

Approximation of Continuous Functions in Normed Spaces

1.1. Introduction

Given a real valued function f defined on $\mathbb{R}$ which is assumed to belong to a class of functions F , approximating f amounts to substitute f by some function φ which belongs to some class V , which we will assume to be included in F . The natural setting is when F is a linear space, which we assume from now. The function φ will be chosen as to be more simple than f . We will call φ the approximation of f , or the approximating function. When f differs from φ , the error

$$R(x) := (f - \varphi)(x)$$

will be considered; the relative error might also be considered, but we will mainly define φ through analysis pertaining to the function R .

In order to make this more meaningful, we have first to define the term “simple” when applied to approximation procedures, and its impact on the behavior of the function R . This requires a precise choice for the class V and for the criterion leading to the choice of φ in V for a given function f .

The choice of the approximating function φ clearly depends on properties to be expected by R or on some function of this error.

A natural setting leads one to define a distance d on F . Therefore, choosing φ follows from the solution of a problem of the kind

$$\min_{\varphi \in V} d_F(f, \varphi).$$

This problem may be difficult to solve when f does not belong to V .

A usual choice is to define the distance d_F through a norm on F , say $\|\cdot\|_F$. The resulting distance satisfies $d_F(f, g) := \|f - g\|_F$.

The advantages of such a choice are clear when considering the properties of the norm. First, the norm is continuous, when seen as a functional on the linear space F . This setting makes all linear operations on F continuous with respect to the norm, and the topology on F is determined by the neighborhoods of the null function, through translation. This setting provides a set of useful tools; for example, the unit ball in F defined by

$$\overline{S_F(\mathbf{0}, 1)} := \{f \in F : \|f\|_F \leq 1\}$$

is compact if and only if F is a finite dimensional linear space (see Appendix 1, [CEL 16] theorem A1.6), a basic property to be used extensively when proving the existence of optimal points for a linear functional. More generally, all topological properties of a linear functional derive from its behavior on $S_F(\mathbf{0}, 1)$. We suppose therefore that the function f to be approximated belongs to a normed space $(F, \|\cdot\|_F)$. We may also require a further assumption, which makes the analytic frame easier; at a certain time, we will require that the norm is defined by an inner product $\langle \cdot, \cdot \rangle_F$, defining $\|f\|_F^2 := \langle f, f \rangle_F$. In this case, F is a pre-Hilbert space.

The approximating function φ is assumed to belong to a finite dimensional subspace of F , denoted by V .

1.2. Some remarks on the meaning of the word “simple”. Choosing the approximation

The definition of simplicity is somehow vague and shares many different meanings. From the analytic point of view, simplicity is related to analytic regularity. From the standpoint of numerical calculation, the function is seen as a number of elementary operations to be performed; then complexity may be captured through the number of such operations. The function is complex when many operations are required to evaluate it. This latest point of view relates to the means at our disposal in terms of memory size, precision of the basic operations in the computer, speed of the processors, etc. Complexity results from the present state of the means at our disposal.

Choosing a function φ in a linear space V with finite dimension n may lead to defining n itself as the simplicity of the problem. It therefore seems that the simplicity of a given problem is defined by the number of parameters which are requested in order to solve it.

We say that the approximation procedure is linear when the approximating function φ belongs to a finite dimensional linear space.

The relation between the simplicity of the approximating function φ and the error R is loosely stated as follows.

Let $\varphi_1, \dots, \varphi_n$ be n functions defined on $\mathbb{R}$, and consider the absolute value of the local error substituting f by

$$\varphi = \sum_{i=1}^n \alpha_i \varphi_i \in V := \text{span} \{ \varphi_i : i = 1, \dots, n \},$$

namely,

$$|R_n(x)| = \left| f(x) - \sum_{i=1}^n \alpha_i \varphi_i(x) \right|.$$

This quantity is transformed into a global index, $d_F(f, \varphi)$, which is a function of the dimension of the linear space V where the approximation is performed. We denote this quantity by $\Psi(n)$. When

$$\Psi(n) = \mathcal{O}(g(n)) \quad [1.1]$$

for some non-negative and non-increasing function g which satisfies $\lim_{n \rightarrow \infty} g(n) = 0$, we say that the approximation procedure is of accuracy $\mathcal{O}(g(n))$.

Let $f \in \mathcal{C}^{(p)}(\mathbb{R})$, and

$$\varphi_N(x) := \sum_{|l| < N} e^{2i\pi lx} \widehat{f}(l),$$

where $\widehat{f}(l)$ are the Fourier coefficients of f

$$\widehat{f}(l) := \int_{\mathbb{R}} f(x) e^{-2i\pi lx} dx.$$

We have

$$f(x) := \sum_{l=-\infty}^{+\infty} e^{2i\pi lx} \widehat{f}(l)$$

and there exists a constant $c > 0$, such that

$$\left| \widehat{f}(l) \right| \leq \frac{c}{l^p}. \quad [1.2]$$

The convergence of the series $\sum_{l=-\infty}^{+\infty} e^{2i\pi lx} \widehat{f}(l)$ to f holds as a consequence of the following result:

THEOREM 1.1.– (Fejér) When f is some integrable function with summable Fourier coefficients, then the series of functions $\sum_{l=-\infty}^{+\infty} e^{2i\pi lx} \widehat{f}(l)$ converges pointwise to f on $\mathbb{R}$.

PROOF.– See Kolmogorov–Fomine [KOL 74] ■

We prove [1.2] and deduce [1.1] for a suitable function g and with $\Psi(N) := \sup_x |f(x) - \varphi_N(x)|$.

Evaluate

$$\int_{\mathbb{R}} f(x) e^{-2i\pi lx} dx$$

by parts, which yields

$$\widehat{f}(l) = \left(\frac{1}{2\pi il} \right) \widehat{f'}(l).$$

Iterating, we obtain

$$\widehat{f}(l) = \left(\frac{1}{2\pi il} \right)^p \widehat{f^{(p)}}(l).$$

Hence, defining

$$c := \left(\frac{1}{2\pi} \right)^p \max_{l \in \mathbb{Z}} \left(\widehat{f^{(p)}}(l) \right)$$

it holds

$$\begin{aligned} \Psi(N) &:= \sup_x |f(x) - \varphi_N(x)| = \sup_x \left| \sum_{|l| > N} \widehat{f}(l) e^{2i\pi lx} \right| \\ &\leq \sum_{|l| > N} \left| \widehat{f}(l) \right| \leq \sum_{|l| > N} \frac{c}{l^p} = \mathcal{O}(N^{-p}). \end{aligned}$$

For a smooth function, convergence of $\Psi(N)$ to 0 is fast; for example, let

$$\mathbb{R} \ni x \mapsto f(x) := \frac{e^2 - 1}{2(e^2 + 1 - 2e \cos 2\pi x)}.$$

Then,

$$\hat{f}(l) = \frac{e^{-|l|}}{2},$$

and therefore

$$\begin{aligned} |f(x) - \varphi(x)| &= \left| \sum_{|l| > N} \hat{f}(l) e^{2i\pi l x} \right| \leq \sum_{|l| > N} \left| \frac{e^{-|l|}}{2} \right| |e^{2i\pi l x}| \\ &\leq \sum_{|l| > N} \left| \frac{e^{-|l|}}{2} \right| = \frac{e^{-N}}{e - 1} =: \Psi(N). \end{aligned}$$

Clearly, analytic simplicity and algorithmic complexity are concepts that interact with each other.

