Statistical Benefits of Convolutional Kernels

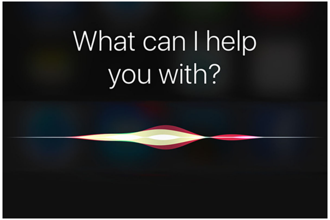
Alberto Bietti

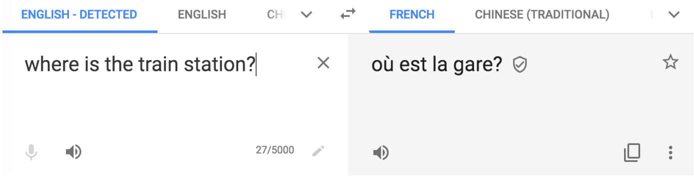
NYU

Physics/AI Journal Club, Harvard. March 8, 2022.

Success of deep learning

State-of-the-art models in various domains (images, speech, text, …)





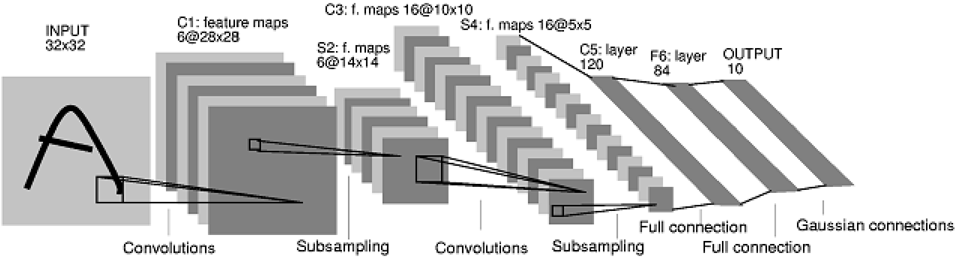
Success of deep learning

State-of-the-art models in various domains (images, speech, text, ...)

$$f(x) = W_n \sigma(W_{n-1} \cdots \sigma(W_1 x) \cdots)$$

Recipe: **huge models** + **lots of data** + **compute** + **simple algorithms**

Exploiting data structure through architectures

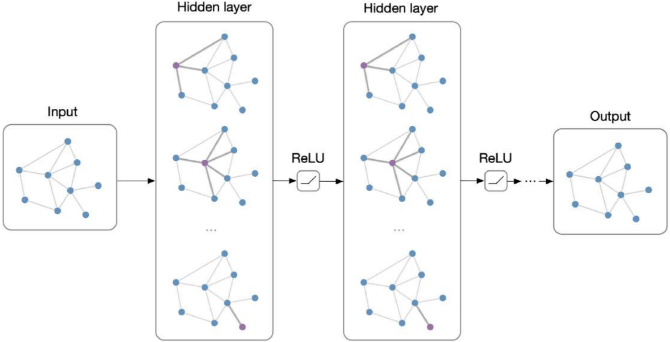


(LeCun et al., 1998)

Modern architectures (CNNs, GNNs, ...)

- Provide some invariance through pooling
- Model (local) interactions at different scales, hierarchically
- Useful **inductive biases** for learning efficiently on structured data

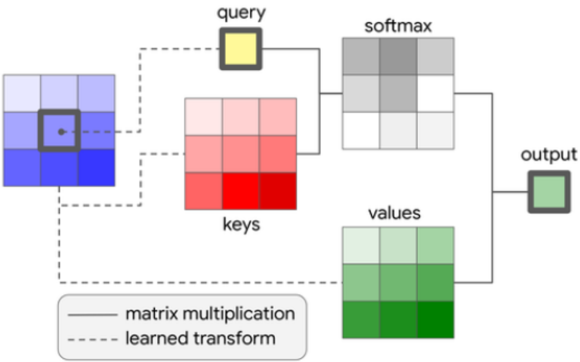
Exploiting data structure through architectures



Modern architectures (CNNs, GNNs, ...)

- Provide some invariance through pooling
- Model (local) interactions at different scales, hierarchically
- Useful **inductive biases** for learning efficiently on structured data

Exploiting data structure through architectures



Modern architectures (CNNs, GNNs, ...)

- Provide some invariance through pooling
- Model (local) interactions at different scales, hierarchically
- Useful **inductive biases** for learning efficiently on structured data

Understanding deep learning

The challenge of deep learning theory

- **Over-parameterized** (millions of parameters)
- **Expressive** (can approximate any function)
- Complex **architectures** for exploiting problem structure
- Yet, **easy to optimize** with (stochastic) gradient descent!

Understanding deep learning

The challenge of deep learning theory

- **Over-parameterized** (millions of parameters)
- **Expressive** (can approximate any function)
- Complex **architectures** for exploiting problem structure
- Yet, **easy to optimize** with (stochastic) gradient descent!

A functional space viewpoint

- View deep networks as functions in some functional space
- Non-parametric models, natural measures of complexity (*e.g.*, norms)
- Optimization performs **implicit regularization** towards

$$\min_f \Omega(f) \quad \text{s.t.} \quad y_i = f(x_i), \quad i = 1, \dots, n$$

Understanding deep learning

The challenge of deep learning theory

- **Over-parameterized** (millions of parameters)
- **Expressive** (can approximate any function)
- Complex **architectures** for exploiting problem structure
- Yet, **easy to optimize** with (stochastic) gradient descent!

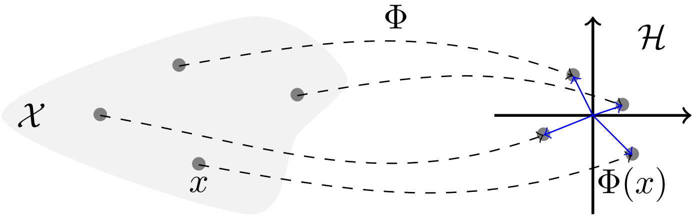
A functional space viewpoint

- View deep networks as functions in some functional space
- Non-parametric models, natural measures of complexity (*e.g.*, norms)
- Optimization performs **implicit regularization** towards

$$\min_f \Omega(f) \text{ s.t. } y_i = f(x_i), \quad i = 1, \dots, n$$

What is an appropriate functional space / norm Ω?

Kernels to the rescue



Kernels?

- Map data x to high-dimensional space, $\Phi(x) \in \mathcal{H}$ ($\mathcal{H}$: "RKHS")
- Functions $f \in \mathcal{H}$ are linear in features: $f(x) = \langle f, \Phi(x) \rangle$ (f can be non-linear in x!)
- Learning with a positive definite kernel $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$
  - $\mathcal{H}$ can be infinite-dimensional! (*kernel trick*)
  - Need to compute kernel matrix $K = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$, or approximations

Why kernels?

Clean and well-developed theory

- Tractable methods (convex optimization)
- Statistical and approximation properties well understood for many kernels
  - *e.g.*, smooth functions (Caponnetto and De Vito, 2007), interaction splines (Wahba, 1990)

Why kernels?

