# Quantification d'incertitudes

## Part IV.1

Approximation

# Introduction

The goal of approximation is to replace a function $u$ in some space $X$ by a simpler function (easy to estimate and to operate with).

An approximation are searched in a set of functions $X_n$ described by $n$ parameters (or $O(n^\alpha)$ parameters), sometimes called a model class or hypothesis set.

A sequence of subsets $(X_n)_{n\geq 1}$ is called an approximation tool.

# Best approximation error

For a certain subset of functions $X_n$, the error of best approximation of $u$ by elements of $X_n$ is defined by

$$e_n(u) = \inf_{v \in X_n} d(u, v)$$

where $d$ is a distance measuring the quality of an approximation, typically

$$d(u, v) = \|u - v\|_X.$$

# Fundamental problems in approximation

- Determine if $e_n(u)$ converges to 0 for a certain class of functions and a certain approximation tool,
- Determine how fast $e_n(u)$ converges to 0 for a certain class of functions and a certain approximation tool, e.g.
$$e_n(u) \leq M\gamma(n)^{-1},$$
- For a given approximation tool, determine the class of functions for which a certain convergence rate will be ensured, e.g.
$$\mathcal{A}^\gamma = \{u : \sup_{n\geq 1} \gamma(n) e_n(u) < +\infty\}$$
where $\gamma$ is a strictly increasing function, or determine the complexity $n = n(\epsilon, u) \geq \gamma^{-1}(\epsilon/M)$ for having $e_n(u) \leq \epsilon$,
- Provide algorithms which produce approximations $u_n \in X_n$ such that
$$d(u, u_n) \leq Ce_n(u)$$
with $C$ independent of $n$ or $C(n)e_n(u) \to 0$ as $n \to \infty$

# Outline

# Approximation tools

We distinguish linear approximation, that is approximation in linear spaces $X_n$, from nonlinear approximation, where $X_n$ are nonlinear spaces.

When $X_n$ is nonlinear (sometimes non smooth), approximation problems

$$\min_{v \in X_n} d(u, v)$$

become nonlinear (possibly non smooth) optimization problems.

# Linear approximation tools

- Algebraic polynomials:
$$X_n = \mathbb{P}_n([a,b]) = span\{1, x, \dots, x^n\}$$
- Trigonometric polynomials
$$X_n = \mathbb{T}_n = span\{1, cos(x), sin(x), \dots, cos(nx), sin(nx)\}$$
- Splines
$$X_n = \mathcal{S}_p^r(T_n) = \{v \in C^r(\Omega) : v_{|K} \in \mathbb{P}_p(K), K \in T_n\},$$
where $T_n$ is a partition of $\Omega$ into $\#T_n = n$ elements. Includes standard finite elements for $r = 0$.
- Kernel-based functions
$$X_n = \{v(x) = \sum_{i=1}^n a_i K(x, x_i) : a_i \in \mathbb{R}\}$$
for given points $\{x_i\}_{i=1}^n$ and given kernel $K(x,y)$ (e.g. $K(x,y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$).
- For $(\varphi_i)_{i=1}^n$ a given basis of functions:
$$X_n = span\{\varphi_1, \dots, \varphi_n\}$$

# Nonlinear approximation tools

- $n$-term approximation on a given basis (or frame) $\{\varphi_i\}_{i\geq 1}$ (e.g. polynomials, wavelets...)

$$X_n = \{v = \sum_{i\in\Lambda} a_i\varphi_i : a_i \in \mathbb{R}, \#\Lambda = n\}$$

- Rational functions

$$X_n = \{\frac{p}{q} : p, q \in \mathbb{P}_n\}$$

- Splines over a free partition

$$X_n = \{v \in \mathcal{S}_p^r(T_n) \text{ with } T_n \text{ a partition of } \Omega \subset \mathbb{R}^d \text{ with } \#T_n = n\}$$

