol{w} \rangle)]$.

... and their corresponding kernels

- Fully-connected (FC) NNs induce inner-product kernels (the NTK):

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d),$$

  where $h(\langle \boldsymbol{x}, \boldsymbol{y} \rangle / d) = \mathbb{E}_{\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)\sigma(\langle \boldsymbol{y}, \boldsymbol{w} \rangle)]$.

- Invariant (IV) NNs induce invariant kernels:

$$K_{\mathsf{IV}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{g \in \mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2 \rangle / d) \pi(\mathrm{d}g).$$

... and their corresponding kernels

- Fully-connected (FC) NNs induce inner-product kernels (the NTK):

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d),$$

  where $h(\langle \boldsymbol{x}, \boldsymbol{y} \rangle / d) = \mathbb{E}_{\boldsymbol{w} \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[\sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)\sigma(\langle \boldsymbol{y}, \boldsymbol{w} \rangle)]$.

- Invariant (IV) NNs induce invariant kernels:

$$K_{\mathsf{IV}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{g \in \mathcal{G}_d} h(\langle \boldsymbol{x}_1, g \cdot \boldsymbol{x}_2 \rangle / d) \pi(\mathrm{d}g).$$

- Goal: quantify the advantage of K_{IV} over K_{FC} learning invariant function.

Test error of KRR with cyclic invariant kernel

Samples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$:

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})), \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2),$$

where $f_\star$ is $\mathcal{G}_d$-invariant, for $\mathcal{G}_d$ to be 1D or 2D cyclic group.

Test error of KRR with cyclic invariant kernel

Samples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$:

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})), \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2),$$

where $f_\star$ is $\mathcal{G}_d$-invariant, for $\mathcal{G}_d$ to be 1D or 2D cyclic group.

Theorem (Mei, Misiakiewicz, Montanari, 2021 (Informal))

Assume sufficiently smooth activation function. Consider kernel ridge regression with FC and IV kernel, $\mathsf{KRR}_{\mathsf{FC}}$ *and* $\mathsf{KRR}_{\mathsf{IV}}$ *respectively. In the regime* $n, d \to \infty$ *with* $d^\ell \ll n \ll d^{\ell+1}$*, w.h.p,*

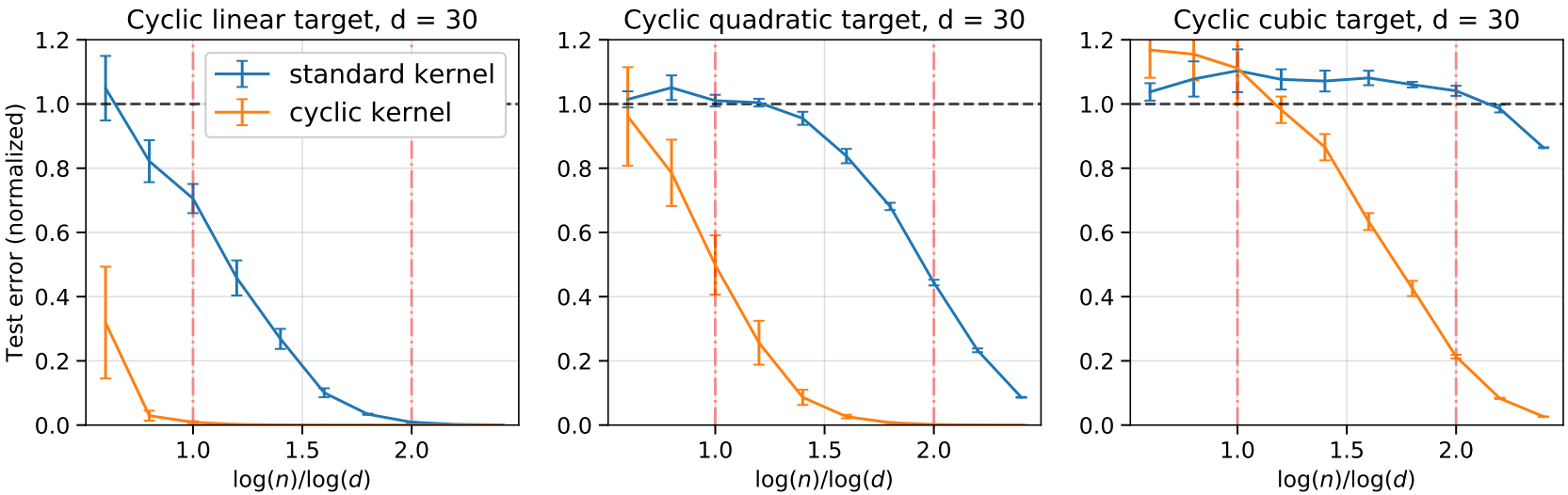
$$\|\mathsf{KRR}_{\mathsf{FC}} - \mathsf{P}_{\le \ell} f_\star\|_{L^2} = o(1),$$
$$\|\mathsf{KRR}_{\mathsf{IV}} - \mathsf{P}_{\le \ell+1} f_\star\|_{L^2} = o(1).$$

Equivalently, the test error satisfies

$$\mathbb{E}[(f_\star(\boldsymbol{x}) - \mathsf{KRR}_{\mathsf{FC}}(\boldsymbol{x}))]^2 = \|\mathsf{P}_{>\ell} f_\star\|_{L^2}^2 + o(1),$$
$$\mathbb{E}[(f_\star(\boldsymbol{x}) - \mathsf{KRR}_{\mathsf{IV}}(\boldsymbol{x}))]^2 = \|\mathsf{P}_{>\ell+1} f_\star\|_{L^2}^2 + o(1).$$

Numerical simulations

$$f_{\mathrm{lin}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i, \qquad f_{\mathrm{quad}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i x_{i+1}, \qquad f_{\mathrm{cube}} = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} x_i x_{i+1} x_{i+2}.$$



Another interpretation of the theoretical result

To fit a degree ℓ invariant polynomial:

$\mathsf{KRR}_{\mathsf{FC}}$ require sample size $n \asymp d^{\ell}$,

$\mathsf{KRR}_{\mathsf{IV}}$ require sample size $n \asymp d^{\ell-1}$.

