# Harmonic Decompositions of Convolutional Networks

M. Scetbon[1] Z. Harchaoui[2]

[1]CREST, ENSAE

[2]Department of Statistics, University of Washington

**Q1 :** How can a Convolutional Network achieve impressive prediction performance with high dimensional data ?

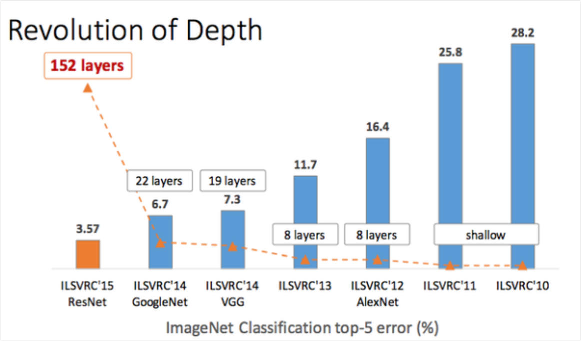
**Q2 :** What is the effect of depth on the statistical performance of a Convolutional Network ?

1. Image taken from https ://medium.com/@Lidinwise/the-revolution-of-depth-facf174924f5

**Q1 :** How can a Convolutional Network achieve impressive prediction performance with high dimensional data ?

**Q2 :** What is the effect of depth on the statistical performance of a Convolutional Network ?

**ImageNet** : dimension $\simeq 1e6$ [1]




1. Image taken from https ://medium.com/@Lidinwise/the-revolution-of-depth-facf174924f5

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

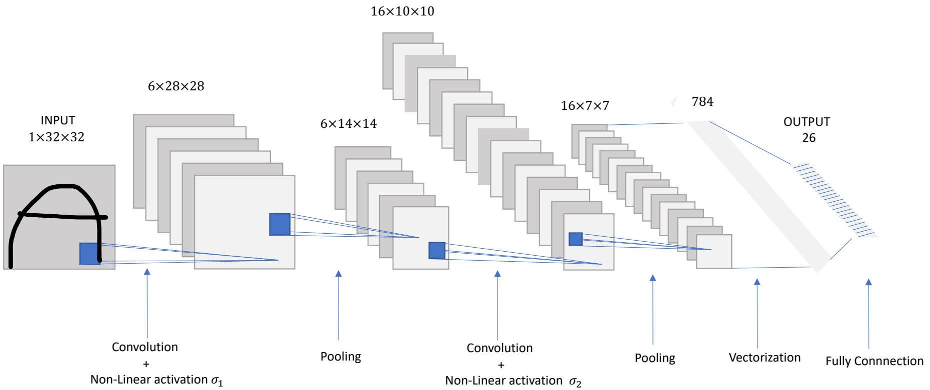
- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

- **Reproducing Kernel Hilbert Space (RKHS) associated to a Convolutional Network (CNN)**
- **Spectral Analysis of a CNN**
  - Functional ANOVA Decomposition
  - Control of the Eigenvalue Decay
- **Statistical Performance of the Regularized Least Squares (RLS)**
  - What is the dimension really captured by the network ?
  - How do the convergence rates scale with respect to the number of layers ?