Clean and well-developed theory

- Tractable methods (convex optimization)
- Statistical and approximation properties well understood for many kernels
  - *e.g.*, smooth functions (Caponnetto and De Vito, 2007), interaction splines (Wahba, 1990)

This talk:

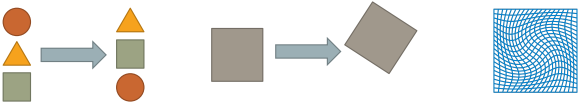
- Formal study of **convolutional kernels** and their RKHS
- **Benefits** of (deep) convolutional structure

Outline

1. Sample complexity under invariance and stability (B., Venturi, and Bruna, 2021)

2. Locality and depth (B., 2021)

Geometric priors

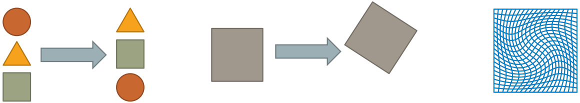


Functions $f : \mathcal{X} \to \mathbb{R}$ that are "smooth" along known transformations of input x

- *e.g.*, translations, rotations, permutations, deformations
- We consider: **permutations** $\sigma \in G$

$$(\sigma \cdot x)[u] = x[\sigma^{-1}(u)]$$

Geometric priors



Functions $f : \mathcal{X} \to \mathbb{R}$ that are "smooth" along known transformations of input x

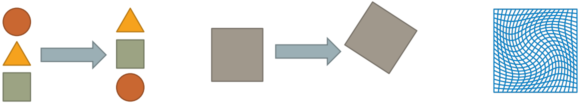
- *e.g.*, translations, rotations, permutations, deformations
- We consider: **permutations** $\sigma \in G$

$$(\sigma \cdot x)[u] = x[\sigma^{-1}(u)]$$

Group invariance: If G is a group (*e.g.*, cyclic shifts, all permutations), we want

$$f(\sigma \cdot x) = f(x), \quad \sigma \in G$$

Geometric priors



Functions $f : \mathcal{X} \to \mathbb{R}$ that are "smooth" along known transformations of input x

- *e.g.*, translations, rotations, permutations, deformations
- We consider: **permutations** $\sigma \in G$

$$(\sigma \cdot x)[u] = x[\sigma^{-1}(u)]$$

Group invariance: If G is a group (*e.g.*, cyclic shifts, all permutations), we want

$$f(\sigma \cdot x) = f(x), \quad \sigma \in G$$

Geometric stability: For other sets G (*e.g.*, local shifts, deformations), we want

$$f(\sigma \cdot x) \approx f(x), \quad \sigma \in G$$

Interlude: Kernels for Wide Shallow Networks

$$f(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \rho(\langle w_i, x \rangle)$$

Interlude: Kernels for Wide Shallow Networks

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{m}} \sum_{i=1}^m v_i \rho(\langle w_i, x \rangle) \\ &= \langle v, \varphi(x) \rangle, \qquad \text{with } \varphi(x) = \frac{1}{\sqrt{m}} \rho(Wx) \in \mathbb{R}^m \end{aligned}$$

Interlude: Kernels for Wide Shallow Networks

$$f(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \rho(\langle w_i, x \rangle)$$
$$= \langle v, \varphi(x) \rangle, \qquad \text{with } \varphi(x) = \frac{1}{\sqrt{m}} \rho(Wx) \in \mathbb{R}^m$$

- **Random Features** (RF, Neal, 1996; Rahimi and Recht, 2007): $w_i \sim \mathcal{N}(0, I)$, learn v

$$K_{RF}(x, x') = \lim_{m \to \infty} \langle \varphi(x), \varphi(x') \rangle$$
$$= \mathbb{E}_w[\rho(\langle w, x \rangle) \rho(\langle w, x' \rangle)] = \kappa_\rho(\langle x, x' \rangle) \;\; \text{when } x, x' \in \mathbb{S}^{d-1}$$

Interlude: Kernels for Wide Shallow Networks

$$f(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \rho(\langle w_i, x \rangle)$$
$$= \langle v, \varphi(x) \rangle, \qquad \text{with } \varphi(x) = \frac{1}{\sqrt{m}} \rho(Wx) \in \mathbb{R}^m$$

- **Random Features** (RF, Neal, 1996; Rahimi and Recht, 2007): $w_i \sim \mathcal{N}(0, I)$, learn v

$$K_{RF}(x, x') = \lim_{m \to \infty} \langle \varphi(x), \varphi(x') \rangle$$
$$= \mathbb{E}_w[\rho(\langle w, x \rangle) \rho(\langle w, x' \rangle)] = \kappa_\rho(\langle x, x' \rangle) \text{ when } x, x' \in \mathbb{S}^{d-1}$$

- A related kernel: **Neural Tangent Kernel** (NTK, Jacot et al., 2018): train both w_i and v_i near random initialization

Group-Invariant Models through Pooling

Pooling operator

$$S_G f(x) := \frac{1}{|G|} \sum_{\sigma \in G} f(\sigma \cdot x)$$

Convolutional network with pooling (group averaging)

$$f_G(x) = \langle v, \underbrace{\frac{1}{|G|} \sum_{\sigma \in G} \varphi(\sigma \cdot x)}_{\Phi(x)} \rangle, \qquad \text{with } \varphi(x) = \frac{1}{\sqrt{m}} \rho(Wx)$$

Group-Invariant Models through Pooling

Pooling operator

$$S_G f(x) := \frac{1}{|G|} \sum_{\sigma \in G} f(\sigma \cdot x)$$

Convolutional network with pooling (group averaging)

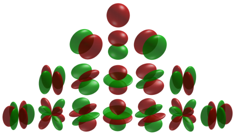
$$f_G(x) = \langle v, \underbrace{\frac{1}{|G|} \sum_{\sigma \in G} \varphi(\sigma \cdot x)}_{\Phi(x)} \rangle, \qquad \text{with } \varphi(x) = \frac{1}{\sqrt{m}} \rho(Wx)$$

Invariant kernel (Haasdonk and Burkhardt, 2007; Mroueh et al., 2015)

$$K_G(x, x') = \frac{1}{|G|} \sum_{\sigma \in G} \kappa(\langle \sigma \cdot x, x' \rangle), \quad \text{when } x, x' \in \mathbb{S}^{d-1}$$

- When $\kappa = \kappa_\rho$, this corresponds to Random Features kernel for f_G

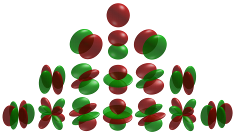
Harmonic analysis on the sphere



- τ: uniform distribution on the sphere $\mathbb{S}^{d-1}$
- $L^2(\tau)$ basis of **spherical harmonics** $Y_{k,j}$
- $N(d,k)$ harmonics of degree k, form a basis of $V_{d,k}$

Harmonic analysis on the sphere

- τ: uniform distribution on the sphere $\mathbb{S}^{d-1}$
- $L^2(\tau)$ basis of **spherical harmonics** $Y_{k,j}$
- $N(d,k)$ harmonics of degree k, form a basis of $V_{d,k}$



