On the Sample Complexity of Learning under Invariance and Geometric Stability

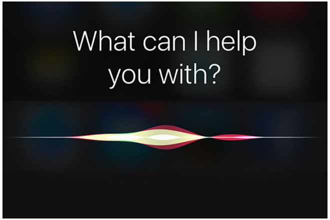
Alberto Bietti

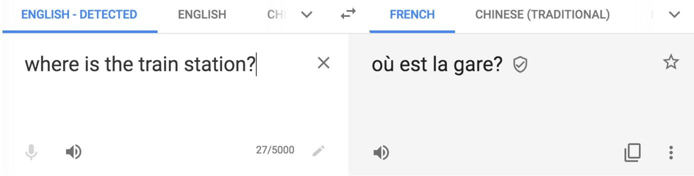
NYU

Flatiron Institute. Sept. 21, 2021.

Success of deep learning

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





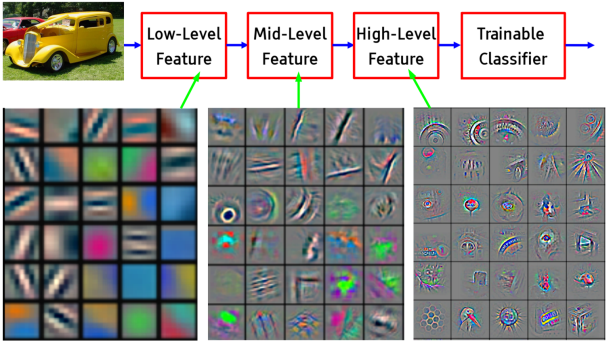
Success of deep learning

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

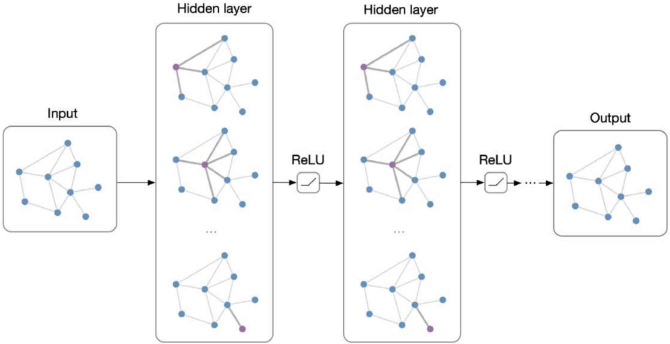


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Modern architectures (CNNs, GNNs, Transformers, ...)

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

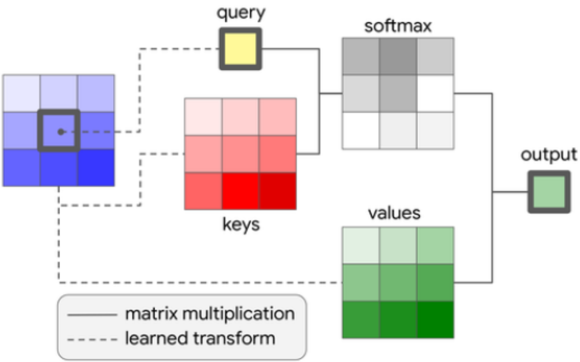
Exploiting data structure through architectures



Modern architectures (CNNs, GNNs, Transformers, ...)

- 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, Transformers, ...)

- 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)

Understanding deep learning

The challenge of deep learning theory

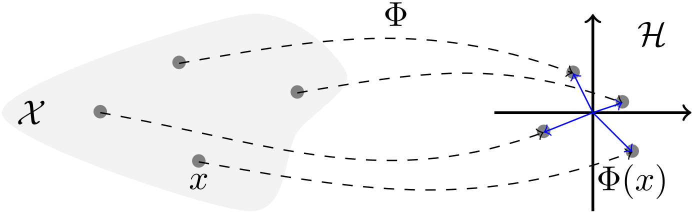
- **Over-parameterized** (millions of parameters)
- **Expressive** (can approximate any function)
- Complex **architectures** for exploiting problem structure
- Yet, **easy to optimize** with (stochastic) gradient descent!