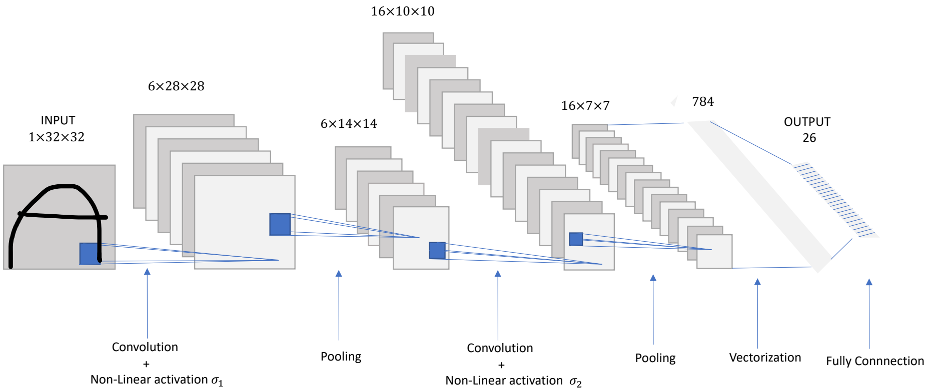
# RKHS associated to a CNN

# Convolutional Network




- $\mathcal{F}_N$ : function space generated by a CNN with a fixed number of layers $N$ and non-linear activations $(\boldsymbol{\sigma}_i)_{i=1}^N$

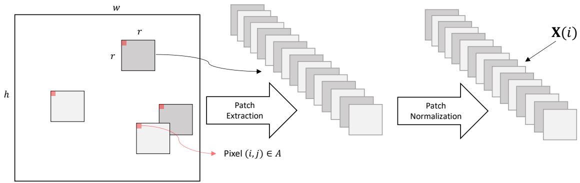
# Convolutional Network




- $\mathcal{F}_N$ : function space generated by a CNN with a fixed number of layers $N$ and non-linear activations $(\boldsymbol{\sigma}_i)_{i=1}^N$

# RKHS induced by a CNN

- Image space : $\mathcal{I} := \prod_{i=1}^{n} S^{d-1} \subset \mathbb{R}^D$




$= \mathbf{X}$

- Define for all $i$, $f_i(x) = \sum_{t\geq 0} \frac{|\sigma_i^{(t)}(0)|}{t!} x^t$

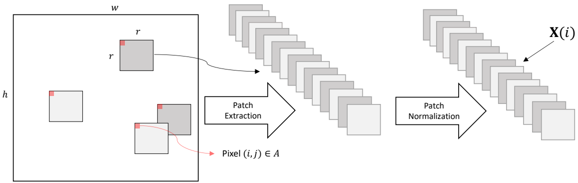
**Convolutional Kernel**

$$K_N(\mathbf{X}, \mathbf{X}') := f_N \circ \ldots \circ f_2 \left( \sum_{i=1}^{n} f_1 \left( \langle \mathbf{X}_i, \mathbf{X}'_i \rangle_{\mathbb{R}^d} \right) \right)$$

- $H_N$ : RKHS associated to convolutional kernel $K_N$

# RKHS induced by a CNN

- Image space : $\mathcal{I} := \prod_{i=1}^{n} S^{d-1} \subset \mathbb{R}^D$




$= \mathbf{X}$

- Define for all $i$, $f_i(x) = \sum_{t\geq 0} \frac{|\boldsymbol{\sigma}_i^{(t)}(0)|}{t!} x^t$

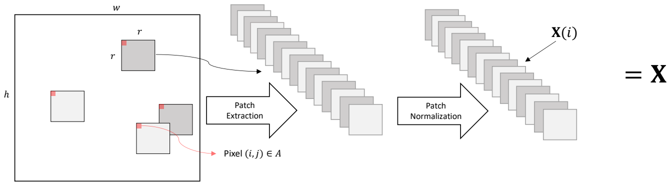
**Convolutional Kernel**

$$K_N(\mathbf{X}, \mathbf{X}') := f_N \circ \ldots \circ f_2 \left( \sum_{i=1}^{n} f_1\left(\langle \mathbf{X}_i, \mathbf{X}'_i \rangle_{\mathbb{R}^d}\right) \right)$$

- $H_N$ : RKHS associated to convolutional kernel $K_N$

## RKHS induced by a CNN

- Image space : $\mathcal{I} := \prod_{i=1}^{n} S^{d-1} \subset \mathbb{R}^D$




- Define for all $i$, $f_i(x) = \sum_{t \geq 0} \frac{|\boldsymbol{\sigma}_i^{(t)}(0)|}{t!} x^t$