Another interpretation of the theoretical result

To fit a degree ℓ invariant polynomial:
$\mathrm{KRR}_{\mathsf{FC}}$ require sample size $n \asymp d^{\ell}$,
$\mathrm{KRR}_{\mathsf{IV}}$ require sample size $n \asymp d^{\ell-1}$.

- We gain a factor d in sample complexity by using a cyclic kernel.
- Similar results hold for invariant random features.
- For general group $\mathcal{G}_d$, we gain a sample size factor p which corresponds to the "effective dimension" of the group $\mathcal{G}_d$.

The mechanism

Principle: KRR fit functions in the top $O(n)$ eigenspace of the kernel.

The mechanism

Principle: KRR fit functions in the top $O(n)$ eigenspace of the kernel.

- For FC kernel, the eigen-functions are **spherical harmonics**

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=0}^{\infty} \lambda_k(h) \sum_{l=1}^{B(d,k)} Y_{k,l}^{(d)}(\boldsymbol{x}_1) Y_{k,l}^{(d)}(\boldsymbol{x}_2).$$

The mechanism

Principle: KRR fit functions in the top $O(n)$ eigenspace of the kernel.

- For FC kernel, the eigen-functions are **spherical harmonics**

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=0}^{\infty} \lambda_k(h) \sum_{l=1}^{B(d,k)} Y_{k,l}^{(d)}(\boldsymbol{x}_1) Y_{k,l}^{(d)}(\boldsymbol{x}_2).$$

- For IV kernel, the eigen-functions are **invariant spherical harmonics**

$$K_{\mathsf{IV}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=0}^{\infty} \lambda_k(h) \sum_{l=1}^{D(d,k)} \overline{Y}_{k,l}^{(d)}(\boldsymbol{x}_1) \overline{Y}_{k,l}^{(d)}(\boldsymbol{x}_2).$$

The mechanism

Principle: KRR fit functions in the top $O(n)$ eigenspace of the kernel.

- For FC kernel, the eigen-functions are **spherical harmonics**

$$K_{\mathsf{FC}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=0}^{\infty} \lambda_k(h) \sum_{l=1}^{B(d,k)} Y_{k,l}^{(d)}(\boldsymbol{x}_1) Y_{k,l}^{(d)}(\boldsymbol{x}_2).$$

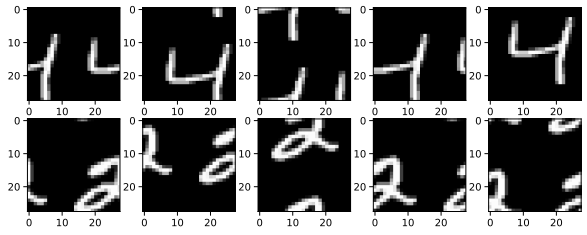
- For IV kernel, the eigen-functions are **invariant spherical harmonics**

$$K_{\mathsf{IV}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=0}^{\infty} \lambda_k(h) \sum_{l=1}^{D(d,k)} \overline{Y}_{k,l}^{(d)}(\boldsymbol{x}_1) \overline{Y}_{k,l}^{(d)}(\boldsymbol{x}_2).$$

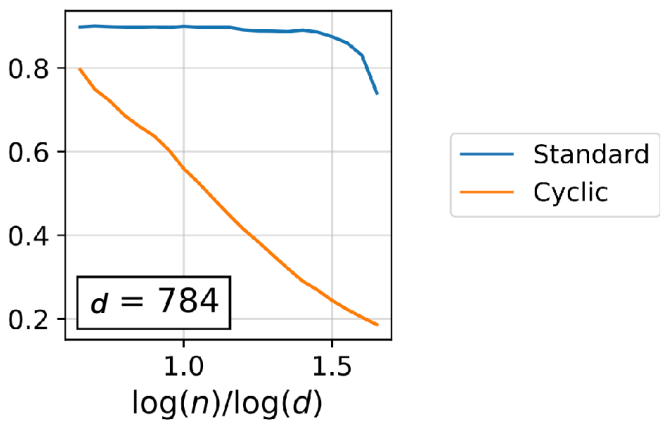
Kernel	FC	IV
Eigenspace $O(d^0) - O(d^1)$	Degree-1 harmonics	Degree-2 cyclic harmonics
Eigenspace $O(d^1) - O(d^2)$	Degree-2 harmonics	Degree-3 cyclic harmonics
Eigenspace $O(d^2) - O(d^3)$	Degree-3 harmonics	Degree-4 cyclic harmonics

Cyclic invariant MNIST

- Make MNIST dataset invariant under cyclic translation in 2 dimensions.

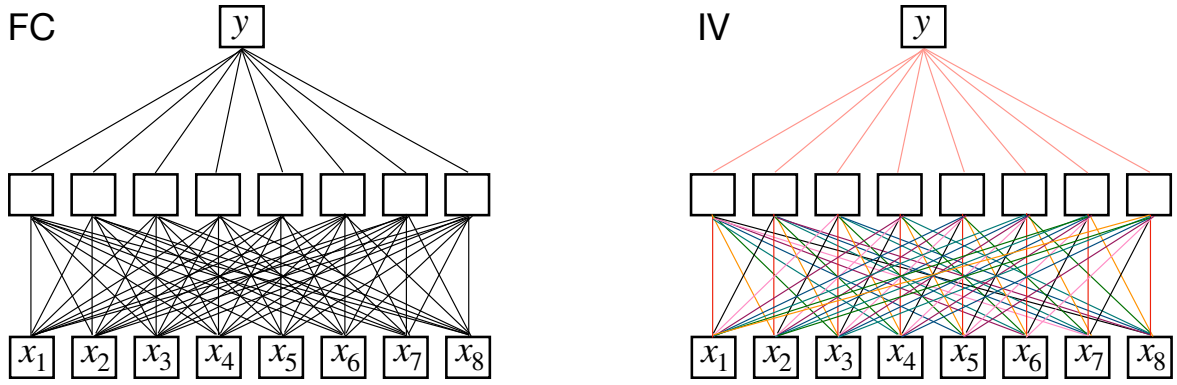


- Compare the classification error of KRR with inner-product kernel and cyclic kernel.