- Kernel-based functions with free points $\{x_i\}_{i=1}^n$ and/or free kernel parameters.
- Neural networks

$$X_n = \{v(x) = \sum_{i=1}^n a_i\sigma(w_i^T x + b_i) : a_i \in \mathbb{R}, w_i \in \mathbb{R}^d, b_i \in \mathbb{R}\}$$

# Linear versus nonlinear approximation: an illustrative example

Let $u \in X = C([0,1])$ and $X_n$ be the set of piecewise constant functions on a partition $T_n = \{I_k\}_{k=1}^n$ of $[0,1]$. Let

$$u_n(x) = \sum_{k=1}^n a_k 1_{I_k}(x) \quad \text{with} \quad a_k = \frac{1}{|I_k|}\int_{I_k} u(x)dx.$$

The local approximation error of $u$ by $u_n$ is such that

$$\|u - u_n\|_{L^\infty(I_k)} \leq \sup_{x,y \in I_k} |u(x) - u(y)| \leq \int_{I_k} |u'(x)|dx$$

- **Linear approximation case** (fixed partition). Taking a uniform partition, and assuming $u' \in L^\infty$, we have

$$\|u - u_n\|_{L^\infty} \leq \|u'\|_{L^\infty} n^{-1}.$$

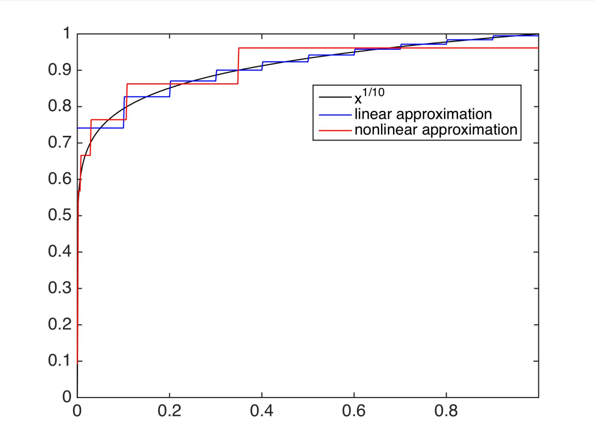
- **Nonlinear approximation case** (free partition). Assume $u' \in L^1$. Choosing a partition which equilibrates $\int_{I_k} |u'|$, we obtain

$$\|u - u_n\|_{L^\infty} \leq \|u'\|_{L^1} n^{-1}$$

# Linear versus nonlinear approximation: an illustrative example

## Illustration.

As an example, consider $u(x) = x^\alpha$ in $C([0,1])$, with $0 < \alpha < 1$. $u'(x) = \alpha x^{\alpha-1}$ is in $L^1$ but not in $L^\infty$. Nonlinear approximation converges in $O(n^{-1})$ while linear approximation converges in $O(n^{-\alpha})$.



# Outline

## The curse of dimensionality

For $X = L^p(\mathcal{X})$, $\mathcal{X} = [0,1]^d$, and $X_n = \mathbb{P}_m(T_h)$ the set of piecewise polynomials on a uniform partition, we have for $u \in W^{k,p}(\mathcal{X})$, $k \leq m+1$,

$$e_n(u)_{L^p} \lesssim n^{-k/d}$$

We observe

- the curse of dimensionality : deterioration of the rate of approximation when $d$ increases. Exponential growth with $d$ of the complexity for reaching a given accuracy.
- the blessing of smoothness : improvement of the rate of approximation when $k$ increases.

We may ask if the curse of dimensionality is due to the particular choice of approximation tool (piecewise polynomials) for approximating functions in $W^{k,p}$ ?

# Optimal linear approximation methods: linear widths

For a compact subset $K$ of $X$, we define the Kolmogorov $n$-width of $K$:

$$d_n(K)_X = \inf_{dim(X_n)=n} \sup_{u \in K} \inf_{v \in X_n} \|u - v\|_X$$

where the infimum is taken over all linear subspaces $X_n$ of dimension $n$.