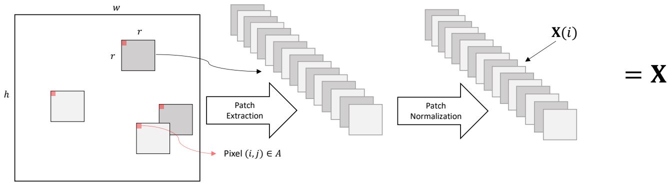
### Convolutional Kernel

$$K_N(\mathbf{X}, \mathbf{X}') := f_N \circ \ldots \circ f_2 \left( \sum_{i=1}^{n} f_1 \left( \langle \mathbf{X}_i, \mathbf{X}'_i \rangle_{\mathbb{R}^d} \right) \right)$$

- $H_N$ : RKHS associated to convolutional kernel $K_N$

## RKHS induced by a CNN

- Image space : $\mathcal{I} := \prod_{i=1}^{n} S^{d-1} \subset \mathbb{R}^D$




- Define for all $i$, $f_i(x) = \sum_{t \geq 0} \frac{|\boldsymbol{\sigma}_i^{(t)}(0)|}{t!} x^t$

**Convolutional Kernel**

$$K_N(\mathbf{X}, \mathbf{X}') := f_N \circ \ldots \circ f_2 \left( \sum_{i=1}^{n} f_1 \left( \langle \mathbf{X}_i, \mathbf{X}'_i \rangle_{\mathbb{R}^d} \right) \right)$$

- $H_N$ : RKHS associated to convolutional kernel $K_N$

## Hilbertian envelope of $\mathcal{F}_N$

$\mathcal{F}_N \subset H_N$ : Any function generated by a CNN is an element of the RKHS $H_N$

- **Network width** : $H_N$ does not depend on the number of filters considered at each hidden layer.
- **Kernel universality** : $H_N$ is dense in $\mathcal{C}(\mathcal{I})$ w.r.t. the uniform norm $\|\cdot\|_\infty$.

As a result we have :

$$\inf_{f \in H_N} R(f) := \mathbb{E}[(f(\mathbf{X}) - Y)^2] = R^*$$

where $R^*$ is the Bayes risk.

## Hilbertian envelope of $\mathcal{F}_N$

$\mathcal{F}_N \subset H_N$ : Any function generated by a CNN is an element of the RKHS $H_N$

- **Network width** : $H_N$ does not depend on the number of filters considered at each hidden layer.

- **Kernel universality** : $H_N$ is dense in $\mathcal{C}(\mathcal{I})$ w.r.t. the uniform norm $\|\cdot\|_\infty$.

  As a result we have :

$$\inf_{f \in H_N} R(f) := \mathbb{E}[(f(\mathbf{X}) - Y)^2] = R^*$$

  where $R^*$ is the Bayes risk.

**Hilbertian envelope of $\mathcal{F}_N$**

$\mathcal{F}_N \subset H_N$ : Any function generated by a CNN is an element of the RKHS $H_N$

- **Network width** : $H_N$ does not depend on the number of filters considered at each hidden layer.
- **Kernel universality** : $H_N$ is dense in $\mathcal{C}(\mathcal{I})$ w.r.t. the uniform norm $\|\cdot\|_\infty$.

As a result we have :

$$\inf_{f \in H_N} R(f) := \mathbb{E}[(f(\mathbf{X}) - Y)^2] = R^*$$

where $R^*$ is the Bayes risk.

### Hilbertian envelope of $\mathcal{F}_N$

$\mathcal{F}_N \subset H_N$ : Any function generated by a CNN is an element of the RKHS $H_N$

- **Network width** : $H_N$ does not depend on the number of filters considered at each hidden layer.
- **Kernel universality** : $H_N$ is dense in $\mathcal{C}(\mathcal{I})$ w.r.t. the uniform norm $\|\cdot\|_\infty$.

  As a result we have :

$$\inf_{f \in H_N} R(f) := \mathbb{E}[(f(\mathbf{X}) - Y)^2] = R^*$$

where $R^*$ is the Bayes risk.