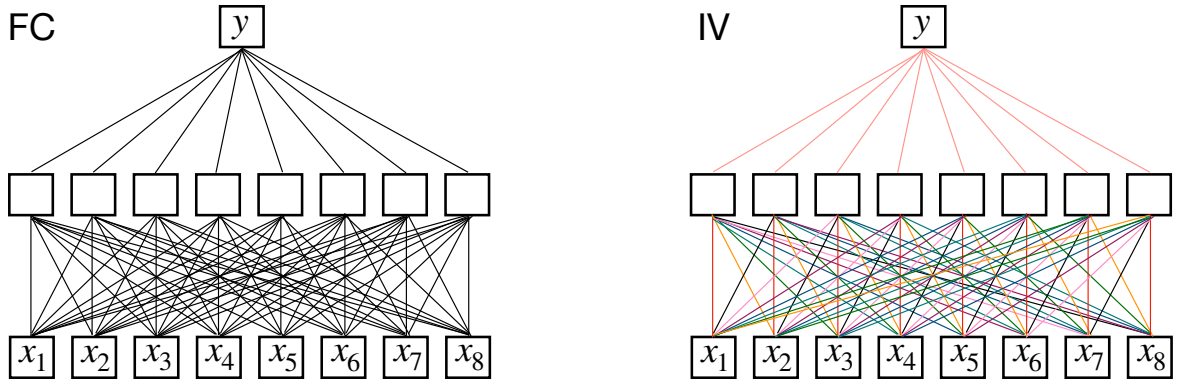
Locality vs invariance

Cyclic NN can be understood as FCNN with weight sharing structure.

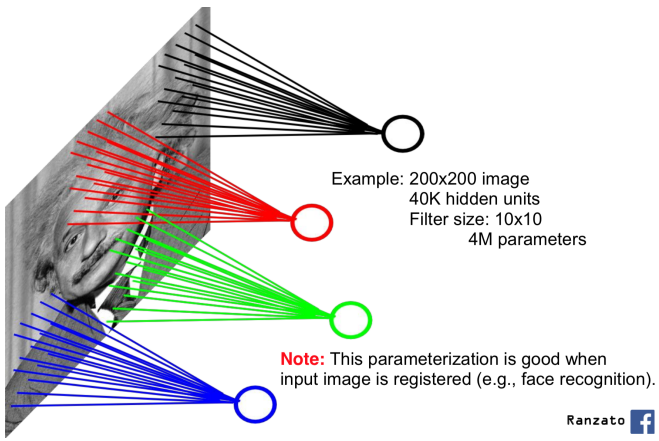


Locality vs invariance

Cyclic NN can be understood as FCNN with weight sharing structure.

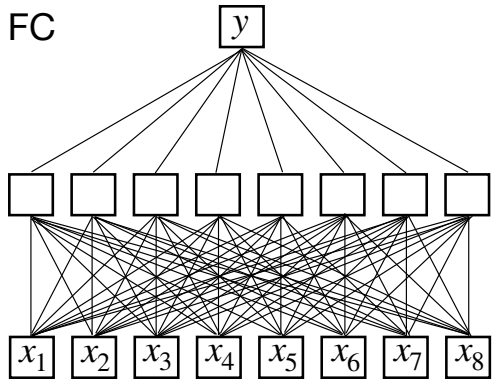


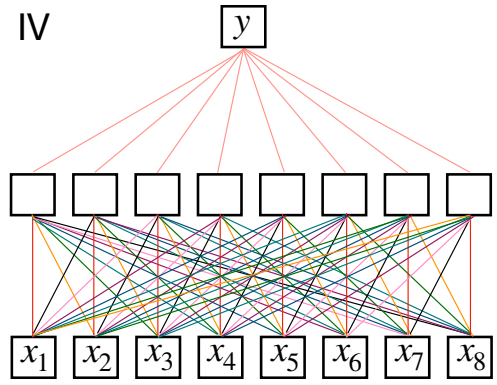
However, in practice, convolutional neural networks are locally-connected



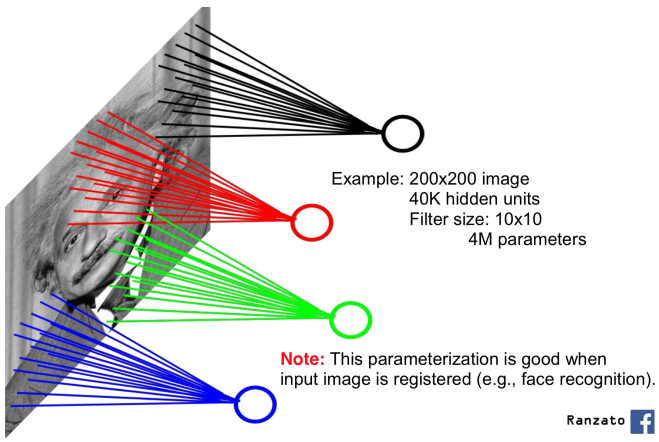
Locality vs invariance

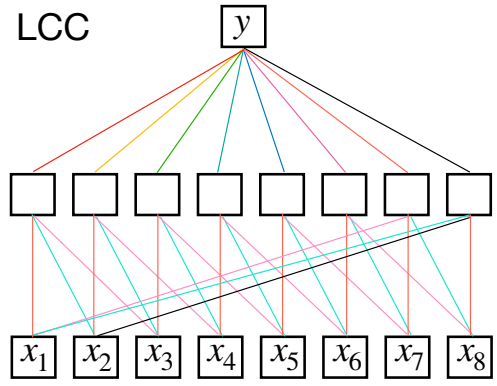
Cyclic NN can be understood as FCNN with weight sharing structure.