A functional space viewpoint

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

What is an appropriate functional space?

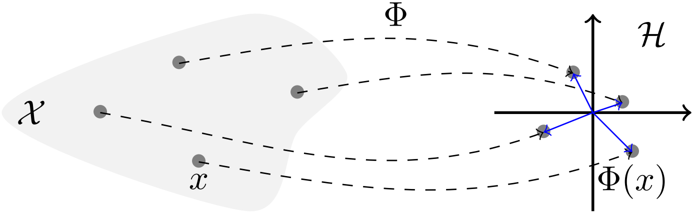
Kernels to the rescue



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 $K = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$

Kernels to the rescue



Clean and well-developed theory

- Tractable methods (convex optimization)
- Statistical and approximation properties well understood for many kernels
- Costly (kernel matrix of size N^2) but approximations are possible

Kernels for neural network architectures

Infinite-width networks (Neal, 1996; Rahimi and Recht, 2007; Jacot et al., 2018)

- *e.g.*, one-layer network: $f(x) = \frac{1}{\sqrt{m}} \sum_{i=1}^m v_i \rho(w_i^\top x)$
- Random Feature kernel: $w_i \sim \mathcal{N}(0, I)$, v_i trained

$$K_\rho(x, x') = \mathbb{E}_w[\rho(w^\top x)\rho(w^\top x')] = \kappa_\rho(x^\top x') \quad \text{when } x, x' \in \mathbb{S}^{d-1}$$

- Neural Tangent kernel: “lazy training” of both layers near random initialization

Kernels for neural network architectures

Hierarchical kernels (Cho and Saul, 2009)

- Kernels can be constructed **hierarchically**

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle \text{ with } \Phi(x) = \varphi_2(\varphi_1(x))$$

- *e.g.*, dot-product kernels on the sphere

$$K(x, x') = \kappa_2(\langle \varphi_1(x), \varphi_1(x') \rangle) = \kappa_2(\kappa_1(x^\top x'))$$

- For κ_ρ, corresponds to infinite-width limit of deep fully-connected net

Kernels for neural network architectures

Hierarchical kernels (Cho and Saul, 2009)

- Kernels can be constructed **hierarchically**

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle \text{ with } \Phi(x) = \varphi_2(\varphi_1(x))$$

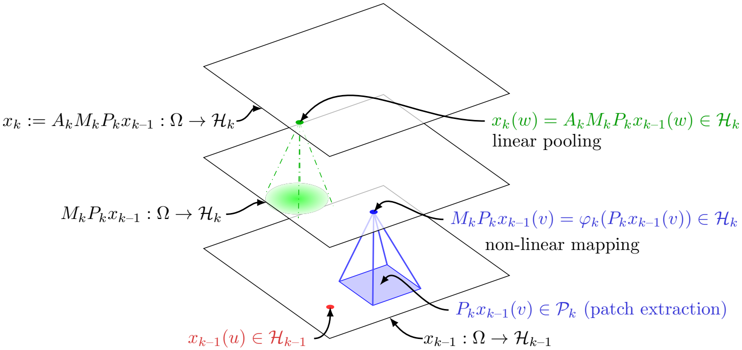
- *e.g.*, dot-product kernels on the sphere

$$K(x, x') = \kappa_2(\langle \varphi_1(x), \varphi_1(x') \rangle) = \kappa_2(\kappa_1(x^\top x'))$$

- For κ_ρ, corresponds to infinite-width limit of deep fully-connected net
- But: deep = shallow (same RKHS), limited picture (B. and Bach, 2021; Chen and Xu, 2021):
- **Can more structure help?**

Kernels for neural network architectures

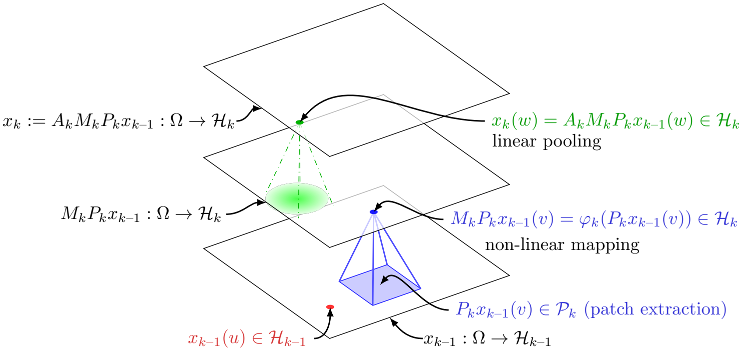
Convolutional kernels for images (Mairal et al., 2014; Mairal, 2016; Shankar et al., 2020)