# Spectral Analysis of CNNs

# Functional ANOVA Decomposition

**Tensor-Product Space ANOVA model.** A $n$-dimensional function $f$ can be decomposed as

$$f(\mathbf{X}_1, ..., \mathbf{X}_n) = C + \sum_{i=1}^{n} f_i(\mathbf{X}_i) + \sum_{i<j}^{n} f_{i,j}(\mathbf{X}_i, \mathbf{X}_j) + \dots$$

- $C$ : a constant.
- $d^*$ : the highest order of interactions allowed by the model.
- $f_i \in H$ where $H$ is an RKHS : ***main effect***
- $\forall A \subset \{1, ..., n\}$ with $|A| \leq d^*$, $f_A \in H^{\otimes |A|}$

**Mercer Decomposition**

$$K_N(\mathbf{X}, \mathbf{X}') = \sum_{\substack{k_1,\ldots,k_n \geq 0 \\ 1 \leq l_{k_i} \leq \alpha_{k_i,d}}} \mu_{(k_i, l_{k_i})_{i=1}^n} e_{(k_i, l_{k_i})_{i=1}^n}(\mathbf{X}) e_{(k_i, l_{k_i})_{i=1}^n}(\mathbf{X}')$$

- $e_{(k_i, l_{k_i})_{i=1}^n}(\mathbf{X}) := \prod_{i=1}^{n} Y_{k_i}^{l_{k_i}}(\mathbf{X}_i)$
- $\mathbf{X}_i \in S^{d-1}$ : a patch
- $(Y_m^l)$ : Orthonormal basis of spherical harmonics

# Functional ANOVA Decomposition

## Proposition

*Let* $N \geq 2$*,* $f_1$ *a real value function that admits a Taylor decomposition around 0* $f_N \circ .... \circ f_2$ *a polynomial of degree* $a \geq 1$*. Then by denoting* $d^* := \min(a, n)$*,*

$$|\{i : k_i \neq 0\}| > d^* \implies \mu_{(k_i, l_{k_i})_{i=1}^n} = 0$$

.

- $\mu_{(k_i, l_{k_i})_{i=1}^n}$ vanish as soon as the interactions captured by the eigenfunctions associated is too large relatively to the *depth of the network.*
- $d^*$ : the highest order of interaction allowed by the network.

# Functional ANOVA Decomposition

**Proposition**

*Let $N \geq 2$, $f_1$ a real value function that admits a Taylor decomposition around 0 $f_N \circ .... \circ f_2$ a polynomial of degree $a \geq 1$. Then by denoting $d^* := \min(a, n)$,*

$$|\{i : k_i \neq 0\}| > d^* \implies \mu_{(k_i, l_{k_i})_{i=1}^n} = 0$$

.

- $\mu_{(k_i, l_{k_i})_{i=1}^n}$ vanish as soon as the interactions captured by the eigenfunctions associated is too large relatively to the *depth of the network*.
- $d^*$ : the highest order of interaction allowed by the network.

# Functional ANOVA Decomposition

**Proposition**

*Let $N \geq 2$, $f_1$ a real value function that admits a Taylor decomposition around 0 $f_N \circ .... \circ f_2$ a polynomial of degree $a \geq 1$. Then by denoting $d^* := \min(a, n)$,*

$$|\{i : k_i \neq 0\}| > d^* \implies \mu_{(k_i, l_{k_i})_{i=1}^n} = 0$$

.

- $\mu_{(k_i, l_{k_i})_{i=1}^n}$ vanish as soon as the interactions captured by the eigenfunctions associated is too large relatively to the *depth of the network*.
- $d^*$ : the highest order of interaction allowed by the network.

**$\mathcal{F}_N$ is highly structured**

A CNN is a constructive way to build a functional ANOVA model where :

- the ***main effects*** live in a Hilbert space completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$
- the highest order of interaction $d^*$ is controlled by the depth of the network.