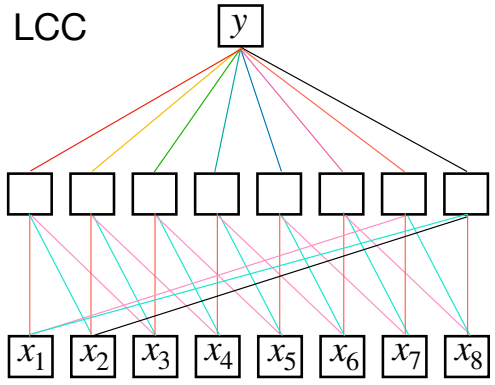


However, in practice, convolutional neural networks are locally-connected



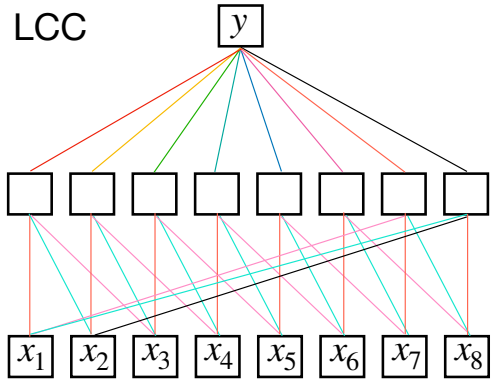


Modeling locality



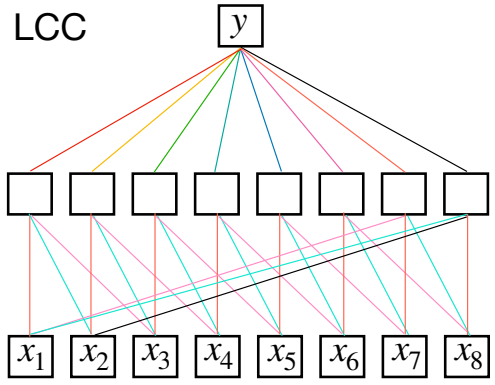
- Covariate $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \{\pm 1\}^d$.

Modeling locality



- Covariate $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \{\pm 1\}^d$.
- The k'th patch $\boldsymbol{x}_{(k)} = (x_{k+1}, \ldots, x_{k+q})^\mathsf{T} \in \{\pm 1\}^q$, $\quad q$: window size.

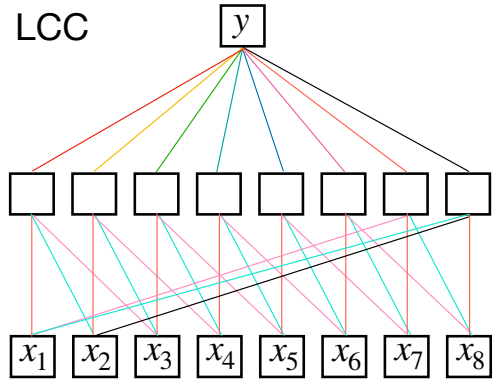
Modeling locality



- Covariate $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \{\pm 1\}^d$.
- The k'th patch $\boldsymbol{x}_{(k)} = (x_{k+1}, \ldots, x_{k+q})^\mathsf{T} \in \{\pm 1\}^q$, $\quad q$: window size.
- Weights (convolutional filters) $\{\boldsymbol{w}_i\}_{i \in [N]} \subseteq \{\pm 1\}^q$.
- Locally-connected and convolutional (LCC) NN with window size q

$$\hat{f}_{\mathsf{LCC}}(\boldsymbol{x}) = \sum_{i \in [N]} \sum_{k \in [d]} a_{ik} \sigma\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k)} \rangle\right).$$

Modeling locality



- Covariate $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\mathsf{T} \in \{\pm 1\}^d$.
- The k'th patch $\boldsymbol{x}_{(k)} = (x_{k+1}, \ldots, x_{k+q})^\mathsf{T} \in \{\pm 1\}^q$, $\quad q$: window size.
- Weights (convolutional filters) $\{\boldsymbol{w}_i\}_{i \in [N]} \subseteq \{\pm 1\}^q$.
- Locally-connected and convolutional (LCC) NN with window size q

$$\hat{f}_{\mathsf{LCC}}(\boldsymbol{x}) = \sum_{i \in [N]} \sum_{k \in [d]} a_{ik} \sigma\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k)} \rangle\right).$$

- Locally-connected and convolutional (LCC) kernel with window size q

$$K_{\mathsf{LCC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k)} \rangle / q).$$

Properties of LCC kernels with window size q

- The range of the kernel is the space of q-local functions

$$\left\{ f(x) = \sum_{k \in [d]} g_k(x_{(k)}) : \{g_k\}_{k \in [d]} \subseteq L^2(\{\pm 1\}^q) \right\}.$$

Properties of LCC kernels with window size q

- The range of the kernel is the space of q-local functions

$$\left\{ f(\boldsymbol{x}) = \sum_{k \in [d]} g_k(\boldsymbol{x}_{(k)}) : \{g_k\}_{k \in [d]} \subseteq L^2(\{\pm 1\}^q) \right\}.$$

- Eigen-functions are q-local harmonics [Misiakiewicz, Mei, 2021]

$$K_{\mathsf{LCC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k)} \rangle / q) = \sum_{\ell=0}^{q} \sum_{S \in \mathcal{E}_\ell} r(S) \xi_{q,\ell} \cdot Y_S(\boldsymbol{x}) Y_S(\boldsymbol{y}),$$

where the eigenfunctions are $Y_S(\boldsymbol{x}) = \prod_{i \in S} x_i$ ($\mathcal{E}_\ell$ is all sets of length ℓ within windows of size q).