$d_n(K)_X$ measures how well the set $K$ can be approximated by a $n$-dimensional space. It measures the ideal performance that we can expect from linear approximation methods.

Upper bound for $d_n(K)_X$ can be obtained by a specific approximation method. Lower bound for $d_n(K)$ comes from the diversity in $K$.

For $X = L^p(\mathcal{X})$, $\mathcal{X} = [0,1]^d$, and $K$ the unit ball of $W^{k,p}(\mathcal{X})$, we have

$$cn^{-k/d} \leq d_n(K)_X \leq Cn^{-k/d}.$$

Exponential growth of the complexity to reach a given accuracy. So the curse of dimensionality can not be avoided by a suitable choice of linear approximation spaces.

# Can extra smoothness help ?

For $X = L^\infty(\mathcal{X})$ with $\mathcal{X} = [0,1]^d$ and

$$K = \{v \in C^\infty(\mathcal{X}) : \sup_\alpha \|D^\alpha u\|_{L^\infty} < \infty\},$$

we have

$$\min\{n : d_n(K)_X \leq 1/2\} \geq c2^{d/2}.$$

Extra smoothness does not help !

## Can nonlinear methods help ?

For evaluating the ideal performance of the approximation of elements of a subset $K$ in $X$ by nonlinear methods, different notions of widths have been proposed.

The following definition of a nonlinear width is relevant for many numerical algorithms:

$$\delta_n(K)_X = \inf_{E,D} \sup_{u \in K} \|u - D(E(u))\|_X$$

where the infimum is taken over all continuous functions $E$ from $K$ to $\mathbb{R}^n$ and all continuous functions $D$ from $\mathbb{R}^n$ to $K$. $\delta_n(K)_X$ quantifies how well the set $K$ can be approximated by $n$-dimensional nonlinear manifolds having continuous parametrizations.

For $X = L^p(\mathcal{X})$, $\mathcal{X} = [0,1]^d$, and $K$ the unit ball of $W^{k,p}(\mathcal{X})$, we have

$$cn^{-k/d} \leq \delta_n(K)_X \leq Cn^{-k/d}.$$

Again, we observe an exponential growth of the complexity to reach a given accuracy. So the curse of dimensionality can not be avoided by nonlinear methods.

# How to beat the curse of dimensionality ?

The key is to consider classes of functions with specific low-dimensional structures and to propose approximation formats (models) which exploit these structures (application-dependent).

Approximations are searched in subsets $X_n$ with a number of parameters

$$n = O(d^p).$$

# Outline

# Classical tools for high-dimensional approximation

- Linear models

$$a_1 x_1 + \ldots + a_d x_d$$

- Polynomial models

$$\sum_{\alpha \in \Lambda} a_\alpha x^\alpha$$

where $\Lambda \subset \mathbb{N}^d$ is a set of multi-indices, either fixed (linear approximation) or free (nonlinear approximation).

- More general expansions

$$\sum_{i=1}^{n} a_i \psi_i(x)$$

where the $\psi_i$ are either fixed (linear approximation) or freely selected in a dictionary of functions (nonlinear approximation).

# Classical tools for high-dimensional approximation

- Additive models

$$u_1(x_1) + \ldots + u_d(x_d)$$

or more generally

$$\sum_{\alpha \subset T} u_\alpha(x_\alpha)$$

where $T \subset 2^{\{1,\ldots,d\}}$ is either fixed (linear approximation) or a free parameter (nonlinear approximation).

- Multiplicative models

$$u_1(x_1) \ldots u_d(x_d)$$

or more generally

$$\prod_{\alpha \in T} u_\alpha(x_\alpha)$$

where $T \subset 2^{\{1,\ldots,d\}}$ is either a fixed or a free parameter.