**$\mathcal{F}_N$ is highly structured**

A CNN is a constructive way to build a functional ANOVA model where :

- the ***main effects*** live in a Hilbert space completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$
- the highest order of interaction $d^*$ is controlled by the depth of the network.

# Functional ANOVA Decomposition

**$\mathcal{F}_N$ is highly structured**

A CNN is a constructive way to build a functional ANOVA model where :

- the ***main effects*** live in a Hilbert space completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$
- the highest order of interaction $d^*$ is controlled by the depth of the network.

## Mercer Decomposition

$$K_N(\mathbf{X},\mathbf{X}') = \sum_{\substack{k_1,\dots,k_n \geq 0 \\ 1 \leq l_{k_i} \leq \alpha_{k_i,d}}} \mu_{(k_i,l_{k_i})_{i=1}^n} e_{(k_i,l_{k_i})_{i=1}^n}(\mathbf{X}) e_{(k_i,l_{k_i})_{i=1}^n}(\mathbf{X}')$$

where

$$\mu_{(k_i,l_{k_i})_{i=1}^n} := \sum_{q \geq 0} a_q \sum_{\substack{\alpha_1,\dots,\alpha_n \geq 0 \\ \sum_{i=1}^n \alpha_i = q}} \binom{q}{\alpha_1,\dots,\alpha_n} \prod_{i=1}^n \lambda_{k_i,\alpha_i}$$

and

$$\lambda_{k,\alpha} = \frac{|S^{d-2}|\Gamma((d-1)/2)}{2^{k+1}} \sum_{s \geq 0} \left[ \frac{d^{2s+k}}{dt^{2s+k}}\Big|_{t=0} \frac{f_1^{\alpha}(t)}{(2s+k)!} \right] \frac{(2s+k)!}{(2s)!} \frac{\Gamma(s+1/2)}{\Gamma(s+k+d/2)}$$

# Control of the Eigenvalue Decay

- $(b_m)_{m\geq 0}$ : Coefficient in the Taylor series of $f_1$.

  Assume that there exists $c_1, c_2, r > 0$ such that for all $m \geq 0$

  $$c_2 r^m \leq b_m \leq c_1 r^m.$$

- $f_N \circ .... \circ f_2$ : polynomial of degree $a \geq 1$.
- $d^* = \min(a, n)$ : highest order of interaction.

**Proposition**

*There exists $C_1, C_2 > 0$ and $0 < \gamma < q$ constants such that for all $m \geq 0$ :*

$$C_2 e^{-qm^{\frac{1}{(d-1)d^*}}} \leq \mu_m \leq C_1 e^{-\gamma m^{\frac{1}{(d-1)d^*}}}$$

# Statistical Performance of RLS

# What is the dimension really captured by the network ?

- $d$ : dimension of the extracted patches in $S^{d-1}$
- $d^*$ : highest order of interaction allowed by the network
- $f^*(x) = \mathbb{E}(Y|X = x)$ : the conditional mean
- $f_{H_N,\lambda}$ : the solution of

$$\min_{f \in H_N} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_N}^2 \right\} .$$

**Learning Rates**

For a well chosen $\lambda_\ell$ we obtain with high probability that :

$$R(f_{H_N,\lambda_\ell}) - R(f^*) \lesssim \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

# What is the dimension really captured by the network ?

- $d$ : dimension of the extracted patches in $S^{d-1}$
- $d^*$ : highest order of interaction allowed by the network
- $f^*(x) = \mathbb{E}(Y|X = x)$ : the conditional mean
- $f_{H_N,\lambda}$ : the solution of

$$\min_{f \in H_N} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_N}^2 \right\} .$$

**Learning Rates**

For a well chosen $\lambda_\ell$ we obtain with high probability that :

$$R(f_{H_N,\lambda_\ell}) - R(f^*) \lesssim \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