Properties of LCC kernels with window size q

- The range of the kernel is the space of q-local functions

$$\left\{ f(\boldsymbol{x}) = \sum_{k \in [d]} g_k(\boldsymbol{x}_{(k)}) : \{g_k\}_{k \in [d]} \subseteq L^2(\{\pm 1\}^q) \right\}.$$

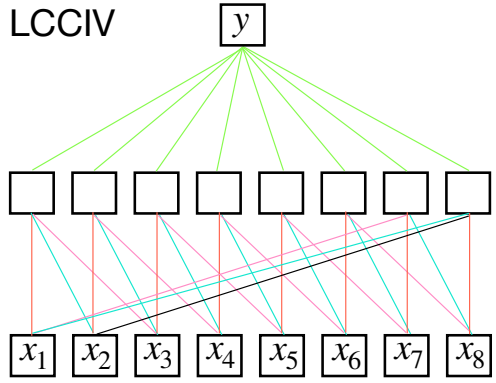
- Eigen-functions are q-local harmonics [Misiakiewicz, Mei, 2021]

$$K_{\mathsf{LCC}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k)} \rangle / q) = \sum_{\ell=0}^{q} \sum_{S \in \mathcal{E}_\ell} r(S) \xi_{q,\ell} \cdot Y_S(\boldsymbol{x}) Y_S(\boldsymbol{y}),$$

where the eigenfunctions are $Y_S(\boldsymbol{x}) = \prod_{i \in S} x_i$ ($\mathcal{E}_\ell$ is all sets of length ℓ within windows of size q).

Kernel	LCC, q-local
Eigenspace $O(1) - O(d)$	Degree-1 q-local harmonics
Eigenspace $O(dq^0) - O(dq^1)$	Degree-2 q-local harmonics
Eigenspace $O(dq^1) - O(dq^2)$	Degree-3 q-local harmonics

Locality + Invariance



- q-locally-connected, convolutional and invariant (LCCIV) neural networks

$$\hat{f}_{\text{LCCIV}}(\boldsymbol{x}) = \sum_{i \in [N]} a_i \sum_{k \in [d]} \sigma\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_{(k)} \rangle\right).$$

- q-locally-connected, convolutional and invariant (LCCIV) kernels

$$K_{\text{LCCIV}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k,k' \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')} \rangle / q).$$

Properties of LCCIV kernels with window size q

- The range of the kernel is the space of cyclic q-local functions

$$\left\{ f(\boldsymbol{x}) = \sum_{k \in [d]} g(\boldsymbol{x}_{(k)}) : g \in L^2(\{\pm 1\}^q) \right\}.$$

Properties of LCCIV kernels with window size q

- The range of the kernel is the space of cyclic q-local functions

$$\left\{ f(\boldsymbol{x}) = \sum_{k \in [d]} g(\boldsymbol{x}_{(k)}) : g \in L^2(\{\pm 1\}^q) \right\}.$$

- Eigen-functions are cyclic q-local polynomials [Misiakiewicz, Mei, 2021]

$$K_{\text{LCCIV}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k,k' \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')} \rangle / q) = \sum_{\ell=0}^{q} \sum_{S \in \mathcal{E}_\ell} r(S) \xi_{q,\ell} \cdot \overline{Y}_S(\boldsymbol{x}) \overline{Y}_S(\boldsymbol{y}),$$

where the eigenfunctions are $\overline{Y}_S = d^{-1} \sum_{k=1}^{d} Y_{S+k}$ ($\mathcal{E}_\ell$ is all sets of length ℓ within windows of size q).

Properties of LCCIV kernels with window size q

- The range of the kernel is the space of cyclic q-local functions

$$\left\{ f(\boldsymbol{x}) = \sum_{k \in [d]} g(\boldsymbol{x}_{(k)}) : g \in L^2(\{\pm 1\}^q) \right\}.$$

- Eigen-functions are cyclic q-local polynomials [Misiakiewicz, Mei, 2021]

$$K_{\text{LCCIV}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k,k' \in [d]} h(\langle \boldsymbol{x}_{(k)}, \boldsymbol{y}_{(k')} \rangle / q) = \sum_{\ell=0}^{q} \sum_{S \in \mathcal{E}_\ell} r(S) \xi_{q,\ell} \cdot \overline{Y}_S(\boldsymbol{x}) \overline{Y}_S(\boldsymbol{y}),$$

where the eigenfunctions are $\overline{Y}_S = d^{-1} \sum_{k=1}^{d} Y_{S+k}$ ($\mathcal{E}_\ell$ is all sets of length ℓ within windows of size q).