- Good empirical performance with tractable approximations (Nyström)

Kernels for neural network architectures

Convolutional kernels for images (Mairal et al., 2014; Mairal, 2016; Shankar et al., 2020)



- Good empirical performance with tractable approximations (Nyström)

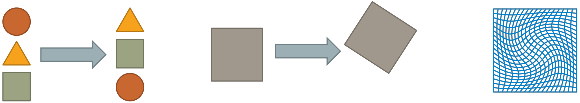
Our goal: study sample complexity benefits of architectures through kernels

Outline

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

2. Locality and depth (B., 2021)

Geometric priors

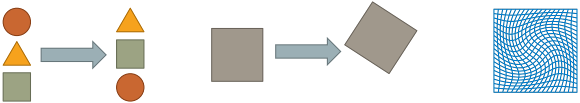


Functions $f : \mathcal{X} \to \mathbb{R}$ that are "smooth" along known transformations of input x

- *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

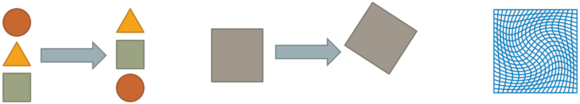
- *e.g.*, translations, rotations, permutations, deformations
- We consider: **permutations** $\sigma \in G$

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

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

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

Geometric priors: symmetrization/pooling operator

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

Geometric priors: symmetrization/pooling operator

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

Assumptions on a target function f^*

- G-invariant: $S_G f^* = f^*$
- G-stable: $f^* = S_G g^*$, for some g^* (more generally, $f^* = S_G^r g^*$)

Geometric priors: symmetrization/pooling operator

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

Assumptions on a target function f^*

- G-invariant: $S_G f^* = f^*$
- G-stable: $f^* = S_G g^*$, for some g^* (more generally, $f^* = S_G^r g^*$)

Dot-product kernels with pooling (Haasdonk and Burkhardt, 2007; Mroueh et al., 2015)

$$K(x, x') = \kappa(\langle x, x' \rangle), \qquad K_G(x, x') = \frac{1}{|G|} \sum_{\sigma \in G} \kappa(\langle \sigma \cdot x, x' \rangle)$$

- If $\kappa = \kappa_\rho$, corresponds to pooling $f(x) = \frac{1}{|G|} \sum_{\sigma \in G} \frac{1}{\sqrt{m}} \sum_{i=1}^m v_i \rho(\langle w_i, \sigma \cdot x \rangle)$

Geometric priors: symmetrization/pooling operator

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

Assumptions on a target function f^*

- G-invariant: $S_G f^* = f^*$
- G-stable: $f^* = S_G g^*$, for some g^* (more generally, $f^* = S_G^r g^*$)

Dot-product kernels with pooling (Haasdonk and Burkhardt, 2007; Mroueh et al., 2015)

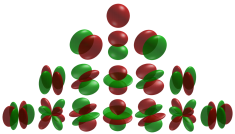
$$K(x, x') = \kappa(\langle x, x' \rangle), \qquad K_G(x, x') = \frac{1}{|G|} \sum_{\sigma \in G} \kappa(\langle \sigma \cdot x, x' \rangle)$$

- If $\kappa = \kappa_\rho$, corresponds to pooling $f(x) = \frac{1}{|G|} \sum_{\sigma \in G} \frac{1}{\sqrt{m}} \sum_{i=1}^m v_i \rho(\langle w_i, \sigma \cdot x \rangle)$

How do these interact with generic smoothness properties of f^*?