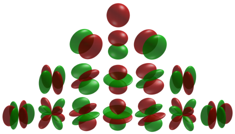
Dot-product kernels and their RKHS $\quad K(x,x') = \kappa(\langle x, x' \rangle)$

$$\mathcal{H} = \left\{ f = \sum_{k=0}^{\infty} \sum_{j=1}^{N(d,k)} a_{k,j} Y_{k,j}(\cdot) \text{ s.t. } \|f\|_{\mathcal{H}}^2 := \sum_{k,j} \frac{a_{k,j}^2}{\mu_k} < \infty \right\}$$

- **integral operator**: $T_K f(x) = \int \kappa(\langle x, y \rangle) f(y) d\tau(y)$
- $\mu_k = c_d \int_{-1}^{1} \kappa(t) P_{d,k}(t) (1-t^2)^{\frac{d-3}{2}} dt$: eigenvalues of T_K, with multiplicity $N(d,k)$
- $P_{d,k}$: **Legendre**/**Gegenbauer** polynomial

Harmonic analysis on the sphere

- τ: uniform distribution on the sphere $\mathbb{S}^{d-1}$
- $L^2(\tau)$ basis of **spherical harmonics** $Y_{k,j}$
- $N(d,k)$ harmonics of degree k, form a basis of $V_{d,k}$



Dot-product kernels and their RKHS $K(x,x') = \kappa(\langle x, x' \rangle)$

$$\mathcal{H} = \left\{ f = \sum_{k=0}^{\infty} \sum_{j=1}^{N(d,k)} a_{k,j} Y_{k,j}(\cdot) \text{ s.t. } \|f\|_{\mathcal{H}}^2 := \sum_{k,j} \frac{a_{k,j}^2}{\mu_k} < \infty \right\}$$

- **integral operator**: $T_K f(x) = \int \kappa(\langle x, y \rangle) f(y) d\tau(y)$
- $\mu_k = c_d \int_{-1}^{1} \kappa(t) P_{d,k}(t)(1-t^2)^{\frac{d-3}{2}} dt$: eigenvalues of T_K, with multiplicity $N(d,k)$
- $P_{d,k}$: **Legendre/Gegenbauer** polynomial
- **decay** $\leftrightarrow$ **regularity**: $\mu_k \asymp k^{-2\beta} \leftrightarrow \|f\|_{\mathcal{H}} = \|T_K^{-1/2} f\|_{L^2(\tau)} \approx \|\Delta_{\mathbb{S}^{d-1}}^{\beta/2} f\|_{L^2(\tau)}$

Invariant harmonics

Key properties of S_G for group-invariant case (Mei, Misiakiewicz, and Montanari, 2021)

- S_G acts as projection from $V_{d,k}$ (dim $N(d,k)$) to $\overline{V}_{d,k}$ (dim $\overline{N}(d,k)$)

Invariant harmonics

Key properties of S_G for group-invariant case (Mei, Misiakiewicz, and Montanari, 2021)

- S_G acts as projection from $V_{d,k}$ (dim $N(d,k)$) to $\overline{V}_{d,k}$ (dim $\overline{N}(d,k)$)
- The number of invariant spherical harmonics $\overline{N}$ can be estimated using:

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

- We have $T_{K_G} = S_G T_K$

Invariant harmonics

Key properties of S_G for group-invariant case (Mei, Misiakiewicz, and Montanari, 2021)

- S_G acts as projection from $V_{d,k}$ (dim $N(d,k)$) to $\overline{V}_{d,k}$ (dim $\overline{N}(d,k)$)
- The number of invariant spherical harmonics $\overline{N}$ can be estimated using:

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

- We have $T_{K_G} = S_G T_K$

Previous work (Mei et al., 2021)

- High-dimensional regime $d \to \infty$ with $n \asymp d^s$
- $\gamma_d(k) = \Theta_d(d^{-\alpha}) \implies$ sample complexity gain by factor d^α
- Studied for translations: gains by a factor d

Invariant harmonics

Key properties of S_G for group-invariant case (Mei, Misiakiewicz, and Montanari, 2021)

- S_G acts as projection from $V_{d,k}$ (dim $N(d,k)$) to $\overline{V}_{d,k}$ (dim $\overline{N}(d,k)$)
- The number of invariant spherical harmonics $\overline{N}$ can be estimated using:

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

- We have $T_{K_G} = S_G T_K$

Previous work (Mei et al., 2021)

- High-dimensional regime $d \to \infty$ with $n \asymp d^s$
- $\gamma_d(k) = \Theta_d(d^{-\alpha}) \implies$ sample complexity gain by factor d^α
- Studied for translations: gains by a factor d
- **Beyond translations? What about groups/sets G exponential in d?**

Invariant harmonics

Key properties of S_G for group-invariant case (Mei, Misiakiewicz, and Montanari, 2021)

- S_G acts as projection from $V_{d,k}$ (dim $N(d,k)$) to $\overline{V}_{d,k}$ (dim $\overline{N}(d,k)$)
- The number of invariant spherical harmonics $\overline{N}$ can be estimated using:

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

- We have $T_{K_G} = S_G T_K$

Previous work (Mei et al., 2021)

- High-dimensional regime $d \to \infty$ with $n \asymp d^s$
- $\gamma_d(k) = \Theta_d(d^{-\alpha}) \implies$ sample complexity gain by factor d^α
- Studied for translations: gains by a factor d
- **Beyond translations? What about groups/sets G exponential in d?**
- tl;dr: we consider d fixed, $n \to \infty$, show (asymptotic) **gains by a factor** $|G|$

Counting invariant harmonics

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

Proposition ((B., Venturi, and Bruna, 2021))

As $k \to \infty$, we have

$$\gamma_d(k) = \frac{1}{|G|} + O(k^{-d+\chi}),$$

where χ is the maximal number of cycles of any permutation $\sigma \in G \setminus \{Id\}$.

Counting invariant harmonics

$$\gamma_d(k) := \frac{\overline{N}(d,k)}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)].$$

Proposition ((B., Venturi, and Bruna, 2021))

As $k \to \infty$, we have

$$\gamma_d(k) = \frac{1}{|G|} + O(k^{-d+\chi}),$$

where χ is the maximal number of cycles of any permutation $\sigma \in G \setminus \{Id\}$.