Kernel	LCCIV, q-local
Eigenspace $O(q^0) - O(q^1)$	Degree-2 cyclic q-local harmonics
Eigenspace $O(q^1) - O(q^2)$	Degree-3 cyclic q-local harmonics
Eigenspace $O(q^2) - O(q^3)$	Degree-4 cyclic q-local harmonics

Test error of KRR with convolutional kernels

Let $f_\star(\boldsymbol{x}) = \sum_{k\in[d]} g(\boldsymbol{x}_{(k)})$ be cyclic q-local. Given iid samples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$,

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim \mathrm{Unif}(\{\pm 1\}^d), \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2).$$

Theorem (Misiakiewicz, Mei, 2021 (Informal))

To fit the degree ℓ polynomial part of $f_\star$,

KRR with K_{LCC} requires sample size

$$dq^{\ell-1} \ll n \ll dq^{\ell}$$

KRR with K_{LCCIV} requires sample size

$$q^{\ell-1} \ll n \ll q^{\ell}.$$

Test error of KRR with convolutional kernels

Let $f_\star(\boldsymbol{x}) = \sum_{k\in[d]} g(\boldsymbol{x}_{(k)})$ be cyclic q-local. Given iid samples $\{(y_i, \boldsymbol{x}_i)\}_{i\in[n]}$,

$$y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i, \qquad \boldsymbol{x}_i \sim \mathrm{Unif}(\{\pm 1\}^d), \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2).$$

Theorem (Misiakiewicz, Mei, 2021 (Informal))

To fit the degree ℓ polynomial part of $f_\star$,

KRR with K_{LCC} requires sample size

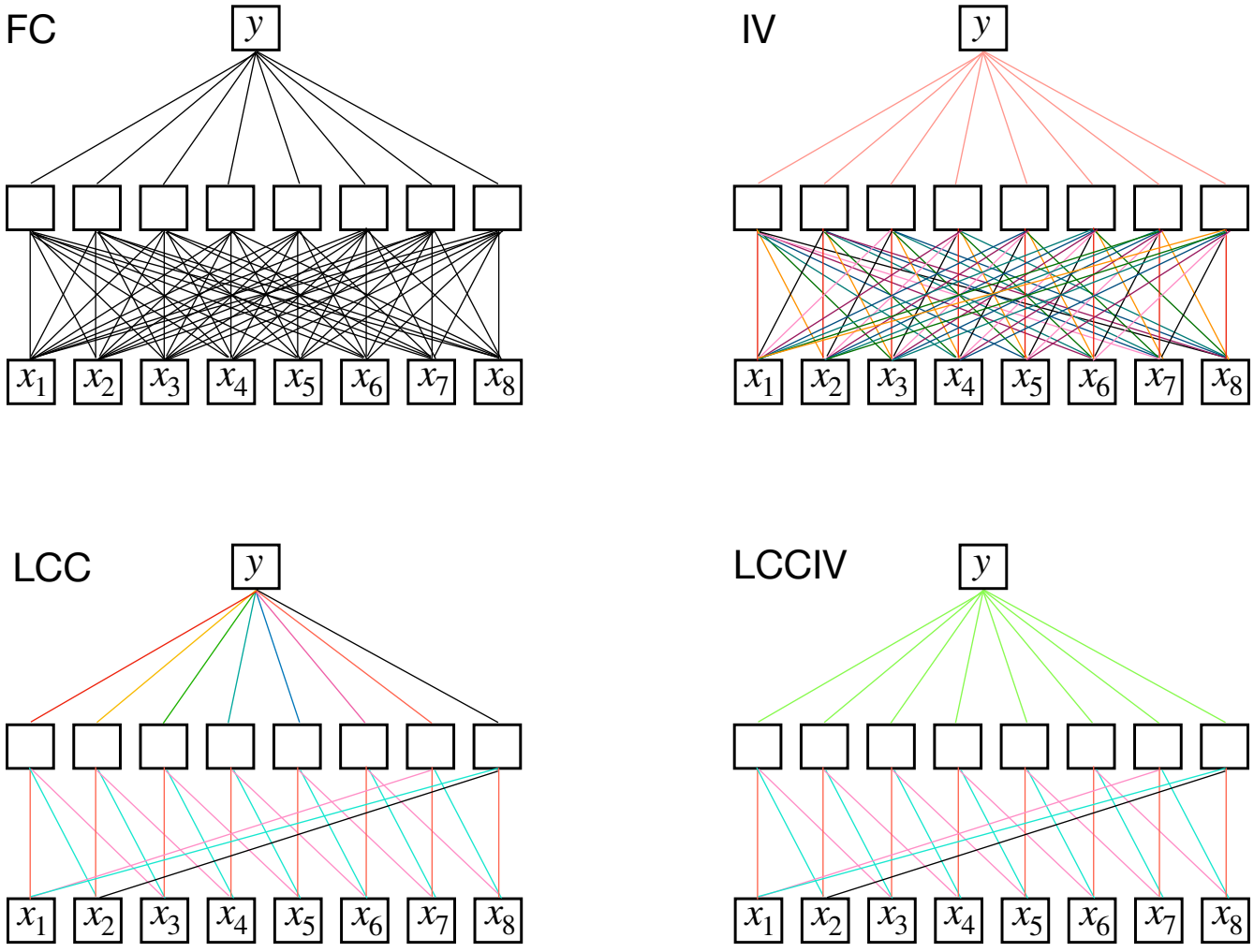
$$dq^{\ell-1} \ll n \ll dq^{\ell}$$

KRR with K_{LCCIV} requires sample size

$$q^{\ell-1} \ll n \ll q^{\ell}.$$

- K_{FC} requires sample size $d^{\ell} \ll n \ll d^{\ell+1}$,
- K_{IV} requires sample size $d^{\ell-1} \ll n \ll d^{\ell}$.

Four architectures



Comparison and numerical simulations

To fit a degree ℓ polynomial	K_{FC}	K_{IV}	K_{LCC}	K_{LCCIV}
Sample complexity	d^{ℓ}	$d^{\ell-1}$	$dq^{\ell-1}$	$q^{\ell-1}$
$\ell = 3$, $q = 10$, $d = 30$	$27,000$	900	3000	100

Table: Sample size n to fit a cyclic q-local polynomial of degree ℓ.