Another principle that may lead our ideas when choosing an approximating procedure is the following: the graph of φ should be as close as possible to that of f . Henceforth, when f has “slow variations”, we would require the same for φ , and the same for “rapid variations” of f .

For a given function f with slow variations, its approximation should be a polynomial with a low degree. On the contrary, a function f with large variations would require an approximating polynomial with a high degree. If f has nearly vertical slopes with abrupt changes in curvature or vertices with angular behavior, then its approximation can be chosen as a rational ratio function, i.e. the ratio of two polynomials. In such cases, the advantage of using an approximating function φ in

$$\mathcal{P}_{n,m} := \left\{ \frac{P_n}{Q_m} : (n, m) \in \mathbb{N}^2 \setminus \{(0, 0)\} \right\}$$

lies in the fact that a tuning of m and n allows us to reproduce very irregular local behaviors of f . This choice therefore allows us to reduce the complexity inherent to the choice of a polynomial with high degree for the same function f , which furthermore may lead to computational difficulties and numerical instability.

For functions f with rapid increase or decrease, the approximating function should include exponential terms of the form e^x, e^{-x} . Accordingly, approximations

for periodic functions should include trigonometric terms. More generally, the choice of the function φ would make φ follow the graph of f as simply as possible, sometimes through the choice of a grid, according to the local behavior of f .

Observe that, except for the class $\mathcal{P}_{n,m}$, all other approximating classes which we considered are linear spaces with finite dimension. For example, the linear space

$$V := \left\{ \varphi_m : \varphi_m(x) = \sum_{k=1}^m a_k e^{b_k x}, a_k, b_k \in \mathbb{R} \right\}$$

has dimension m .

The linear space

$$V := \left\{ \varphi_m : \varphi_m(x) = \sum_{k=1}^m a_k \cos kx + b_k \sin kx, a_k, b_k \in \mathbb{R} \right\}$$

which is used in order to approximate periodic functions has dimension $2n + 1$.

1.2.1. Splines

Assuming that f is defined on some interval $[a, b]$ in $\mathbb{R}$, define a partition of $[a, b]$, namely, $I_1 := [a, z_1], \dots, I_{i+1} := [z_i, z_{i+1}], \dots, I_k := [z_k, b]$. In any of these interval I_j , f can be approximated by a simple function, for example, a straight line, a parabola or a cubic function. The resulting approximating function on $[a, b]$ may not be defined in the same way on all the intervals I_j . Therefore, the restrictions $\varphi|_{I_j}$ of φ on the intervals I_j , $j = 1, \dots, k$, are polynomials with degree less or equal to $m - 1$.

For a given subdivision $\{z_1, \dots, z_k\}$ of $[a, b]$, the class $S_m(z_1, \dots, z_k)$ of all functions with continuous derivatives on $[a, b]$ up to the order $m - 2$ in $[a, b]$, which satisfy that their restrictions $\varphi|_{I_j}$ are polynomial functions with degree less or equal to $m - 1$, is the class of polynomial splines. The set of points $\{z_1, \dots, z_k\}$ is the set of nodes for the spline.

When $m = 2$, the spline functions are continuous and are linear on each I_j . For $m = 4$, we use the word “cubic spline”. These are functions with first and second continuous derivatives and they coincide with polynomials with degree less or equal to 3 between the nodes.

Denote

$$x_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

PROPOSITION 1.1.— The set of functions $S_m(z_1, \dots, z_k)$ defined on $[a; b]$ is a finite dimensional linear space. A basis of $S_m(z_1, \dots, z_k)$ is given by the family of functions

$$\left\{ 1|_{[a,b]}, x|_{[a,b]}, \dots, x^{r-1}|_{[a,b]}, (x - z_1)_+^{r-1}|_{[a,b]}, \dots, (x - z_k)_+^{r-1}|_{[a,b]} \right\}$$

where $1|_{[a,b]}, \dots, (x - z_k)_+^{r-1}|_{[a,b]}$ denotes the restrictions of the various functions on $[a, b]$.

PROOF.— For clearness, we denote 1 instead of $1|_{[a,b]}, \dots, (x - z_k)_+^{r-1}$ instead of $(x - z_k)_+^{r-1}|_{[a,b]}$. We prove that the family $\left\{ 1, x, \dots, x^{r-1}, (x - z_1)_+^{r-1}, \dots, (x - z_k)_+^{r-1} \right\}$ is a linearly independent family. Let $x \in [a, b]$, and assume that

$$\sum_{j=0}^{r-1} \alpha_j x^j + \sum_{i=1}^k \beta_i (x - z_i)_+^{r-1} = 0.$$

Assume $a \leq x < z_1 \leq \dots \leq z_k$. It holds that $x - z_1 < 0, \dots, x - z_k < 0$ and therefore $(x - z_i)_+ = 0$ for $i = 1, \dots, k$. Hence, for $x \in [a, z_1)$, it holds $\sum_{j=0}^{r-1} \alpha_j x^j = 0$. Since $\{x^j, j = 0, \dots, r-1\}$ is a family of independent functions, $\alpha_0 = \dots = \alpha_{r-1} = 0$. Now, let x such that $z_1 \leq x < z_2$. Then, $x - z_2 < 0, \dots, x - z_k < 0$, and therefore $\sum_{i=1}^k \beta_i (x - z_i)_+^{r-1} = 0$ becomes