Spherical harmonics, dot-product kernels

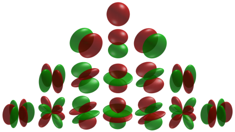
Harmonic analysis on the sphere



- 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 $V_{d,k}$

Spherical harmonics, dot-product kernels

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**: $T_K f(x) = \int \kappa(\langle x, y \rangle) f(y) d\tau(y)$
- $\mu_k = \frac{\omega_{d-2}}{\omega_{d-1}} \int_{-1}^{1} \kappa(t) P_{d,k}(t) (1-t^2)^{\frac{d-3}{2}} dt$: eigenvalues of T_K, each 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 $\overline{N}(d,k)$)
- 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) $x \sim \tau$, $\mathbb{E}[y|x] = f^*(x)$, $\text{Var}(y|x) \leq \sigma^2$
- (G-invariance) f^* is G-invariant

Sample complexity of invariant kernel: assumptions

Kernel Ridge Regression

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

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
- (*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

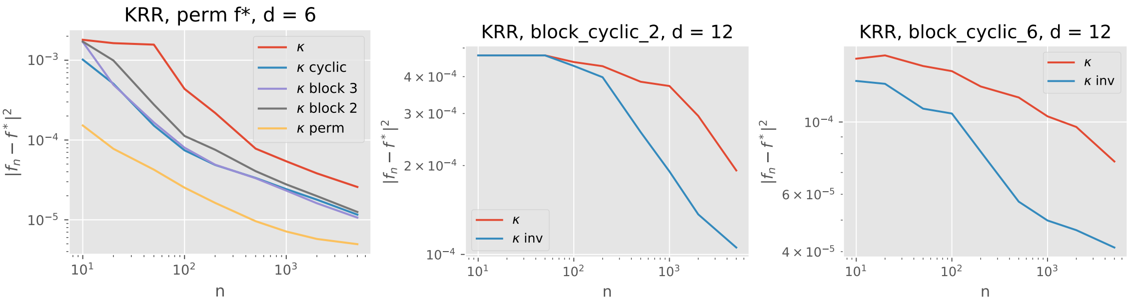
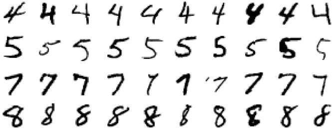


Figure: Comparison of KRR with invariant and non-invariant kernels.

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 $\ell_n := \sup\{\ell : D(\ell) \lesssim \nu_d(\ell)^{\frac{2r}{2\alpha r+1}} n^{\frac{1}{2\alpha r+1}}\}$.

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

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!)

Other limitations

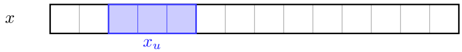
- Gains are asymptotic, constants in $O(\cdot)$ may be large
- Requires knowledge of the group for the invariant kernel
- For large groups, the pooling operation is costly
  - More structure may help, *e.g.*, stability through depth (B. and Mairal, 2019; Bruna and Mallat, 2013; Mallat, 2012)

Outline

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

2. Locality and depth (B., 2021)

Breaking the curse of dimensionality with locality

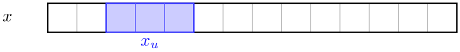


One-layer local convolutional kernel: localized patches $x_u = (x[u], \dots, x[u+s])$ (1D)

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

- RKHS $\mathcal{H}_K$ contains functions $f(x) = \sum_{u \in \Omega} g_u(x_u)$ with $g_u \in \mathcal{H}_k$
- **No curse**: smoothness requirement on g_u scales with s instead of d

Breaking the curse of dimensionality with locality



One-layer local convolutional kernel: localized patches $x_u = (x[u], \ldots, x[u+s])$ (1D)

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

- RKHS $\mathcal{H}_K$ contains functions $f(x) = \sum_{u \in \Omega} g_u(x_u)$ with $g_u \in \mathcal{H}_k$
- **No curse**: smoothness requirement on g_u scales with s instead of d
- **Pooling**: same functions, RKHS norm encourages similarities between the g_u