- Relies on singularity analysis of density of $\langle \sigma \cdot x, x \rangle$ (Saldanha and Tomei, 1996)
  - Decay $\leftrightarrow$ nature of singularities $\leftrightarrow$ eigenvalue multiplicities $\leftrightarrow$ cycle statistics
- χ can be large ($= d - 1$) for some groups (*e.g.*, $\sigma = (1\ 2)$)
- Can use upper bounds with faster decays but larger constants

Counting invariant harmonics: examples

Translations (cyclic group)

$$\gamma_d(k) = d^{-1} + O(k^{-d/2+6})$$

Only linear gain in d, but with a fast rate

Counting invariant harmonics: examples

Translations (cyclic group)

$$\gamma_d(k) = d^{-1} + O(k^{-d/2+6})$$

Only linear gain in d, but with a fast rate

Block translations: $d = s \cdot r$, with r cycles of length s

$$\gamma_d(k) = \frac{1}{s^r} + O(k^{-s/2+1})$$

For $s = 2$, exponential gains ($|G| = 2^{d/2}$) but slow rate

Counting invariant harmonics: examples

Translations (cyclic group)

$$\gamma_d(k) = d^{-1} + O(k^{-d/2+6})$$

Only linear gain in d, but with a fast rate

Block translations: $d = s \cdot r$, with r cycles of length s

$$\gamma_d(k) = \frac{1}{s^r} + O(k^{-s/2+1})$$

For $s = 2$, exponential gains ($|G| = 2^{d/2}$) but slow rate

Full permutation group: For any s,

$$\gamma_d(k) \leq \frac{2}{(s+1)!} + O(k^{-d/2+\max(s/2,6)})$$

For $s = d/2$, exponential gains with fast rate

Sample complexity of invariant kernel: assumptions

Kernel Ridge Regression

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_G} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_G}^2$$

Problem assumptions

- (data) $x \sim \tau$, $\mathbb{E}[y|x] = f^*(x)$, $\mathrm{Var}(y|x) \leq \sigma^2$
- (G-invariance) f^* is G-invariant

Sample complexity of invariant kernel: assumptions

Kernel Ridge Regression

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_G} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_G}$$

Problem assumptions

- (data) $x \sim \tau$, $\mathbb{E}[y|x] = f^*(x)$, $\mathrm{Var}(y|x) \leq \sigma^2$
- (G-invariance) f^* is G-invariant
- (*capacity*) $\lambda_m(T_K) \leq C_K m^{-\alpha}$
  - *e.g.*, $\alpha = \frac{2s}{d-1}$ for Sobolev space of order s with $s > \frac{d-1}{2}$

Sample complexity of invariant kernel: assumptions

Kernel Ridge Regression

$$\hat{f}_\lambda := \arg\min_{f \in \mathcal{H}_G} \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_G}$$

Problem assumptions

- (data) $x \sim \tau$, $\mathbb{E}[y|x] = f^*(x)$, $\mathrm{Var}(y|x) \leq \sigma^2$
- (G-invariance) f^* is G-invariant
- (*capacity*) $\lambda_m(T_K) \leq C_K m^{-\alpha}$
  - *e.g.*, $\alpha = \frac{2s}{d-1}$ for Sobolev space of order s with $s > \frac{d-1}{2}$
- (*source*) $\|T_K^{-r} f^*\|_{L^2} \leq C_{f^*}$
  - *e.g.*, if $2\alpha r = \frac{2s}{d-1}$, f^* belongs to Sobolev space of order s

Sample complexity of invariant kernel: generalization

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} \overline{N}(d,k) \lesssim \nu_d(\ell)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$, *where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$.

$$\mathbb{E}\, \|\hat{f} - f^*\|^2_{L^2(d\tau)} \leq C \left(\frac{\nu_d(\ell_n)}{n} \right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Sample complexity of invariant kernel: generalization

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} \overline{N}(d,k) \lesssim \nu_d(\ell)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$*, where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$*.*

$$\mathbb{E}\,\|\hat{f} - f^*\|^2_{L^2(d\tau)} \leq C \left(\frac{\nu_d(\ell_n)}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Replace $\nu_d(\ell_n)$ *by* 1 *for non-invariant kernel.*

Sample complexity of invariant kernel: generalization

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} \overline{N}(d,k) \lesssim \nu_d(\ell)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$*, where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$*.*

$$\mathbb{E}\,\|\hat{f} - f^*\|^2_{L^2(d\tau)} \leq C \left(\frac{\nu_d(\ell_n)}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Replace $\nu_d(\ell_n)$ *by* 1 *for non-invariant kernel.*

- We have $\nu_d(\ell_n) = \frac{1}{|G|} + O\left(n^{\frac{-\beta}{(d-1)(2\alpha r+1)+2\beta\alpha r}}\right)$ when $\gamma_d(k) = 1/|G| + O(k^{-\beta})$
- $\implies$ **Improvement in sample complexity** by a factor $|G|$!

Sample complexity of invariant kernel: generalization

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} \overline{N}(d,k) \lesssim \nu_d(\ell)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$*, where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$.

$$\mathbb{E}\,\|\hat{f} - f^*\|^2_{L^2(d\tau)} \leq C \left(\frac{\nu_d(\ell_n)}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Replace $\nu_d(\ell_n)$ *by* 1 *for non-invariant kernel.*

- We have $\nu_d(\ell_n) = \frac{1}{|G|} + O\left(n^{\frac{-\beta}{(d-1)(2\alpha r+1)+2\beta\alpha r}}\right)$ when $\gamma_d(k) = 1/|G| + O(k^{-\beta})$
- $\implies$ **Improvement in sample complexity** by a factor $|G|$!
- C may depend on d, but is **optimal** in a minimax sense over non-invariant f^*

Sample complexity of invariant kernel: generalization

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} \overline{N}(d,k) \lesssim \nu_d(\ell)^{\frac{2\alpha r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$*, where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$.

$$\mathbb{E}\,\|\hat{f} - f^*\|^2_{L^2(d\tau)} \leq C \left(\frac{\nu_d(\ell_n)}{n}\right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Replace $\nu_d(\ell_n)$ *by* 1 *for non-invariant kernel.*

- We have $\nu_d(\ell_n) = \frac{1}{|G|} + O\left(n^{\frac{-\beta}{(d-1)(2\alpha r+1)+2\beta\alpha r}}\right)$ when $\gamma_d(k) = 1/|G| + O(k^{-\beta})$
- $\implies$ **Improvement in sample complexity** by a factor $|G|$!
- C may depend on d, but is **optimal** in a minimax sense over non-invariant f^*
- Main ideas:
  - Approximation error: same as non-invariant kernel
  - Estimation error: pick ℓ_n such that $\mathcal{N}_{K_G}(\lambda_n) \lesssim \nu_d(\ell_n)\mathcal{N}_K(\lambda_n)$ ($\mathcal{N}(\lambda_n)$: degrees of freedom)

Synthetic experiments

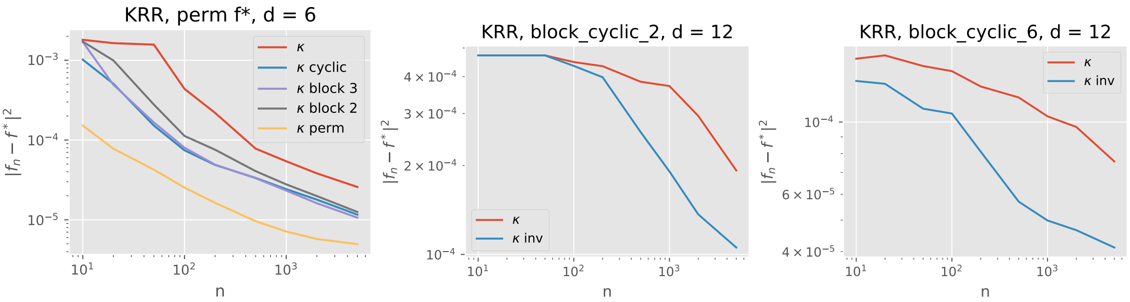
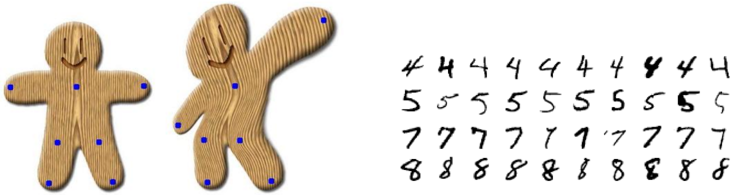


Figure: Comparison of KRR with invariant and non-invariant kernels.