$$\beta_1 (x - z_1)_+^{r-1} = 0.$$

Hence, $\beta_1 = 0$. For $z_2 \leq x < z_3$ it holds $\sum_{i=2}^k \beta_i (x - z_i)_+^{r-1} = \beta_2 (x - z_2)_+^{r-1} = 0$. Hence, $\beta_2 = 0$. In the same way looking at the various cases $x \in [z_i, z_{i+1})$, $i = 3, 4, \dots, k-1$, we get $\beta_2 = \dots = \beta_k = 0$. Therefore, the system $\left\{ 1, x, \dots, x^{r-1}, (x - z_1)_+^{r-1}, \dots, (x - z_k)_+^{r-1} \right\}$ is linearly independent. It remains to prove that it generates all the splines. Let $s \in S_m(z_1, \dots, z_k)$ be some spline. Rename the nodes $z_1, \dots, z_k$, as follows: $a = z_0, \dots, z_{k+1} = b$. Consider the polynomials $P_i \in \mathcal{P}_{r-1}([a, b])$, $i = 1, \dots, k+1$, such that $s(x) = P_i(x)$ in $[z_{i-1}; z_i]$ for $i = 1, \dots, k+1$. The conditions of differentiability yield

$$P_{i+1}^{(j)}(z_i) = P_i^{(j)}(z_i), j = 1, \dots, r-2.$$

Therefore, there exist coefficients γ_i such that

$$P_{i+1}(x) - P_i(x) = \gamma_i (x - z_i)^{r-1}.$$

Define

$$P_1(x) = \sum_{i=0}^{r-1} \alpha_i x^i$$

with $s(x) = P_1(x)$ for $x \in [a, z_1]$. In the interval $[z_1, z_2]$, we have

$$s(x) = P_1(x) + (P_2 - P_1)(x) = \sum_{i=0}^{r-1} \alpha_i x^i + \gamma_1 (x - z_1)^{r-1}.$$

Therefore, in $[a, z_1] \cup [z_1, z_2] = [a, z_2]$, we may write

$$s(x) = \sum_{i=0}^{r-1} \alpha_i x^i + \gamma_1 (x - z_1)_+^{r-1}.$$

Proceeding this way, we prove that

$$S_m(z_1, \dots, z_k) = \text{span} \left\{ \begin{matrix} 1|_{[a,b]}, x|_{[a,b]}, \dots, x^{r-1}|_{[a,b]} \\ (x - z_1)_+^{r-1}|_{[a,b]}, \dots, (x - z_k)_+^{r-1}|_{[a,b]} \end{matrix} \right\}.$$

Therefore,

$$\dim S_m(z_1, \dots, z_k) = r + k. \quad \blacksquare$$

This proof can be found in [DRO 11].

1.3. The choice of the norm in order to specify the error

An important family of norms of functions defined on some interval $[a, b]$ is defined by

$$L_p(\omega, f - \varphi) := \left(\int_a^b |f(x) - \varphi(x)|^p \omega(x) dx \right)^{1/p}, \text{ for } p \in (0, +\infty].$$

The function

$$\omega : [a, b] \rightarrow (0, \infty)$$

is called the weight function; it satisfies

$$\int_a^b \omega(x) dx = 1.$$

In the case when $x \mapsto \omega(x) = 1$, for all x in $[a, b]$, the above norm is denoted by $L_p(f - \varphi)$ or L_p instead of $L_p(1, f - \varphi)$.

As known

$$L_\infty(f - \varphi) = \sup_{x \in [a, b]} |f(x) - \varphi(x)|.$$

Norms of current use for applications are L_2, L_∞ . The norm L_2 is called the least square norm, or the Hilbertian norm. The norm L_∞ is the uniform norm, or Chebyshev's norm, or sup norm; some authors name it the Lagrangian norm or minimax norm.

The symbols $\|\cdot\|_2$ and $\|\cdot\|_\infty$ will be used to denote L_2 and L_∞ , respectively.

Although no general rule exists for the choice of a specific norm, some common sense arguments prevail. First, the norm should be such that the problem of approximating some function f has at least one solution. If such a solution can be made explicit, this would be a clear advantage. Such is the case for the L_2 norm, which furthermore induces uniqueness of the approximation.

The choice of the norm obviously depends upon the quality of the approximation which we intend to reach and also on the characteristics of the function f . For example, assume that we have only some information on the L_2 norm of the function f ; clearly, we will not be able to produce any information on the local error $R(x)$ at some point x . For this sake, we should consider the norm L_∞ . The resulting information is that in any point x , the error $|R(x)|$ is less or equal $L_\infty(f - \varphi)$.

The following remarks and examples may illustrate the relation which should link the choice of the norm and the properties of the function f .

EXAMPLE 1.1.— Assume that a polynomial $\varphi \in \mathcal{P}_n$ with given degree n is chosen for the approximation of the exponential function e^x on $\mathbb{R}$. For fixed x , decompose e^x as the sum of its integer part and its mantissa in base 2

$$\lg_2 e^x = [\lg_2 e^x] + m, \text{ where } [\lg_2 e^x] \in \mathbb{N}, \text{ and } m \in (0; 1).$$

Therefore, it holds

$$e^x = 2^{[\lg_2 e^x]} 2^m.$$

The integer power of 2, namely, $2^{[\lg_2 e^x]}$, does not require any approximation, since this is known without error. The only term to be approximated is 2^m . Since $m \in (0; 1)$, it holds $2^m \in (1; 2)$. It follows that it is enough to have at hand a good

approximation of the mantissa of the function e^x when it takes its values in $(1; 2)$, in order to obtain a good approximation of the function e^x on whole $\mathbb{R}$. In other words, accurately evaluating e^x on $\mathbb{R}$ requires only an accurate approximation of x on $(\ln 1; \ln 2) = (0; 0,693)$. Clearly, the norm which guarantees an upper bound ε for the error at any point x in $(0; 0,693)$ is the uniform norm on $[0; 0,693]$. It follows that the approximating problem can be stated as

$$\varphi_n^* = \arg \min_{\varphi \in \mathcal{P}_n} \max_{x \in [0; 0,693]} |e^x - \varphi_n(x)|.$$

EXAMPLE 1.2.– Assume now that we intend to approximate the discontinuous function $f(x) := \text{sign}(x)$, for $x \in [-5; 5]$, using a function φ in $\mathcal{C}^{(0)}([-5; 5])$. With the uniform norm, there exists an infinity of solutions for this problem

$$\varphi^* = \arg \min_{\varphi \in \mathcal{C}^{(0)}([-5; 5])} \max_{x \in [-5; 5]} \left| \frac{x}{|x|} - \varphi(x) \right|.$$

Indeed, it holds

$$\min_{\varphi \in \mathcal{C}^{(0)}([-5; 5])} \max_{x \in [-5; 5]} \left| \frac{x}{|x|} - \varphi(x) \right| \geq 1$$

as seen now. We consider different cases:

– *First case:* using the null function $\varphi(x) = 0$ as approximation, we have

$$\max_x |\text{sign}(x) - \varphi(x)| = \max_x |\text{sign}(x) - 0| = 1.$$

– *Second case:* if we use a strictly positive function $\varphi(x) > 0$ for all x , then since for any $x < 0$, $\text{sign}(x) = -1$, it holds

$$\begin{aligned} \max_{x \in [-5; 0)} |\text{sign}(x) - \varphi(x)| &= \max_{x \in [-5; 0)} |-1 - \varphi(x)| \\ &= \max_{x \in [-5; 0)} |1 + \varphi(x)| = \max_{x \in [-5; 0)} (1 + \varphi(x)) \\ &= 1 + \max_{x \in [-5; 0)} \varphi(x) \geq 1. \end{aligned}$$

– *Third case:* similar to case 2, if we use a strictly negative function.