# Composition of functions

$$f(g(x))$$

with $g : \mathbb{R}^d \to \mathbb{R}^m$ is a simple map to a low-dimensional space ($m \ll d$), and $f : \mathbb{R}^m \to \mathbb{R}$ has a low-dimensional parametrization.

- Linear transformations (ridge functions)

$$f(Wx), \quad W \in \mathbb{R}^{m \times d}$$

  A typical example is the perceptron

$$f(y) = a\sigma(w^T x + b)$$

- For large $m$, requires specific models for $f$, e.g.

$$f(g(x)) = f_1(g_1(x)) + \ldots + f_m(g_m(x))$$

  A sum of $m$ perceptrons is a shallow neural network (with one hidden layer of width $m$)

$$\sum_{i=1}^{m} a_i \sigma({w_i}^T x + b_i) = a^T \sigma(A_i x + b_i)$$

# More compositions... deep neural networks

Deep neural networks consist of a composition of functions

$$g_L \circ g_{L-1} \circ \ldots \circ g_2 \circ g_1(x)$$

with

$$g_\ell(x) = \sigma_\ell(A_\ell x + b_\ell), \quad A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}, \quad b_\ell \in \mathbb{R}^{n_\ell}$$

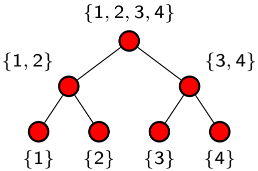
and $\sigma_\ell$ a given function (so called activation function).

# More compositions... deep neural networks

Structured sparsity can be imposed on matrices $A_\ell$, leading to specific architectures.

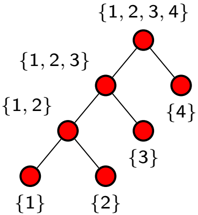
- Deep convolutional networks

$$f_{1,2,3,4}\left(f_{1,2}\left(f_1(x_1), f_2(x_2)\right), f_{3,4}\left(f_3(x_3), f_4(x_4)\right)\right)$$




- Deep recurrent networks

$$f_{1,2,3,4}\left(f_{1,2,3}\left(f_{1,2}\left(f_1(x_1), f_2(x_2)\right), f_3(x_3)\right), f_4(x_4)\right)$$



# Low-rank formats

A multivariate function $v(x_1, \dots, x_d)$ is identified with an order-$d$ tensor.

- Approximation with rank one (multiplicative model)

$$v(x) = u_1(x_1) \dots u_d(x_d)$$

- Approximation with canonical rank $r$

$$v(x) = \sum_{i=1}^{r} u_1^i(x_1) \dots u_d^i(x_d)$$

- For a subset of variables $\alpha \subset \{1, \dots, d\} := D$, $v(x)$ can be identified with a bivariate function $v(x_\alpha, x_{\alpha^c})$, where $x_\alpha$ and $x_{\alpha^c}$ are complementary groups of variables.

  The canonical rank of this bivariate function is the $\alpha$-rank of $v$, denoted $\mathrm{rank}_\alpha(v)$, and such that

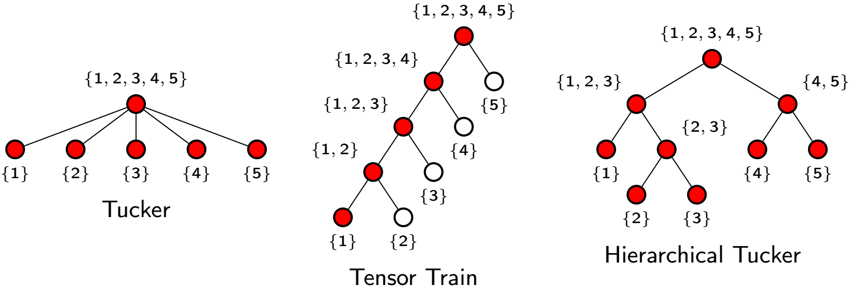
$$v(x) = \sum_{i=1}^{\mathrm{rank}_\alpha(v)} v_\alpha^i(x_\alpha) w_{\alpha^c}^i(x_{\alpha^c})$$