Comparison and numerical simulations

To fit a degree ℓ polynomial	K_{FC}	K_{IV}	K_{LCC}	K_{LCCIV}
Sample complexity	d^{ℓ}	$d^{\ell-1}$	$dq^{\ell-1}$	$q^{\ell-1}$
$\ell = 3$, $q = 10$, $d = 30$	27,000	900	3000	100

Table: Sample size n to fit a cyclic q-local polynomial of degree ℓ.

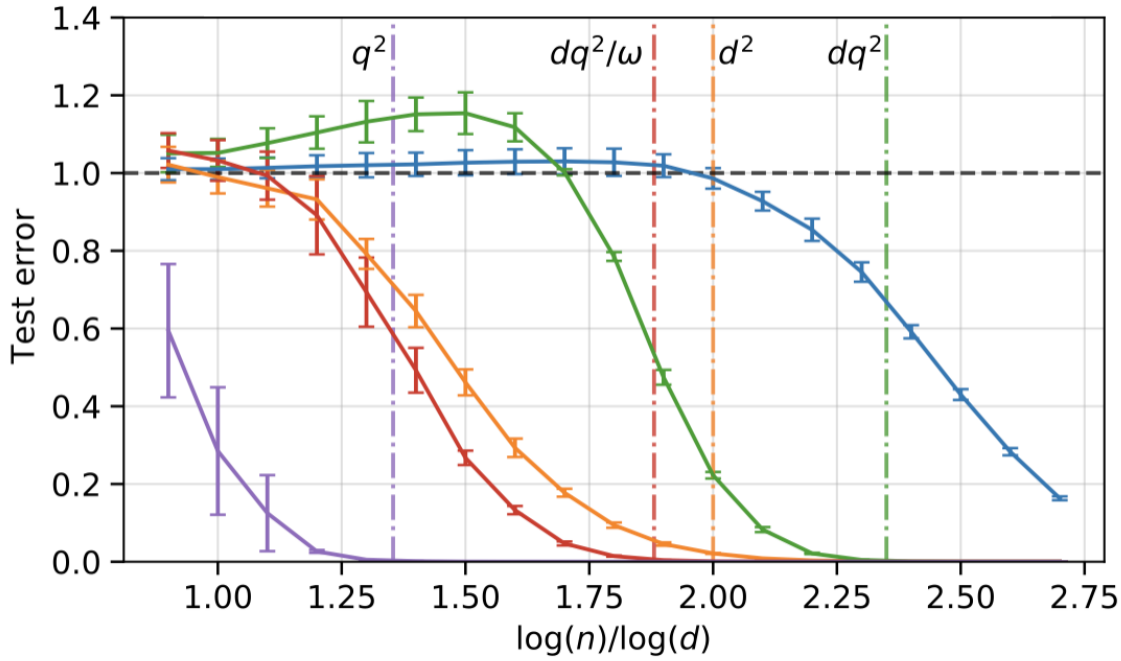


Figure: Simulation for fitting a cubic cyclic 3-local polynomial $f(x) = d^{-1/2} \sum_{i=1}^{d} x_i x_{i+1} x_{i+2}$. Here $d = 30$, $q = 10$.

The proof machinery

A general framework for analyzing the performance of random features regression and kernel ridge regression in the high dimensional regime.

Theorem (Mei, Misiakiewicz, Montanari, 2022 (Very informal))

Suppose the kernel has eigen-decomposition

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{z}).$$

Assume λ_i satisfies the "spectral gap" and "decaying" assumptions, and ψ_i satisfies "hypercontractivity" and "concentration" assumptions. Then KRR with kernel K with sample size n fits the top $O(n)$ eigenspace $\{\psi_i\}_{i \le O(n)}$.

The proof machinery

A general framework for analyzing the performance of random features regression and kernel ridge regression in the high dimensional regime.

Theorem (Mei, Misiakiewicz, Montanari, 2022 (Very informal))

Suppose the kernel has eigen-decomposition

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\boldsymbol{x}) \psi_i(\boldsymbol{z}).$$

Assume λ_i satisfies the "spectral gap" and "decaying" assumptions, and ψ_i satisfies "hypercontractivity" and "concentration" assumptions. Then KRR with kernel K with sample size n fits the top $O(n)$ eigenspace $\{\psi_i\}_{i \le O(n)}$.

Proof idea

Everything can be expressed as summation, product, and inversion of matrices. E.g.,

$$\mathbb{E}[f_\star(\boldsymbol{x})\mathsf{KRR}(\boldsymbol{x})] = \boldsymbol{u}^{\mathsf{T}}(K + \lambda I)^{-1}\boldsymbol{y}, \qquad \text{where } u_i = \mathbb{E}[f_\star(\boldsymbol{x})K(\boldsymbol{x}, \boldsymbol{x}_i)].$$

The technical difficulty lies in analyzing spectral properties of random matrices.

Comparison

Classical non-asymptotic results: Oracle inequality and minimax lower bound
[Caponnetto, de Vito, 2007], [Rahimi, Recht, 2009], [Bach, 2017], [E, Ma, Wu, 2018]

$$\mathrm{R}(f_d) \leq \min_{f_\star \in \mathcal{F}} \|f_d - f_\star\|^2 + \mathcal{G}(\delta_n; \mathcal{F}).$$

Comparison

Classical non-asymptotic results: Oracle inequality and minimax lower bound
[Caponnetto, de Vito, 2007], [Rahimi, Recht, 2009], [Bach, 2017], [E, Ma, Wu, 2018]

$$\mathrm{R}(f_d) \leq \min_{f_\star \in \mathcal{F}} \|f_d - f_\star\|^2 + \mathcal{G}(\delta_n; \mathcal{F}).$$