Breaking the curse of dimensionality with locality

Simple generalization bound

- Slow rate with Rademacher complexity and 1-Lipschitz loss, $f^* \in \mathcal{H}_K$

$$\mathbb{E}\, L(\hat{f}_n) - L(f^*) \lesssim \|f^*\|_{\mathcal{H}_K} \sqrt{\frac{\mathbb{E}_x\, K(x,x)}{n}}$$

- For invariant targets $f^* = \sum_{u \in \Omega} g^*(x_u)$: $\|f^*\|_{\mathcal{H}_K}$ independent of pooling
- If $\mathbb{E}_x\, k(x_u, x_v) \ll 1$ for $u \neq v$:
  - No pooling: $\mathbb{E}_x\, K(x,x) = |\Omega|$
  - Global pooling: $\mathbb{E}_x\, K(x,x) \approx 1 \implies$ **gain by factor $|\Omega|$**

Breaking the curse of dimensionality with locality

Simple generalization bound

- Slow rate with Rademacher complexity and 1-Lipschitz loss, $f^* \in \mathcal{H}_K$

$$\mathbb{E}\, L(\hat{f}_n) - L(f^*) \lesssim \|f^*\|_{\mathcal{H}_K} \sqrt{\frac{\mathbb{E}_x\, K(x,x)}{n}}$$

- For invariant targets $f^* = \sum_{u \in \Omega} g^*(x_u)$: $\|f^*\|_{\mathcal{H}_K}$ independent of pooling
- If $\mathbb{E}_x\, k(x_u, x_v) \ll 1$ for $u \neq v$:
  - No pooling: $\mathbb{E}_x\, K(x,x) = |\Omega|$
  - Global pooling: $\mathbb{E}_x\, K(x,x) \approx 1 \implies$ **gain by factor $|\Omega|$**
  - General pooling filter $\|h\|_1 = 1$: $\mathbb{E}_x\, K(x,x) \approx \|h\|_2^2 |\Omega|$

Breaking the curse of dimensionality with locality

Simple generalization bound

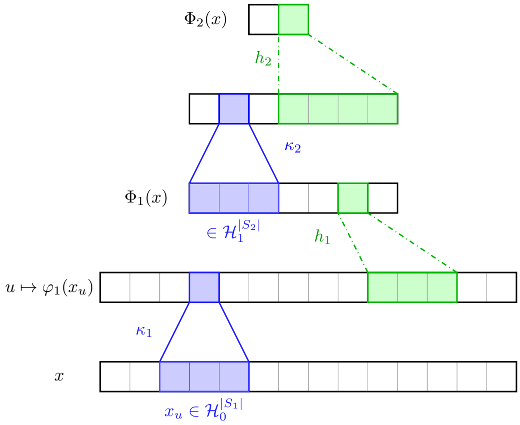
- Slow rate with Rademacher complexity and 1-Lipschitz loss, $f^* \in \mathcal{H}_K$

$$\mathbb{E}\, L(\hat{f}_n) - L(f^*) \lesssim \|f^*\|_{\mathcal{H}_K} \sqrt{\frac{\mathbb{E}_x\, K(x,x)}{n}}$$

- For invariant targets $f^* = \sum_{u \in \Omega} g^*(x_u)$: $\|f^*\|_{\mathcal{H}_K}$ independent of pooling
- If $\mathbb{E}_x\, k(x_u, x_v) \ll 1$ for $u \neq v$:
  - No pooling: $\mathbb{E}_x\, K(x,x) = |\Omega|$
  - Global pooling: $\mathbb{E}_x\, K(x,x) \approx 1 \implies$ **gain by factor $|\Omega|$**
  - General pooling filter $\|h\|_1 = 1$: $\mathbb{E}_x\, K(x,x) \approx \|h\|_2^2 |\Omega|$
- Fast rates possible (Favero et al., 2021)