– *Fourth case:* for a function φ with changing sign, we must consider various subcases.

a) Assume that $\varphi(0) = 0$. Then,

$$\left| \lim_{x \rightarrow 0} (\text{sign}(x) - \varphi(x)) \right| = |1 - 0| = 1.$$

Hence,

$$\sup_{x \in [-5; 5]} |\text{sign}(x) - \varphi(x)| = 1.$$

b) Assume that $\varphi(c) = 0$ for some $c < 0$. Then, there exists a neighborhood on the right of c , say $\mathcal{I}_{c+}$, on which $x < 0$ and $\varphi(x) > 0$ hold. Hence,

$$\begin{aligned} \sup_{x \in \mathcal{I}_{c+}} |\text{sign}(x) - \varphi(x)| &= \sup_{x \in \mathcal{I}_{c+}} |-1 - \varphi(x)| \\ &= \sup_{x \in \mathcal{I}_{c+}} |(-1)(1 + \varphi(x))| = 1 + \sup_{x \in \mathcal{I}_{c+}} \varphi(x) \geq 1 \end{aligned}$$

c) Assume that $\varphi(c) = 0$ for some $c > 0$. Then, there exists a neighborhood on the left of c , $\mathcal{I}_{c-}$, on which $x > 0$ and $\varphi(x) < 0$ hold. Hence,

$$\begin{aligned} \sup_{x \in \mathcal{I}_{c-}} |\text{sign}(x) - \varphi(x)| &= \sup_{x \in \mathcal{I}_{c-}} |1 - \varphi(x)| \\ &= \sup_{x \in \mathcal{I}_{c-}} |1 + (-\varphi(x))| \\ &= \sup_{x \in \mathcal{I}_{c-}} (1 + |\varphi(x)|) \\ &= 1 + \sup_{x \in \mathcal{I}_{c-}} |\varphi(x)| \geq 1 \end{aligned}$$

In all those cases, we have

$$\max_{x \in [-5; 5]} |\text{sign}(x) - \varphi(x)| \geq 1.$$

Note that the (continuous) null function $\mathbf{0} : [-5; 5] \rightarrow \mathbb{R}, x \mapsto \mathbf{0}(x) = 0$, has the same norm $L_\infty(\mathbf{0})$ as has φ_ε , defined by

$$\varphi_\varepsilon(x) := \begin{cases} -1 & \text{for } x \in [-5; -\varepsilon], \varepsilon > 0 \\ \frac{x}{\varepsilon} & \text{for } x \in [-\varepsilon; \varepsilon] \\ 1 & \text{for } x \in [\varepsilon; 5] \end{cases}.$$

It is, however, clear that φ_ε is a better approximation of $f(x) = \text{sign}(x)$ than $\mathbf{0}$ is. This latest fact can be made explicit through the L_2 distance between φ_ε and the null function $\mathbf{0}$. Indeed, in this case, it holds

$$L_2(\text{sign}(\cdot) - \varphi_\varepsilon) < L_2(\text{sign}(\cdot) - \mathbf{0}) = L_2(\text{sign}(\cdot)).$$

Therefore, φ_ε improves on $\mathbf{0}$ for the L_2 distance when approximating the sign function. However, for the sup norm, the null function is as bad as an approximation

of the sign function as φ_ε , for example. The choice of the norm should therefore be defined with respect to the properties of the function to be approximated. The sup norm appears to be a bad choice for discontinuous functions. It results that the Hilbertian norm is a better choice than the uniform norm in this case.

1.4. Optimality with respect to a norm

We now consider what may be expected in terms of uniqueness and existence for an optimal approximation for a given norm $\|\cdot\|$.

1.4.1. Existence of an optimal solution

By Weierstrass' Theorem, continuity of a function and compactness of its domain are a set of sufficient conditions in order to assess minimal and maximal points for this function on its domain.

The norm is a continuous mapping defined on F and therefore also on V . As a result, the existence of a best approximating function φ in V may hold when seen as an optimization problem on a compact subset of V .

Since $V \subset F$ is a linear space with finite dimension n the closed ball

$$\overline{S_F(f, \|f\|_F)} := \{\varphi \in V : \|f - \varphi\|_F \leq \|f\|_F\}$$

with center f and radius $\|f\|_F$ is a compact set. Indeed $S_F(f, \|f\|_F)$ is clearly a bounded set in V .

Therefore the problem

$$\min_{\varphi \in V \cap \overline{S_F(f, \|f\|_F)}} \|f - \varphi\|_F$$

has indeed at least one solution in $V \cap \overline{S_F(f, \|f\|_F)}$.

We now consider as potential approximating functions of f , functions φ in V ; we prove that the solutions of

$$\min_{\varphi \in V \subset F} \|f - \varphi\|_F$$

coincide with those of

$$\min_{\varphi \in V \cap \overline{S_F(f, \|f\|_F)}} \|f - \varphi\|_F.$$

We then conclude that the problem

$$\min_{\varphi \in V \cap S_F(f, \|f\|_F)} \|f - \varphi\|_F$$

has a solution which will at least solve

$$\min_{\varphi \in V \subset F} \|f - \varphi\|_F.$$

THEOREM 1.2.— It holds

$$\varphi^* = \arg \min_{\varphi \in V} \|f - \varphi\|_F = \arg \min_{\varphi \in V \cap S_F(f, \|f\|_F)} \|f - \varphi\|_F.$$

PROOF.— Assume $f \neq \varphi$ (otherwise f is the solution of the problem). It holds $\mathbf{0} \in \overline{S_F(f, \|f\|_F)}$, since

$$\|f - \mathbf{0}\|_F = \|f\|_F \leq \|f\|_F.$$

Furthermore $\mathbf{0} \in V$, since V is a linear space. It follows that $\mathbf{0} \in I := V \cap \overline{S_F(f, \|f\|_F)} \neq \emptyset$.