# What is the dimension really captured by the network ?

- $d$ : dimension of the extracted patches in $S^{d-1}$
- $d^*$ : highest order of interaction allowed by the network
- $f^*(x) = \mathbb{E}(Y|X = x)$ : the conditional mean
- $f_{H_N,\lambda}$ : the solution of

$$\min_{f \in H_N} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_N}^2 \right\} .$$

**Learning Rates**

For a well chosen $\lambda_\ell$ we obtain with high probability that :

$$R(f_{H_N,\lambda_\ell}) - R(f^*) \lesssim \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

# What is the dimension really captured by the network ?

- $d$ : dimension of the extracted patches in $S^{d-1}$
- $d^*$ : highest order of interaction allowed by the network
- $f^*(x) = \mathbb{E}(Y|X = x)$ : the conditional mean
- $f_{H_N,\lambda}$ : the solution of

$$\min_{f \in H_N} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_N}^2 \right\} .$$

**Learning Rates**

For a well chosen $\lambda_\ell$ we obtain with high probability that :

$$R(f_{H_N,\lambda_\ell}) - R(f^*) \lesssim \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

# What is the dimension really captured by the network ?

- $d$ : dimension of the extracted patches in $S^{d-1}$
- $d^*$ : highest order of interaction allowed by the network
- $f^*(x) = \mathbb{E}(Y|X = x)$ : the conditional mean
- $f_{H_N,\lambda}$ : the solution of

$$\min_{f \in H_N} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} (f(x_i) - y_i)^2 + \lambda \|f\|_{H_N}^2 \right\} .$$

**Learning Rates**

For a well chosen $\lambda_\ell$ we obtain with high probability that :

$$R(f_{H_N,\lambda_\ell}) - R(f^*) \lesssim \frac{\log(\ell)^{(d-1)d^*}}{\ell}$$

## What is the dimension really captured by the network ?

### The effect of the dimension

- The learning rates are minimax optimal
- The learning rates are free-dimension with respect to the number of parameters (or filters).
- The dimension captured by the network : $(d-1) \times d^* \ll D$

# What is the dimension really captured by the network ?

## The effect of the dimension

- The learning rates are minimax optimal
- The learning rates are free-dimension with respect to the number of parameters (or filters).
- The dimension captured by the network : $(d-1) \times d^* \ll D$

# What is the dimension really captured by the network ?

### The effect of the dimension

- The learning rates are minimax optimal
- The learning rates are free-dimension with respect to the number of parameters (or filters).
- The dimension captured by the network : $(d-1) \times d^* \ll D$

# How do the rates scale with respect to the number of layers ?

The learning rates obtained exhibit two regimes of interest :

- **Regime 1 :** if $d^* \ll n$, the optimal rates obtained are close the optimal rates for estimating multivariate functions in $d$ dimensions where $d$ is the patch size. Therefore the rates obtained are almost *dimension free*.
- **Regime 2 :** As soon as $f_N \circ .... \circ f_2$ is a polynomial function with degree higher than $n$, then adding layers to the network will not change change the rates. Thus there is a regime in which adding layers does not affect the rates and allows the function space of target functions to grow.

## Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\sigma_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

# Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\sigma_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free.*
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

# Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

# Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

## Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

# Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

# Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free*.
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

## Conclusion

- Designed a *universal* kernel $K_N$ such that its RKHS $H_N$ contains $\mathcal{F}_N$.
- Obtained a Mercer decomposition of the kernel $K_N$.
  - the functional ANOVA structure of $F_N$ where $d^*$ and the main effects are completely determined by $(\boldsymbol{\sigma}_i)_{i=1}^N$.
  - control of the eigenvalue decay in several decay regimes.
- Showed the learning rates of RLS on hypothesis space $H_N$
  - Convergence rates are minimax optimal from a nonparametric learning viewpoint.
  - **Regime 1 :** The rates are almost *dimension free.*
  - **Regime 2 :** The rates remain unchanged, while approximation power is increased, as we added more layers.

**Thank you**