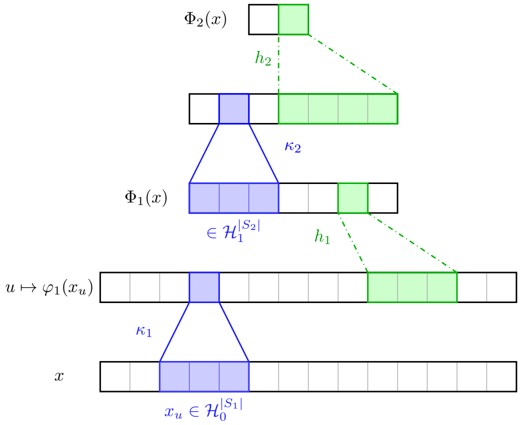
Multi-layer convolutional kernels

Convolutional Kernel Networks (Mairal, 2016) $K_2(x, x') = \langle \Phi_2(x), \Phi_2(x') \rangle$



Multi-layer convolutional kernels

Convolutional Kernel Networks (Mairal, 2016) $K_2(x, x') = \langle \Phi_2(x), \Phi_2(x') \rangle$

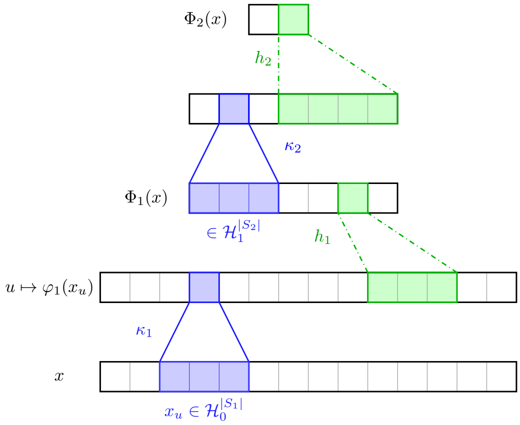


- Consider $\kappa_2(u) = u^2$
- Associated feature map (for $|S_2| = 2$):

$$\varphi_2 \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} z_1 \otimes z_1 & z_1 \otimes z_2 \\ z_2 \otimes z_1 & z_2 \otimes z_2 \end{pmatrix} \in (\mathcal{H}_1 \otimes \mathcal{H}_1)^{|S_2|^2}$$

Multi-layer convolutional kernels

Convolutional Kernel Networks (Mairal, 2016) $K_2(x, x') = \langle \Phi_2(x), \Phi_2(x') \rangle$



- Consider $\kappa_2(u) = u^2$
- Associated feature map (for $|S_2| = 2$):

$$\varphi_2 \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} z_1 \otimes z_1 & z_1 \otimes z_2 \\ z_2 \otimes z_1 & z_2 \otimes z_2 \end{pmatrix} \in (\mathcal{H}_1 \otimes \mathcal{H}_1)^{|S_2|^2}$$

- Captures **interactions** between different patches (Wahba, 1990)
- Pooling h_1: extends range of interactions
- Pooling h_2: builds invariance

Some experiments on Cifar10

2-layers, 3x3 patches, pooling/downsampling sizes (2,5). Patch kernels κ_1, κ_2.

κ_1	κ_2	Test acc. (10k examples)
Exp	Exp	80.5%
Exp	Poly3	80.5%
Exp	Poly2	79.4%
Poly2	Exp	77.4%
Poly2	Poly2	75.1%
Exp	- (Lin)	74.2%

Best performance on full Cifar10 dataset: **88.3%**, with 2-layer architecture, larger patches at 2nd layer. Comparable to (Shankar et al., 2020), which uses more layers.

Structured interaction models via depth and pooling

RKHS with quadratic κ_2: Contains functions

$$f(x) = \sum_{p,q \in S_2} \sum_{u,v \in \Omega} g_{u,v}^{pq}(x_u, x_v),$$

with $g_{u,v}^{pq} = 0$ if $|u - v - (p - q)| > \text{diam}(\text{supp}(h_1))$.

Structured interaction models via depth and pooling

RKHS with quadratic κ_2: Contains functions

$$f(x) = \sum_{p,q \in S_2} \sum_{u,v \in \Omega} g_{u,v}^{pq}(x_u, x_v),$$

with $g_{u,v}^{pq} = 0$ if $|u - v - (p - q)| > \text{diam}(\text{supp}(h_1))$.