Geometric stability to deformations

Deformations

- $\phi : \Omega \to \Omega$: C^1-diffeomorphism (*e.g.*, $\Omega = \mathbb{R}^2$)
- $\phi \cdot x(u) = x(\phi^{-1}(u))$: action operator
- Much richer group of transformations than translations



- Studied for wavelet-based scattering transform (Mallat, 2012; Bruna and Mallat, 2013)

Geometric stability to deformations

Deformations

- $\phi : \Omega \to \Omega$: C^1-diffeomorphism (*e.g.*, $\Omega = \mathbb{R}^2$)
- $\phi \cdot x(u) = x(\phi^{-1}(u))$: action operator
- Much richer group of transformations than translations

Geometric stability

- A function $f(\cdot)$ is **stable** (Mallat, 2012) if:

$$f(\phi \cdot x) \approx f(x) \quad \text{when} \quad \|\nabla \phi - I\|_\infty \leq \epsilon$$

- In particular, near-invariance to translations ($\nabla \phi = I$)

Geometric stability to deformations

Deformations

- $\phi : \Omega \to \Omega$: C^1-diffeomorphism (*e.g.*, $\Omega = \mathbb{R}^2$)
- $\phi \cdot x(u) = x(\phi^{-1}(u))$: action operator
- Much richer group of transformations than translations

Toy model for deformations ("small $\|\nabla\sigma - Id\|$")

$$G_\epsilon := \{\sigma \in \mathcal{S}_d \ : \ |\sigma(u) - \sigma(u') - (u - u')| \leq \varepsilon |u - u'|\}$$

- For $\epsilon = 2$, we have $\gamma_d(k) \leq \tau^d + O(k^{-\Theta(d)})$, with $\tau < 1$
  - gains by a factor **exponential** in d with a fast rate

Stability

- S_G is no longer a projection, but its eigenvalues $\lambda_{k,j}$ on $V_{d,k}$ satisfy

$$\gamma_d(k) := \frac{\sum_{j=1}^{N(d,k)} \lambda_{k,j}}{N(d,k)} = \frac{1}{|G|} \sum_{\sigma \in G} \mathbb{E}_x[P_{d,k}(\langle \sigma \cdot x, x \rangle)]$$

- Source condition adapted to S_G: $f^* = S_G^r T_K^r g^*$ with $\|g^*\|_{L^2} \leq C_{f^*}$

Theorem ((B., Venturi, and Bruna, 2021))

Let $\ell_n := \sup\{\ell : \sum_{k \leq \ell} N(d,k) \lesssim \nu_d(\ell)^{\frac{2r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$, *where* $\nu_d(\ell) := \sup_{k \geq \ell} \gamma_d(k)$.

$$\mathbb{E}\,\|\hat{f} - f^*\|^2_{L^2(\tau)} \leq C \left(\frac{\nu_d(\ell_n)^{1/\alpha}}{n} \right)^{\frac{2\alpha r}{2\alpha r+1}}$$

Discussion

Curse of dimensionality

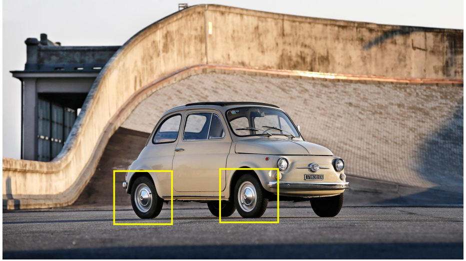
- For Lipschitz targets, cursed rate $n^{-\frac{2\alpha r}{2\alpha r+1}} = n^{-\frac{2}{2+d-1}}$ (unimprovable)
- Improving this rate requires more structural assumptions, which may be exploited with adaptivity (Bach, 2017), or better architectures (up next!)

- Gains are asymptotic, can we get non-asymptotic?
- For large groups, pooling is computationally costly
  - More structure may help, *e.g.*, stability through depth (B. and Mairal, 2019; Bruna and Mallat, 2013; Mallat, 2012)

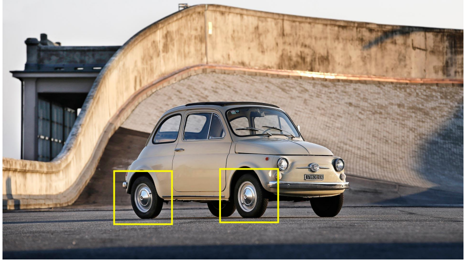
Outline

1. Sample complexity under invariance and stability (B., Venturi, and Bruna, 2021)
2. Locality and depth (B., 2021)

Locality

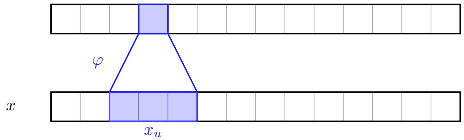


Locality



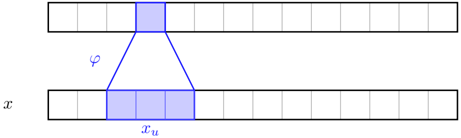
Q: Can locality improve statistical efficiency?