High dimensional asymptotic results:
[El Karoui, 2010], [Fan, and Montanari, 2017], [Ghorbani, Mei, Misiakiewicz, and Montanari, 2020], [Mei, Montanari, 2021]

$$\lim_{d,n \to \infty} \mathrm{R}(f_d) = R^\star.$$

Comparison

Classical non-asymptotic results: Oracle inequality and minimax lower bound
[Caponnetto, de Vito, 2007], [Rahimi, Recht, 2009], [Bach, 2017], [E, Ma, Wu, 2018]

$$\mathrm{R}(f_d) \leq \min_{f_\star \in \mathcal{F}} \|f_d - f_\star\|^2 + \mathcal{G}(\delta_n; \mathcal{F}).$$

High dimensional asymptotic results:
[El Karoui, 2010], [Fan, and Montanari, 2017], [Ghorbani, Mei, Misiakiewicz, and Montanari, 2020], [Mei, Montanari, 2021]

$$\lim_{d,n\to\infty} \mathrm{R}(f_d) = R^\star.$$

Difference between:

HD results	v.s.	Classical results
$n = d^k$ as $d \to \infty$,	v.s.	fixed d large n,
Constant asymptotic error,	v.s.	Vanishing upper bound,
Pointwise lower bound,	v.s.	Minimax lower bound.

Related works

- Analyzing kernel inner product matrices:
  [Ghorbani, Mei, Misiakiewicz, Montanari, 2019, 2020], [Misiakiewicz, 2022], [Hu, Lu, 2022], [Lu, Yau, 2022]

- Learning with invariance and locality in the classical regime:
  [Li, Zhang, Arora, 2020], [Bietti, Venturi, Bruna, 2021], [Bietti, 2021], [Favero, Cagnetta, and Wyart, 2021]

- Extension to multi-layer networks:
  [Xiao, 2021]

Open questions

- We assumed the image covariates x has an isotropic distribution, which is not realistic. The image covariates are sparse in the wavelet domain. Can we also model such properties of images and derive similar results?
- In classification tasks, how to characterize the interplay of invariance and locality in dataset and kernels?
- Consider non-linear neural networks. How to characterize the invariance and locality in neural network training?

Summary

To fit a degree ℓ polynomial	K_{FC}	K_{IV}	K_{LCC}	K_{LCCIV}
Sample complexity	d^ℓ	$d^{\ell-1}$	$dq^{\ell-1}$	$q^{\ell-1}$
$\ell = 3,\ q = 10,\ d = 30$	$27,000$	900	3000	100

Table: Sample size n to fit a cyclic q-local polynomial of degree ℓ.

- Cyclic kernels save a factor of d in learning cyclic functions.
  [Mei, Misiakiewicz, Montanari, 2021]
- Local kernels reduce the sample complexity from d^ℓ to dq^ℓ in learning q-local functions.
  [Misiakiewicz and Mei, 2021]
- General proof machineary.
  [Mei, Misiakiewicz, Montanari, 2022]

Thank you!

Backup Slides

Assumption ($\{n(d), \mathsf{m}(d)\}_{d\geq 1}$-Kernel Concentration Property)

We say that the sequence of operators $\{\mathbb{H}_d\}_{d\geq 1}$ *satisfies the Kernel Concentration Property (KCP) with respect to the sequence* $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$ *if there exists a sequence of integers* $\{u(d)\}_{d\geq 1}$ *with* $u(d) \geq \mathsf{m}(d)$ *such that the following conditions hold.*

(a) *(Hypercontractivity of finite eigenspaces.) For any fixed* $q \geq 1$*, there exists a constant* C *such that, for any* $h \in \mathcal{D}_{d,\leq u(d)} = \mathbf{span}(\psi_s, 1 \leq s \leq u(d))$*, we have*

$$\|h\|_{L^{2q}} \leq C \cdot \|h\|_{L^2}. \tag{1}$$

(b) *(Properly decaying eigenvalues.) There exists fixed* $\delta_0 > 0$*, such that, for all* d *large enough,*

$$n(d)^{2+\delta_0} \leq \frac{(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^8}, \tag{2}$$

$$n(d)^{2+\delta_0} \leq \frac{(\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^2)^2}{\sum_{j=u(d)+1}^{\infty} \lambda_{d,j}^4}. \tag{3}$$

(c) *(Concentration of diagonal elements of kernel) For* $(\boldsymbol{x}_i)_{i\in[n(d)]} \sim_{iid} \nu_d$*, we have:*

$$\max_{i\in[n(d)]} |\mathbb{E}_{\boldsymbol{x}\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i, \boldsymbol{x})^2] - \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x}')^2]| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim\nu_d}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x}')^2], \tag{4}$$

$$\max_{i\in[n(d)]} |H_{d,>\mathsf{m}(d)}(\boldsymbol{x}_i, \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x})]| = o_{d,\mathbb{P}}(1) \cdot \mathbb{E}_{\boldsymbol{x}}[H_{d,>\mathsf{m}(d)}(\boldsymbol{x}, \boldsymbol{x})]. \tag{5}$$

Back

Assumption (Eigenvalue condition at level $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$)

We say that the sequence of Kernel operators $\{\mathbb{H}_d\}_{d\geq 1}$ satisfies the Eigenvalue Condition at level $\{(n(d), \mathsf{m}(d))\}_{d\geq 1}$ if the following conditions hold for all d large enough.

(a) *There exists fixed $\delta_0 > 0$, such that*

$$n(d)^{1+\delta_0} \leq \frac{1}{\lambda^4_{d,\mathsf{m}(d)+1}} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda^4_{d,k}, \tag{6}$$

$$n(d)^{1+\delta_0} \leq \frac{1}{\lambda^2_{d,\mathsf{m}(d)+1}} \sum_{k=\mathsf{m}(d)+1}^{\infty} \lambda^2_{d,k}. \tag{7}$$

(b) *There exists fixed $\delta_0 > 0$, such that*

$$\mathsf{m}(d) \leq n(d)^{1-\delta_0}.$$