For any $h \in V \setminus \overline{S_F(f, \|f\|_F)}$, it holds

$$\|f - h\|_F > \|f\|_F = \|f - \mathbf{0}\|_F. \quad [1.3]$$

Indeed if

$$\|f - h\|_F \leq \|f\|_F,$$

then, by the very definition of $S_F(f, \|f\|_F)$ we should have $h \in \overline{S_F(f, \|f\|_F)}$.

It follows that $\mathbf{0}$ approximates f in a better way than any $h \notin \overline{S_F(f, \|f\|_F)}$ does, by [1.3].

This implies that the optimal approximation of f should belong to $\overline{S_F(f, \|f\|_F)}$. ■

1.4.2. Uniqueness of the optimal solution

We have seen that the solutions to the problem

$$\min_{\varphi \in V \subset F} \|f - \varphi\|_F$$

always exist when V is a finite dimensional linear space.

The problem of uniqueness of such solutions is more complex.

We first state a definition.

DEFINITION 1.1.– Let $(F, \|\cdot\|_F)$ be a normed linear space. Let

$$Fr(S_F(f, r)) := \{g \in F : \|f - g\|_F = r\}$$

be the frontier of the ball $S_F(f, r)$, $r > 0$. The norm $\|\cdot\|_F$ is called a strict norm if and only if for any couple of functions $(g_1, g_2) \in (Fr(S_F(f, r)))^2$, (i.e. such that $\|f - g_1\|_F = \|f - g_2\|_F$) any function $g := \alpha g_1 + (1 - \alpha)g_2$, $\alpha \in (0; 1)$ on the open segment defined by g_1 and g_2 satisfies $\|f - g\|_F < r$. In other words, the open segment belongs to the open ball.

DEFINITION 1.2.– A normed linear space is strict whenever its norm is strict.

The following theorem holds.

THEOREM 1.3.– Let $(F, \|\cdot\|_F)$ be a strict normed linear space and let V be a linear subspace of F with finite dimension n . Given a function f in F , the best approximation $\varphi \in V$ with respect to the norm $\|\cdot\|_F$, is unique.

PROOF.– We prove this result by contradiction, assuming that there exist two distinct solutions φ_1 and φ_2 . Define $\varphi_3 := (\varphi_1 + \varphi_2)/2$. We may assume that $f \in F \setminus V$. Let

$$d^* := \inf_{\varphi \in V} \|f - \varphi\|_F = \|f - \varphi_1\|_F = \|f - \varphi_2\|_F.$$

Then, $\|f - \varphi_1\|_F = \|f - \varphi_2\|_F = d^*$ implicates $(\varphi_1, \varphi_2) \in (Fr(S_F(f, d^*)))^2$. Since any norm is a convex function, it holds

$$\begin{aligned} \|f - \varphi_3\|_F &= \left\| \frac{f}{2} + \frac{f}{2} - \frac{\varphi_1}{2} - \frac{\varphi_2}{2} \right\|_F \\ &= \left\| \frac{f - \varphi_1}{2} + \frac{f - \varphi_2}{2} \right\|_F \\ &\leq \frac{1}{2} \|f - \varphi_1\|_F + \frac{1}{2} \|f - \varphi_2\|_F = d^*. \end{aligned}$$

Now $\|\cdot\|_F$ is strict. Hence, since φ_3 is on the segment from φ_1 to φ_2 and since both φ_1 and φ_2 belong to $(Fr(S_F(f, d^*)))^2$, the inequality above is strict. This implies

$$\|f - \varphi_3\|_F < d^*.$$

Henceforth, neither φ_1 nor φ_2 can be solution of the problem, which concludes the proof. ■

The above theorem has various consequences.

The norms L_∞ and L_1 are not strict norms. Therefore, uniqueness for problems of the kind

$$\min_{\varphi \in V} \int |f(x) - \varphi(x)| dx \text{ and } \min_{\varphi \in V} \sup_x |f(x) - \varphi(x)|$$

may not hold.

However, the above theorem only provides a sufficient condition for uniqueness.

1.4.3. Examples

– We consider the case when

$$F := (\mathbb{R}^2, \|\cdot\|_\infty)$$

and

$$\|(x, y)'\|_\infty := \max\{|x|, |y|\}.$$

The linear space V is defined by

$$V := \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} =: \text{span} \{u\}.$$

The vector u in $\mathbb{R}^2$ has infinitely approximating vectors in V , given by the family of all vectors $v_x := (x, 0)$ where $-1 \leq x \leq 1$, for which $\|v_x - u\|_\infty = 1$ for all x .

– The space $(C^{(0)}([a, b]), \|\cdot\|_\infty)$ is a normed linear space ($\|\cdot\|_\infty := L_\infty$). Let $V := \mathcal{P}_n([a, b])$.

For all f in $C^{(0)}([a, b])$, the optimal solution exists and is unique, as will be seen in Chapter 2 and it satisfies

$$\varphi^* = \arg \min_{\varphi \in V} \|f - \varphi\|_\infty.$$

– Non-existence of the optimal solution.

We explore the approximation of bounded real valued sequences, a context slightly different from the case of continuous functions. Consider the set of all bounded sequences of real numbers f_x , with $X = \mathbb{N}$, namely

$$l^\infty := \left\{ \begin{array}{l} f : X = \mathbb{N} \rightarrow \mathbb{R}, x \mapsto f_x \\ \text{and there exists } c \in \mathbb{R}^+, \text{ such that for all } x, |f_x| \leq c \end{array} \right\}.$$

Elements in l^∞ are denoted by $(f_x)_x$.

Define

$$\|(f_x)_x\|_{l^\infty} := \sum_{x=1}^{\infty} \frac{1}{2^x} |f_x|,$$

which defines a normed linear space $(l^\infty, \|\cdot\|_{l^\infty})$.

Let M denote the subset of l^∞

$$M := \left\{ (\varphi_x)_x \in l^\infty : \lim_{x \rightarrow \infty} \varphi_x = 0 \right\}. \quad [1.4]$$

This set is convex, being a linear subspace in l^∞ ; indeed when

$$\left((\varphi_x^{(1)})_x, (\varphi_x^{(2)})_x \right) \in M^2,$$

then for any $(\alpha_1, \alpha_2) \in \mathbb{R}^2$,

$$\lim_{x \rightarrow \infty} \left(\alpha_1 \varphi_x^{(1)} + \alpha_2 \varphi_x^{(2)} \right) = 0.$$

Consider the approximation of the element of l^∞ which is defined by

$$(1)_x : X = \mathbb{N} \rightarrow \mathbb{R}, x \mapsto 1,$$

the sequence with constant value 1.

Clearly, $(1)_x \in l^\infty$.