## Low-rank formats

- For $T \subset 2^D$ a collection of subsets of $D$, approximation in a subset

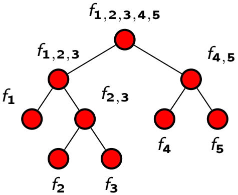
$$\mathcal{T}_r^T = \{v : \mathrm{rank}_\alpha(v) \leq r_\alpha, \alpha \in T\}.$$

- Tree-based tensor formats correspond to a tree-structured $T$



# Tree-based low-rank formats

- A tensor $v$ in $\mathcal{T}_r^T$ admits a multilinear parametrization with parameters $\{f_\alpha\}_{\alpha \in T \cup \{D\}}$ forming a tree network of low dimensional functions (tensors).




$$v(x) = f_{1,2,3,4,5}\left(f_{1,2,3}\left(f_1(x_1), f_{2,3}(f_2(x_2), f_3)(x_3)\right), f_{4,5}\left(f_4(x_4), f_5(x_5)\right)\right)$$

where for $1 \leq \nu \leq d$,

$$f_\nu : \mathcal{X}^\nu \to \mathbb{R}^{r_\nu},$$

and for any node $\alpha$ with children $\beta_1$ and $\beta_2$,

$$f_\alpha : \mathbb{R}^{r_{\beta_1}} \times \mathbb{R}^{r_{\beta_2}} \to \mathbb{R}^{r_\alpha}$$

is a bilinear function, which is identified with a tensor in $\mathbb{R}^{r_\alpha \times r_{\beta_1} \times r_{\beta_2}}$.

- Corresponds to a deep network with a particular architecture and multilinear functions.

# References

R. DeVore and G. Lorentz.
Constructive approximation.
Springer, 1993.

R. A. DeVore.
Nonlinear approximation.
*Acta Numerica*, 7:51–150, 1998.

# Lebesgue spaces

Let $\mu$ be measure on $\mathcal{X} \subset \mathbb{R}^d$.

- $L^p_\mu(\mathcal{X})$, $1 \leq p < \infty$ the set of measurable functions $u : \mathcal{X} \to \mathbb{R}$ with bounded norm

$$\|u\|_{L^p_\mu(\mathcal{X})} = \left( \int_{\mathcal{X}} |u(x)|^p d\mu(x) \right)^{1/p}$$

- $L^\infty_\mu(\mathcal{X})$, the set of measurable functions $u : \mathcal{X} \to \mathbb{R}$ with bounded norm

$$\|u\|_{L^\infty_\mu(\mathcal{X})} = \operatorname*{ess\,sup}_{x \in \mathcal{X}} |u(x)|$$

- $L^p_\mu(\mathcal{X})$, for $1 \leq p \leq \infty$, are Banach spaces.

# Sobolev spaces

For a $d$-variate function $u$, and $\alpha \in \mathbb{N}^d$, the $\alpha$-derivative of $u$ is

$$D^\alpha u(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} u(x_1, \dots, x_d)$$

- $W_\mu^{k,p}(\mathcal{X})$, the Sobolev space of functions $u$ in $L_\mu^p(\mathcal{X})$ whose derivatives $D^\alpha u$ (in the sense of distributions) are in $L_\mu^p(\mathcal{X})$, for $|\alpha| \leq k$.
  One defines on $W_\mu^{k,p}(\mathcal{X})$ a semi-norm

  $$|u|_{W_\mu^{k,p}} = \max_{|\alpha|=k} \|D^\alpha u\|_{L_\mu^p}$$

  and a norm

  $$\|u\|_{W_\mu^{k,p}} = \max_{|\alpha|\leq k} \|D^\alpha u\|_{L_\mu^p}$$

- $W_\mu^{k,p}(\mathcal{X})$, for $1 \leq p \leq \infty$, are Banach spaces.