One-Layer Convolutional Kernels on Patches



- 1D signal: $x[u]$, $u \in \Omega$
- **Patches**: $x_u = (x[u], \ldots, x[u + p - 1]) \in \mathbb{R}^p$, features $\varphi(x_u) = \frac{1}{\sqrt{m}} \rho(W x_u)$, $m \to \infty$

One-Layer Convolutional Kernels on Patches



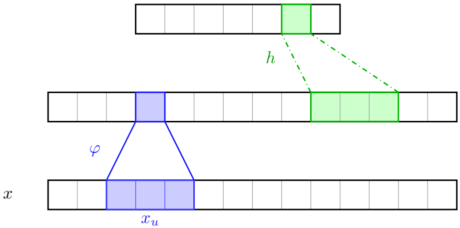
- 1D signal: $x[u]$, $u \in \Omega$
- **Patches**: $x_u = (x[u], \dots, x[u+p-1]) \in \mathbb{R}^p$, features $\varphi(x_u) = \frac{1}{\sqrt{m}} \rho(W x_u)$, $m \to \infty$
- **Convolutional network**:

$$f(x) = \sum_{u \in \Omega} \langle v_u, \varphi(x_u) \rangle =: \langle v, \Phi(x) \rangle$$

- **Convolutional kernel** (with $k(z, z') = \langle \varphi(z), \varphi(z') \rangle$ the RF **patch kernel**)

$$K(x, x') = \sum_{u \in \Omega} k(x_u, x'_u)$$

One-Layer Convolutional Kernels on Patches



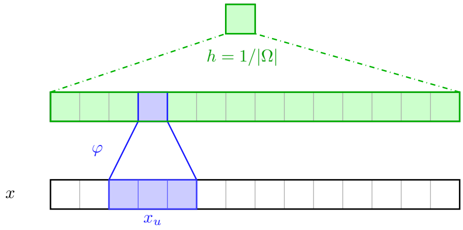
- 1D signal: $x[u]$, $u \in \Omega$
- **Patches**: $x_u = (x[u], \dots, x[u+p-1]) \in \mathbb{R}^p$, features $\varphi(x_u) = \frac{1}{\sqrt{m}}\rho(Wx_u)$, $m \to \infty$
- **Convolutional network**: with pooling filter h

$$f_h(x) = \sum_{u \in \Omega} \langle v_u, \sum_v h[u-v]\varphi(x_v) \rangle$$

- **Convolutional kernel** (with $k(z, z') = \langle \varphi(z), \varphi(z') \rangle$ the RF **patch kernel**)

$$K_h(x, x') = \sum_{u \in \Omega} \sum_{v, v'} h[u-v] h[u-v'] k(x_v, x'_{v'})$$

One-Layer Convolutional Kernels on Patches



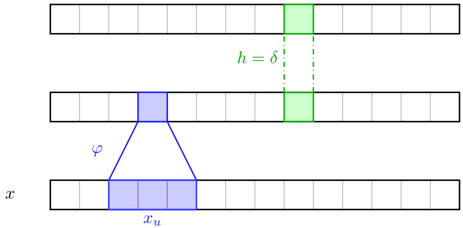
- 1D signal: $x[u]$, $u \in \Omega$
- **Patches**: $x_u = (x[u], \ldots, x[u+p-1]) \in \mathbb{R}^p$, features $\varphi(x_u) = \frac{1}{\sqrt{m}}\rho(Wx_u)$, $m \to \infty$
- **Convolutional network**: with global pooling ($h = 1/|\Omega|$)

$$f_h(x) = \sum_{u \in \Omega} \langle v_u, |\Omega|^{-1} \sum_v \varphi(x_v) \rangle$$

- **Convolutional kernel** (with $k(z, z') = \langle \varphi(z), \varphi(z') \rangle$ the RF **patch kernel**)

$$K_h(x, x') = |\Omega|^{-1} \sum_{v, v'} k(x_v, x'_{v'})$$

One-Layer Convolutional Kernels on Patches



- 1D signal: $x[u]$, $u \in \Omega$
- **Patches**: $x_u = (x[u], \dots, x[u + p - 1]) \in \mathbb{R}^p$, features $\varphi(x_u) = \frac{1}{\sqrt{m}} \rho(W x_u)$, $m \to \infty$
- **Convolutional network**: with no pooling (Dirac $h = \delta$)

$$f_h(x) = \sum_{u \in \Omega} \langle v_u, \varphi(x_u) \rangle$$

- **Convolutional kernel** (with $k(z, z') = \langle \varphi(z), \varphi(z') \rangle$ the RF **patch kernel**)

$$K_h(x, x') = \sum_{u \in \Omega} k(x_u, x'_u)$$

Benefits of Locality and Pooling

- Assume **additive**, **invariant** target $f^*(x) = \sum_{u \in \Omega} g^*(x_u)$
- Consider the kernels:

$$\text{(global pool) } K_g(x, x') = \sum_{v, v'} k(x_v, x'_{v'}) \quad \text{vs} \quad \text{(no pool) } K_\delta(x, x') = \sum_u k(x_u, x'_u)$$

Benefits of Locality and Pooling

- Assume **additive**, **invariant** target $f^*(x) = \sum_{u \in \Omega} g^*(x_u)$
- Consider the kernels:

$$\text{(global pool)}\ K_g(x, x') = \sum_{v, v'} k(x_v, x'_{v'}) \qquad \text{vs} \qquad \text{(no pool)}\ K_\delta(x, x') = \sum_u k(x_u, x'_u)$$

Theorem (Statistical rates with one-layer (B., 2021))

*Assume g^* is s-**smooth**, non-overlapping patches on $\mathbb{S}^{p-1}$. KRR with K_h yields*

$$\mathbb{E}\, R(\hat{f}_{g,n}) - R(f^*) \leq C_p \left(\frac{1}{n}\right)^{\frac{2s}{2s+p-1}} \quad \text{vs} \quad \mathbb{E}\, R(\hat{f}_{\delta,n}) - R(f^*) \leq C_p \left(\frac{|\Omega|}{n}\right)^{\frac{2s}{2s+p-1}}$$

Benefits of Locality and Pooling

- Assume **additive**, **invariant** target $f^*(x) = \sum_{u \in \Omega} g^*(x_u)$
- Consider the kernels:

$$\text{(global pool) } K_g(x, x') = \sum_{v,v'} k(x_v, x'_{v'}) \quad \text{vs} \quad \text{(no pool) } K_\delta(x, x') = \sum_{u} k(x_u, x'_u)$$

Theorem (Statistical rates with one-layer (B., 2021))

*Assume g^* is s-**smooth**, non-overlapping patches on $\mathbb{S}^{p-1}$. KRR with K_h yields*

$$\mathbb{E}\, R(\hat{f}_{g,n}) - R(f^*) \leq C_p \left(\frac{1}{n}\right)^{\frac{2s}{2s+p-1}} \quad \text{vs} \quad \mathbb{E}\, R(\hat{f}_{\delta,n}) - R(f^*) \leq C_p \left(\frac{|\Omega|}{n}\right)^{\frac{2s}{2s+p-1}}$$

- Patch dimension $p \ll d = p|\Omega|$ in the rate (**breaks the curse**!)