The best approximation $(\varphi_x^*)_x$ of $(1)_x$ in M should satisfy

$$\|(1)_x - (\varphi_x^*)_x\|_{l^\infty} = \inf_{\varphi \in M} \|(1)_x - (\varphi_x)_x\|_{l^\infty}.$$

Now, we have

$$\inf_{\varphi \in M} \|(1)_x - (\varphi_x)_x\|_{l^\infty} = 0.$$

Therefore,

$$(\varphi_x)_x = (1)_x.$$

Obviously,

$$\lim_{n \rightarrow \infty} (1)_x = 1,$$

which entails

$$(1)_x \notin M.$$

In order to obtain

$$\inf_{\varphi \in M} \|(1)_x - (\varphi_x)_x\|_{l^\infty} = 0,$$

it is enough to consider, for any positive n in $\mathbb{N}$

$$(\varphi_{x,n})_x := \begin{cases} 1 & \text{for } n < x \\ 0 & \text{for } n \geq x + 1 \end{cases}$$

which belongs to M (see [1.4]).

We then have

$$\|(1)_x - (\varphi_{x,n})_x\|_{l^\infty} = \sum_{x=1}^{\infty} \frac{1}{2^x} |1 - \varphi_{x,n}| = \frac{1}{2^x}.$$

Choosing n arbitrarily yields

$$\lim_{n \rightarrow \infty} \|(1)_x - (\varphi_{x,n})_x\|_{l^\infty} = \lim_{x \rightarrow \infty} \frac{1}{2^x} = 0.$$

The L_2 norm is strict; since it is defined through an inner product, one may ask whether all norms defined through an inner product are strict.

The following theorem answers this question.

THEOREM 1.4.— Let $\langle \cdot, \cdot \rangle_F$ be an inner product on some linear space F . Then, the norm $\|f\|_F$ defined by $\|f\|_F := \langle f, f \rangle_F$ is a strict norm.

PROOF.— Let f_1 and f_2 be two distinct elements in F , such that $\|f_1\|_F = \|f_2\|_F = 1$ assuming that $r = 1$ in definition 1.1, without loss of generality. By Schwartz inequality, $\langle f_1, f_2 \rangle_F \leq \|f_1\|_F \|f_2\|_F$. Equality holds if and only if f_1 is a multiple of f_2 . First, consider the case when equality holds. Then,

$$\left\| \frac{f_1 + f_2}{2} \right\|_F^2 = \frac{1}{4} \langle f_1 + f_2, f_1 + f_2 \rangle_F = \frac{1}{2} (1 + \langle f_1, f_2 \rangle_F) < 1.$$

Since $f_1 = \alpha f_2$ and $\|f_1\|_F = \|f_2\|_F$, it follows that $\|f_1\|_F = |\alpha| \|f_2\|_F$. Hence, $|\alpha| = 1$. Now, $f_1 \neq f_2$. Therefore, we cannot have $\alpha = 1$. Hence, $f_1 = -f_2$. In this case, $\|f_1 + f_2\|_F^2 = 0 < 2$. Strict convexity then holds. Consider, now the case when $f_1 \neq f_2$, with $\|f_1\|_F = \|f_2\|_F = 1$ (in the case when the norm is not one, consider $f_1/\|f_1\|_F \neq f_2/\|f_2\|_F$). Since $\langle f_1, f_2 \rangle_F < \|f_1\|_F \|f_2\|_F = 1$ (strict inequality follows from the fact that $f_1 \neq f_2$)

$$\frac{1}{4} < \langle f_1 + f_2, f_1 + f_2 \rangle_F = \frac{1}{2} (1 + \langle f_1, f_2 \rangle_F) < 1$$

which concludes the proof. ■

1.5. Characterizing the optimal solution

1.5.1. The Hilbertian case

THEOREM 1.5.— Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space and M a non-void closed and convex subset in H . Assume further $M \subseteq V$, where V is a finite dimensional subspace in H . Then, for any $f \in H$, there exists a best approximation $\varphi_M \in M$ with respect to the norm (defined by the inner product). Further to this, φ_M satisfies

$$\langle f - \varphi_M, \varphi - \varphi_M \rangle_H \leq 0 \tag{1.5}$$

for any φ in M . Reciprocally if [1.5] holds for all φ in M , then φ_M is the best approximation of f with respect to the norm.

PROOF.— Existence and uniqueness of φ_M follow from arguments developed above. We have to prove that the two following statements,

φ_M is the best approximation of f in M

and

$$\langle f - \varphi_M, \varphi - \varphi_M \rangle_H \leq 0, \text{ for any } \varphi \in M,$$

are equivalent. Simple calculus proves that

$$\|f - \varphi_M\|_H^2 - \|f - \varphi\|_H^2 + \|\varphi - \varphi_M\|_H^2 = 2 \langle f - \varphi_M, \varphi - \varphi_M \rangle_H.$$

Hence, assuming [1.5], we get

$$\|f - \varphi_M\|_H^2 = \|f - \varphi\|_H^2 + 2$$

$$\begin{aligned} & \langle f - \varphi_M, \varphi - \varphi_M \rangle_H - \|\varphi - \varphi_M\|_H^2 \\ & \leq \|f - \varphi\|_H, \text{ for all } \varphi \in M \end{aligned}$$

which proves that φ_M is the best approximation of f in M . Suppose now that $\varphi_M \in M$ is the best approximation of f in M and argue by contradiction. Suppose that

$$\langle f - \varphi_M, \varphi - \varphi_M \rangle_H > 0.$$

Consider the function

$$g(\alpha) := \|f - [(1 - \alpha)\varphi_M + \alpha\varphi]\|_H^2$$

as $\alpha \rightarrow 0$. In a neighborhood $\mathcal{I}_0$ of 0, $g'(\alpha)$ takes negative values. Therefore, g is decreasing on $\mathcal{I}_0$. There exists some $\alpha' \in (0; 1)$ such that the function $\tilde{\varphi} := (1 - \alpha')\varphi_M + \alpha'\varphi$ satisfies

$$\|f - \tilde{\varphi}\|_H^2 = g(\alpha') < g(0) = \|f - \varphi_M\|_H^2.$$

Since M is convex $\tilde{\varphi} \in M$. Henceforth, $\tilde{\varphi}$ is a better approximation to f than φ_M , which concludes the proof. ■

When M is a finite dimensional linear subspace of H , and not any convex subset in H as previously, then the preceding result takes a simpler form, as shown below.