- Additive and interaction model with $g_{u,v}^{pq} \in \mathcal{H}_k \otimes \mathcal{H}_k$ (still no curse if $s \ll d$)
- Pooling layers encourage similarities between different $g_{u,v}^{pq}$

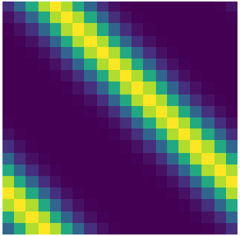
Structured interaction models via depth and pooling

RKHS with quadratic κ_2: Contains functions

$$f(x) = \sum_{p,q \in S_2} \sum_{u,v \in \Omega} g_{u,v}^{pq}(x_u, x_v),$$

with $g_{u,v}^{pq} = 0$ if $|u - v - (p - q)| > \mathrm{diam}(\mathrm{supp}(h_1))$.

- Additive and interaction model with $g_{u,v}^{pq} \in \mathcal{H}_k \otimes \mathcal{H}_k$ (still no curse if $s \ll d$)
- Pooling layers encourage similarities between different $g_{u,v}^{pq}$
  - h_1 captures "2D" invariance
  - h_2 captures invariance along diagonals



Improvements in generalization

$$\mathbb{E}\, L(\hat{f}_n) - L(f^*) \lesssim \|f^*\|_{\mathcal{H}_K} \sqrt{\frac{\mathbb{E}_x\, K(x,x)}{n}}$$

- Consider $f^*(x) = \sum_{u,v \in \Omega} g^*(x_u, x_v)$ with $g^* \in \mathcal{H}_k \otimes \mathcal{H}_k$
- Assume $\mathbb{E}_x[k(x_u, x_{u'})k(x_v, x_{v'})] \leq \epsilon$ if $u \neq u'$ or $v \neq v'$

Improvements in generalization

$$\mathbb{E}\, L(\hat{f}_n) - L(f^*) \lesssim \|f^*\|_{\mathcal{H}_K} \sqrt{\frac{\mathbb{E}_x\, K(x,x)}{n}}$$

- Consider $f^*(x) = \sum_{u,v\in\Omega} g^*(x_u, x_v)$ with $g^* \in \mathcal{H}_k \otimes \mathcal{H}_k$
- Assume $\mathbb{E}_x[k(x_u, x_{u'})k(x_v, x_{v'})] \leq \epsilon$ if $u \neq u'$ or $v \neq v'$
- Obtained bound for different pooling layers (h_1, h_2) and patch sizes $(|S_2|)$:

h_1	h_2	$\|S_2\|$	$\\|f^*\\|_K$	$\mathbb{E}_x\, K(x,x)$	Bound $(\epsilon = 0)$
δ	δ	$\|\Omega\|$	$\|\Omega\|\\|g\\|$	$\|\Omega\|^3 + \epsilon\|\Omega\|^3$	$\\|g\\|\|\Omega\|^{2.5}/\sqrt{n}$
δ	$\mathbf{1}$	$\|\Omega\|$	$\|\Omega\|\\|g\\|$	$\|\Omega\|^2 + \epsilon\|\Omega\|^3$	$\\|g\\|\|\Omega\|^2/\sqrt{n}$
$\mathbf{1}$	$\mathbf{1}$	$\|\Omega\|$	$\sqrt{\|\Omega\|}\\|g\\|$	$\|\Omega\| + \epsilon\|\Omega\|^3$	$\\|g\\|\|\Omega\|/\sqrt{n}$
$\mathbf{1}$	δ or $\mathbf{1}$	1	$\sqrt{\|\Omega\|}\\|g\\|$	$\|\Omega\|^{-1} + \epsilon\|\Omega\|$	$\\|g\\|/\sqrt{n}$

Note: larger polynomial improvements in $|\Omega|$ possible with higher-order interactions.

Conclusion and perspectives

Summary

- Improved sample complexity for invariance and stability through pooling
- Locality breaks the curse
- 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
- 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 F. Bach. Deep equals shallow for relu networks in kernel regimes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 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.

L. Chen and S. Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

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

References II

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

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.

J. Mairal. End-to-End Kernel Learning with Supervised Convolutional Kernel Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

J. Mairal, P. Koniusz, Z. Harchaoui, and C. Schmid. Convolutional kernel networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

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.

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.

References III

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

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.