Benefits of Locality and Pooling

- Assume **additive**, **invariant** target $f^*(x) = \sum_{u \in \Omega} g^*(x_u)$
- Consider the kernels:

$$\text{(global pool)}\ K_g(x, x') = \sum_{v, v'} k(x_v, x'_{v'}) \qquad \text{vs} \qquad \text{(no pool)}\ K_\delta(x, x') = \sum_u k(x_u, x'_u)$$

Theorem (Statistical rates with one-layer (B., 2021))

*Assume g^* is s-**smooth**, non-overlapping patches on $\mathbb{S}^{p-1}$. KRR with K_h yields*

$$\mathbb{E}\, R(\hat{f}_{g,n}) - R(f^*) \leq C_p \left(\frac{1}{n}\right)^{\frac{2s}{2s+p-1}} \quad vs \quad \mathbb{E}\, R(\hat{f}_{\delta,n}) - R(f^*) \leq C_p \left(\frac{|\Omega|}{n}\right)^{\frac{2s}{2s+p-1}}$$

- Patch dimension $p \ll d = p|\Omega|$ in the rate (**breaks the curse**!)
- With localized pooling, we can also learn $f^*(x) = \sum_{u \in \Omega} g^*_u(x_u)$ with different g^*_u

Benefits of Locality and Pooling

- Assume **additive**, **invariant** target $f^*(x) = \sum_{u \in \Omega} g^*(x_u)$
- Consider the kernels:

$$\text{(global pool) } K_g(x, x') = \sum_{v, v'} k(x_v, x'_{v'}) \quad \text{vs} \quad \text{(no pool) } K_\delta(x, x') = \sum_u k(x_u, x'_u)$$

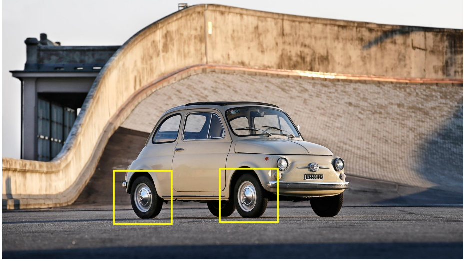
Theorem (Statistical rates with one-layer (B., 2021))

*Assume g^* is s-**smooth**, non-overlapping patches on $\mathbb{S}^{p-1}$. KRR with K_h yields*

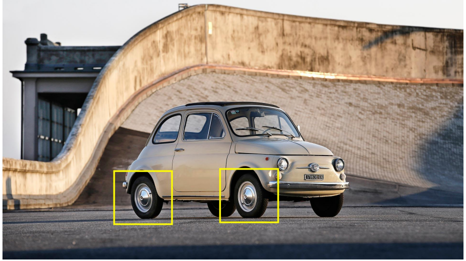
$$\mathbb{E}\, R(\hat{f}_{g,n}) - R(f^*) \leq C_p \left(\frac{1}{n}\right)^{\frac{2s}{2s+p-1}} \quad vs \quad \mathbb{E}\, R(\hat{f}_{\delta,n}) - R(f^*) \leq C_p \left(\frac{|\Omega|}{n}\right)^{\frac{2s}{2s+p-1}}$$

- Patch dimension $p \ll d = p|\Omega|$ in the rate (**breaks the curse**!)
- With localized pooling, we can also learn $f^*(x) = \sum_{u \in \Omega} g^*_u(x_u)$ with different g^*_u
- For overlapping patches, see (Favero et al., 2021; Misiakiewicz and Mei, 2021)

Long-Range Interactions

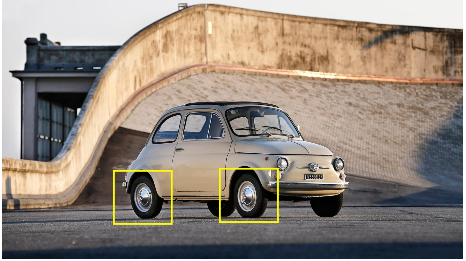


Long-Range Interactions



Q: How to capture interactions between multiple patches?

Long-Range Interactions

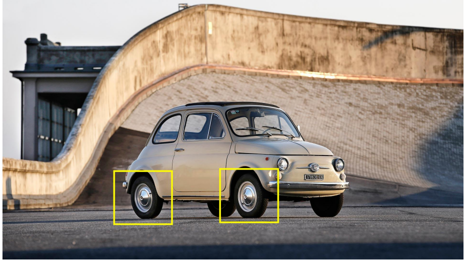


Q: How to capture interactions between multiple patches?

$\rightarrow$ **"add more layers"! Hierarchical kernels** (Cho and Saul, 2009):

$$K(x, x') = \langle \varphi_2(\varphi_1(x)), \varphi_2(\varphi_1(x')) \rangle$$

Long-Range Interactions

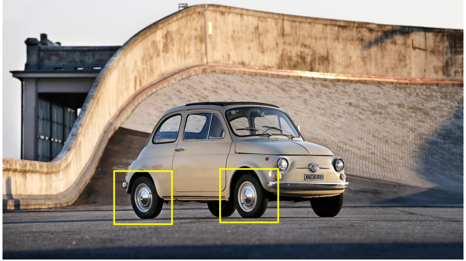


Q: How to capture interactions between multiple patches?

$\rightarrow$ **"add more layers"! Hierarchical kernels** (Cho and Saul, 2009):

$$K(x, x') = \kappa_2(\langle \varphi_1(x), \varphi_1(x') \rangle)$$

Long-Range Interactions



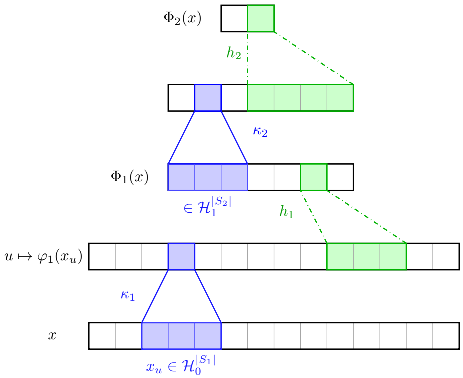
Q: How to capture interactions between multiple patches?

→ **"add more layers"! Hierarchical kernels** (Cho and Saul, 2009):

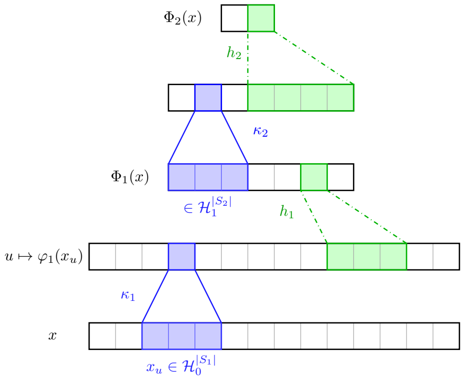
$$K(x, x') = \kappa_2(\kappa_1(\langle x, x' \rangle))$$

RKHS of Two-Layer Convolutional Kernels (B., 2021)

- φ_2/κ_2 captures **interactions** between patches



RKHS of Two-Layer Convolutional Kernels (B., 2021)

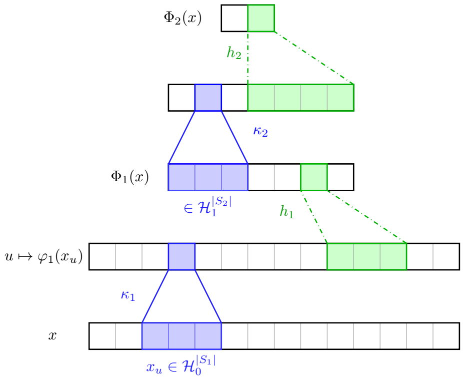


- φ_2/κ_2 captures **interactions** between patches
- Take $\kappa_2(u) = u^2$. RKHS contains

$$f(x) = \sum_{|u-v| \leq r} g_{u,v}(x_u, x_v)$$

- Receptive field r depends on h_1 and s_2
- $g_{u,v} \in \mathcal{H}_1 \otimes \mathcal{H}_1$