THEOREM 1.6.— Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space and M be a linear subspace of H with finite dimension. Then, there exists a unique best approximation φ_M of f in H with respect to the norm. It is characterized by the following property

$$\langle f - \varphi_M, \varphi \rangle_H = 0, \text{ for any } \varphi \text{ in } M.$$

PROOF.— *Direct part.* Assume that φ_M is optimal. We prove that $\langle f - \varphi_M, \varphi \rangle_H = 0$ holds for all $\varphi \in M$. Since M is a linear space, it contains some g defined by

$$g = \varphi_M + \alpha\varphi,$$

where $\alpha \in \mathbb{R}$ and $\varphi \in M$. It holds

$$\begin{aligned} \|f - g\|_H^2 &= \|f - (\varphi_M + \alpha\varphi)\|_H^2 \\ &= \langle f - \varphi_M - \alpha\varphi, f - \varphi_M - \alpha\varphi \rangle_H \\ &= \langle f - \varphi_M, f - \varphi_M \rangle_H - \alpha \langle f - \varphi_M, \varphi \rangle_H \\ &\quad - \alpha \langle f - \varphi_M, \varphi \rangle_H + \alpha^2 \langle \varphi, \varphi \rangle_H \end{aligned}$$

$$\begin{aligned}
&= \langle f - \varphi_M, f - \varphi_M \rangle_H - 2\alpha \langle f - \varphi_M, \varphi \rangle_H + \alpha^2 \langle \varphi, \varphi \rangle_H \\
&= \|f - \varphi_M\|_H^2 - 2\alpha \langle f - \varphi_M, \varphi \rangle_H + \alpha^2 \|\varphi\|_H^2.
\end{aligned}$$

By hypothesis, φ_M is the optimal solution. Therefore, since g belongs to M , $\|f - \varphi_M\|_H^2 \leq \|f - g\|_H^2$. Therefore,

$$\|f - \varphi_M\|_H^2 \leq \|f - \varphi_M\|_H^2 - 2\alpha \langle f - \varphi_M, \varphi \rangle_H + \alpha^2 \|\varphi\|_H^2$$

i.e.

$$0 \leq -2\alpha \langle f - \varphi_M, \varphi \rangle_H + \alpha^2 \|\varphi\|_H^2,$$

for any $\alpha \in \mathbb{R}$, and any $\varphi \in M$. We consider the two following cases:

1) For α , choose the k -th term α_k of any positive convergent sequence with limit 0. Therefore,

$$-2\alpha_k \langle f - \varphi_M, \varphi \rangle_H + \alpha_k^2 \|\varphi\|_H^2 \geq 0$$

i.e.

$$-2 \langle f - \varphi_M, \varphi \rangle_H + \alpha_k \|\varphi\|_H^2 \geq 0.$$

Going to the limit in k yields

$$-2 \langle f - \varphi_M, \varphi \rangle_H \geq 0,$$

hence,

$$\langle f - \varphi_M, \varphi \rangle_H \leq 0, \text{ for any } \varphi \in M.$$

2) The second case is similar with the choice of a negative sequence going to 0. It holds

$$-2 \langle f - \varphi_M, \varphi \rangle_H \leq 0$$

i.e.

$$\langle f - \varphi_M, \varphi \rangle_H \geq 0, \text{ for all } \varphi \in M.$$

Therefore, we obtain

$$\langle f - \varphi_M, \varphi \rangle_H \leq 0, \text{ for all } \varphi \in M$$

and $\langle f - \varphi_M, \varphi \rangle_H \leq 0$, for all $\varphi \in M$

which yields

$$\langle f - \varphi_M, \varphi \rangle_H = 0, \text{ for all } \varphi \in M.$$

Reciprocal. We now turn to the reciprocal statement, namely stating that when

$$\langle f - \varphi_M, \varphi \rangle_H = 0,$$

for any $\varphi \in M$, then φ_M is the best approximation of f in M . Since M is a linear subspace of H it is convex. From the preceding theorem whenever $\langle f - \varphi_M, \varphi \rangle_H = 0$ for all φ in M , then φ_M is the best approximation of f in M with respect to the norm. This concludes the proof. ■

REMARK.— We note that

$$\langle f - \varphi_M, \varphi \rangle_H = 0$$

for any $\varphi \in M$ if and only if $(f - \varphi_M) \in M^\perp$. Furthermore, since $\varphi_M \in V := \text{span}\{\varphi_1, \dots, \varphi_n\}$, there exists a unique set of coefficients $a^* := (a_1^*, \dots, a_n^*)' \in \mathbb{R}^n$ such that $\varphi_M = \sum_{i=1}^n a_i^* \varphi_i$. The relation $\langle f - \varphi_M, \varphi \rangle_H = 0$, for all $\varphi \in M$ identifies the vector a^* . Indeed, $a^* = G^{-1}\mathbf{f}$, where

$$\mathbf{f} := (\langle f, \varphi_i \rangle_{i=1, \dots, n})$$

and

$$G := (\langle \varphi_i, \varphi_j \rangle)_{i=1, \dots, n; j=1, \dots, n}.$$

Clearly, G (the Gram matrix) is diagonal whenever the basis is orthogonal.

1.5.2. The non-Hilbertian case

We consider the case when F is not a Hilbert space but merely a normed linear space and when V is a finite dimensional subspace of F .

The following theorem then holds.

THEOREM 1.7.— Let $(F, \|\cdot\|_F)$ be a normed linear space and V a subspace of F . Let $\pi_V : F \rightarrow V$ be a projection of F onto V . Then,

$$\|f - \pi_V(f)\|_F \leq (1 + \|\pi_V\|) \inf_{\varphi \in V} \|f - \varphi\|_F.$$

In the above display, $|||\pi_V|||$ is the norm of the linear operator, defined by

$$|||\pi_V||| := \sup_{f \in F, f \neq 0} \frac{\|\pi_V(f)\|_F}{\|f\|_F}.$$

PROOF.— Obviously, $\pi_V(\varphi) = \varphi$, for any $\varphi \in V$. Therefore,

$$\begin{aligned} \|f - \pi_V(f)\|_F &= \|f - \pi_V(f) + \varphi - \varphi\|_F \\ &= \|(f - \varphi) + \pi_V(\varphi - f)\|_F \\ &\leq \|f - \varphi\|_F + \|\pi_V(f - \varphi)\|_F. \end{aligned}$$

Since π_V is linear and continuous, there exists some constant

$$|||\pi_V||| := \sup_{f \in F, f \neq 0} \frac{\|\pi_V(f)\|_F}{\|f\|_F}$$

such that

$$\|\pi_V(f - \varphi)\|_F \leq |||\pi_V||| \|f - \varphi\|_F.$$

Therefore,

$$\begin{aligned} \|f - \pi_V(f)\|_F &\leq \|f - \varphi\|_F + \|\pi_V(\varphi - f)\|_F \\ &\leq \|f - \varphi\|_F + |||\pi_V||| \|f - \varphi\|_F \\ &= \|f - \varphi\|_F (1 + |||\pi_V|||). \end{aligned}$$

Consider in both sides of the above display, the infimum with respect to all choices of $\varphi \in V$; this yields

$$\|f - \pi_V(f)\|_F \leq (1 + |||\pi_V|||) \inf_{\varphi \in V} \|f - \varphi\|_F. \quad \blacksquare$$

The above theorem proves that up to the constant $((1 + |||\pi_V|||))$, which depends on π_V only, the error $\|f - \pi_V(f)\|_F$ is minimal.