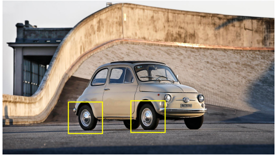
RKHS of Two-Layer Convolutional Kernels (B., 2021)



- φ_2/κ_2 captures **interactions** between patches
- Take $\kappa_2(u) = u^2$. RKHS contains

$$f(x) = \sum_{|u-v|\leq r} g_{u,v}(x_u, x_v)$$

- Receptive field r depends on h_1 and s_2
- $g_{u,v} \in \mathcal{H}_1 \otimes \mathcal{H}_1$



- Pooling h_1: invariance to **relative** position
- Pooling h_2: invariance to **global** position

Is it a Good Model for Cifar10? (B., 2021)

2-layers, patch sizes (3, 5), Gaussian pooling factors (2,5).

κ_1	κ_2	Test acc.
Gauss	Gauss	88.3%
Gauss	Poly4	88.3%
Gauss	Poly3	88.2%
Gauss	Poly2	87.4%
Gauss	Linear	80.9%

Is it a Good Model for Cifar10? (B., 2021)

2-layers, patch sizes (3, 5), Gaussian pooling factors (2,5).

κ_1	κ_2	Test acc.
Gauss	Gauss	88.3%
Gauss	Poly4	88.3%
Gauss	Poly3	88.2%
Gauss	Poly2	87.4%
Gauss	Linear	80.9%

- **Polynomial kernels at second layer suffice!**
- **State-of-the-art for kernels on Cifar10** (at a large computational cost...)
  - Shankar et al. (2020): 88.2% with 10 layers (90% with data augmentation)

Statistical Benefits with Two Layers (B., 2021)

- Consider invariant $f^*(x) = \sum_{u,v \in \Omega} g^*(x_u, x_v)$
- Assume $\mathbb{E}_x[k(x_u, x_{u'})k(x_v, x_{v'})] \leq \epsilon$ if $u \neq u'$ or $v \neq v'$
- Compare different pooling layers ($h_1, h_2 \in \{\text{global}, \delta\}$) and patch sizes ($s_2$):

Statistical Benefits with Two Layers (B., 2021)

- Consider invariant $f^*(x) = \sum_{u,v \in \Omega} g^*(x_u, x_v)$
- Assume $\mathbb{E}_x[k(x_u, x_{u'})k(x_v, x_{v'})] \leq \epsilon$ if $u \neq u'$ or $v \neq v'$
- Compare different pooling layers ($h_1, h_2 \in \{\text{global}, \delta\}$) and patch sizes ($s_2$):

Excess risk bounds when $g^* \in \mathcal{H}_k \otimes \mathcal{H}_k$

h_1	h_2	s_2	$R(\hat{f}_n) - R(f^*)$ (for $\epsilon \to 0$)
δ	δ	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert^{2.5} / \sqrt{n}$
δ	global	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert^{2} / \sqrt{n}$
global	global	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert / \sqrt{n}$
global	global or δ	1	$\lVert g^* \rVert / \sqrt{n}$

Statistical Benefits with Two Layers (B., 2021)

- Consider invariant $f^*(x) = \sum_{u,v \in \Omega} g^*(x_u, x_v)$
- Assume $\mathbb{E}_x[k(x_u, x_{u'})k(x_v, x_{v'})] \leq \epsilon$ if $u \neq u'$ or $v \neq v'$
- Compare different pooling layers ($h_1, h_2 \in \{\text{global}, \delta\}$) and patch sizes ($s_2$):

Excess risk bounds when $g^* \in \mathcal{H}_k \otimes \mathcal{H}_k$

h_1	h_2	s_2	$R(\hat{f}_n) - R(f^*)$ (for $\epsilon \to 0$)
δ	δ	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert^{2.5}/\sqrt{n}$
δ	global	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert^{2}/\sqrt{n}$
global	global	$\lvert\Omega\rvert$	$\lVert g^* \rVert \lvert\Omega\rvert/\sqrt{n}$
global	global or δ	1	$\lVert g^* \rVert/\sqrt{n}$

Polynomial gains in $\lvert\Omega\rvert$ when using the right architecture!

Conclusion and perspectives

Summary

- Improved sample complexity for invariance and stability through pooling
- Locality breaks the curse of dimensionality
- Depth and pooling in convolutional models captures rich interaction models with invariances

Future directions

- Empirical benefits for kernels beyond two-layers?
- Invariance groups need to be built-in, can we adapt to them?
- Adaptivity to structures in multi-layer models:
  - Low-dimensional structures (Gabor) at first layer?
  - More structured interactions at second layer?
  - Can optimization achieve these?

Conclusion and perspectives

Summary

- Improved sample complexity for invariance and stability through pooling
- Locality breaks the curse of dimensionality
- Depth and pooling in convolutional models captures rich interaction models with invariances

Future directions

- Empirical benefits for kernels beyond two-layers?
- Invariance groups need to be built-in, can we adapt to them?
- Adaptivity to structures in multi-layer models:
  - Low-dimensional structures (Gabor) at first layer?
  - More structured interactions at second layer?
  - Can optimization achieve these?

Thank you!

References I

A. B. Approximation and learning with deep convolutional models: a kernel perspective. *arXiv preprint arXiv:2102.10032*, 2021.

A. B. and J. Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research (JMLR)*, 20(25):1–49, 2019.

A. B., L. Venturi, and J. Bruna. On the sample complexity of learning with geometric stability. *arXiv preprint arXiv:2106.07148*, 2021.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research (JMLR)*, 18(19):1–53, 2017.

J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(8):1872–1886, 2013.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. Favero, F. Cagnetta, and M. Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. *arXiv preprint arXiv:2106.08619*, 2021.

References II

B. Haasdonk and H. Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine learning*, 68(1):35–61, 2007.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10): 1331–1398, 2012.

S. Mei, T. Misiakiewicz, and A. Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory (COLT)*, 2021.

T. Misiakiewicz and S. Mei. Learning with convolution and pooling operations in kernel methods. *arXiv preprint arXiv:2111.08308*, 2021.

Y. Mroueh, S. Voinea, and T. A. Poggio. Learning with group invariant features: A kernel perspective. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

R. M. Neal. *Bayesian learning for neural networks*. Springer, 1996.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

References III

N. C. Saldanha and C. Tomei. The accumulated distribution of quadratic forms on the sphere. *Linear algebra and its applications*, 245:335–351, 1996.

V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, L. Schmidt, J. Ragan-Kelley, and B. Recht. Neural kernels without tangents. *arXiv preprint arXiv:2003.02237*, 2020.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.