1.5.2.1. The case L_∞

We now assume that $V := \text{span} \{\varphi_1, \dots, \varphi_n\}$ is a linear subspace of (F, L_∞) , the normed linear space F equipped with the L_∞ norm. We assume that $\dim V = n$ and denote $[a, b]$ the common domain of all functions in F .

Since $\{\varphi_1, \dots, \varphi_n\}$ is a basis of V for all $\varphi \in V$, $\varphi = \sum_{i=1}^n a_i \varphi_i$. In order to emphasize the fact that φ is characterized by its coefficients in the basis $\{\varphi_1, \dots, \varphi_n\}$, we denote it by $\varphi_{\mathbf{a}}$.

The error committed when approximating f by $\varphi_{\mathbf{a}}$ is measured by $L_\infty(f - \varphi_{\mathbf{a}})$. In order to minimize this error, we optimize φ by an adequate tuning of the coefficients a_i 's. This can be achieved choosing n nodes $x_1, \dots, x_n$ in the domain X of φ . We therefore fix n values $(y_1, \dots, y_n)$ and we define φ as the element in V which satisfies $\varphi(x_i) = y_i, i = 1, \dots, n$. We thus solve the system

$$\varphi(x_j) = \sum_{i=1}^n a_i \varphi_i(x_j), \quad j = 1, \dots, n$$

with respect to the coefficients $a_1, \dots, a_n$. The resulting function is $\varphi_{\mathbf{a}}(x) = \sum_{i=1}^n a_i \varphi_i(x)$.

Assume that $f \in \mathcal{C}^{(0)}([a, b])$. The approximation problem is, as usual

$$\min_{\mathbf{a} \in \mathbb{R}^n} \sup_{x \in [a, b]} \left| f(x) - \sum_{i=1}^n a_i \varphi_i(x) \right|.$$

Clearly, the error is minimal when the error curve

$$x \rightarrow R(\mathbf{a}, x) := f(x) - \sum_{i=1}^n a_i \varphi_i(x),$$

is the null function, i.e. when φ coincides with f .

If $f \notin \text{span} \{\varphi_1, \dots, \varphi_n\}$, this never happens. We can alter this in such a way that f and φ intersect in at least n points, imposing the conditions

$$f(x_j) = \sum_{i=1}^n a_i \varphi_i(x_j), \quad i, j = 1, \dots, n$$

i.e. $y_j = f(x_j)$. Therefore, we require that φ coincides with f on the n nodes x_j .

Therefore, the curve $x \rightarrow R(\mathbf{a}, x)$ has n zeros and changes sign $n + 1$ times.

Consider the $n + 1$ abscissas $z_1, \dots, z_{n+1}$ where the curve $x \rightarrow R(\mathbf{a}, x)$ takes its maximal values with alternating signs, namely denote $R(\mathbf{a}, z_j)$ the maximal value (up to the sign) of the error on the interval (x_j, x_{j+1}) . Up to now, their values would not be equal, in absolute value. The striking result of the Borel–Chebyshev theorem 3.4, to be stated in Chapter 3, is that optimality of this procedure is attained when all the $|R(\mathbf{a}, z_j)|$'s are equal, and the $R(\mathbf{a}, z_j)$'s have alternating signs. We state this loosely as:

If f is a continuous function defined on $[a, b]$ and $\{\varphi_1, \dots, \varphi_n\}$ are continuous independent functions defined on $[a, b]$, a necessary and sufficient condition in order that $\sum_{i=1}^n a_i^* \varphi_i(x)$ be the best uniform approximation of f in $\text{span}\{\varphi_1, \dots, \varphi_n\}$, supposing $f \notin \text{span}\{\varphi_1, \dots, \varphi_n\}$, is provided by the fact that the error curve $x \rightarrow R(\mathbf{a}^*, x)$ changes sign at least at $n + 1$ points in $[a, b]$. In those points, $x \rightarrow |R(\mathbf{a}^*, x)|$ should assume its maximal values in $[a, b]$.

1.5.3. Optimization, L_p norms and robustness

In the preceding discussions, the function f was considered fixed and known, or at least could be evaluated at any point of its domain, without any uncertainty. We introduce some notion of robustness which will come to be of upmost importance in the statistical treatment of the approximation problem.

Consider a function f defined on some interval X . We assume that f is approximated by some function φ . Let

$$R(x) := f(x) - \varphi(x)$$

be the local error due to φ .

Let $x \rightarrow \delta(x)$ be some error committed measuring f . Denoting $f_\delta := f + \delta$ the corresponding function, we consider the error committed approximating f_δ by φ . Denote therefore

$$R_\delta(x) := f_\delta(x) - \varphi(x) = R(x) + \delta(x).$$

The question to be answered is the following: according to the choice of p , which type of measurement error $\delta(x)$ will make the norm $L_p(R_\delta)$ of the error committed using φ for f_δ close to $L_p(R)$? In this case, the couple (p, δ) is called robust. It holds, as $\delta(x)/R(x) \rightarrow 0$

$$\frac{(R(x) + \delta(x))^p - (R(x))^p}{\delta(x)} \sim p(R(x))^{p-1}$$

from which, we have

$$(R(x) + \delta(x))^p - (R(x))^p \sim p(R(x))^{p-1} \delta(x)$$

assuming $\delta(x)$ is small with respect to $R(x)$.

Integrating,

$$|L_p(R + \delta) - L_p(R)| \lesssim p \int_X |\delta(x) (R(x))^{p-1}| dx$$

which means that $L_p(R + \delta)$ is nearly between

$$L_p(R) - p \int_X |\delta(x) (R(x))^{p-1}| dx$$

and

$$L_p(R) + p \int_X |\delta(x) (R(x))^{p-1}| dx.$$

We are interested in the role of a small variation δ of the approximating function φ on the L_p norm of the error. When $p > 1$, the L_p norm does not change significantly when δ assumes even large values at points x where $R(x) := f(x) - \varphi(x)$ is small. When $p < 1$, then a contrary statement holds: the L_p norm of the error does not change when $\delta(x)$ takes small values and $R(x)$ is large. When $p = 1$, the change in the norm of the error is of the order of the integral of δ on X .

