



HAL
open science

Optimal Smoothing and Gaussian Processes with noisy data under constraints

Laurence Grammont, Xavier Bay, Hassan Maatouk

► **To cite this version:**

Laurence Grammont, Xavier Bay, Hassan Maatouk. Optimal Smoothing and Gaussian Processes with noisy data under constraints. 2022. hal-03625227

HAL Id: hal-03625227

<https://hal.science/hal-03625227v1>

Preprint submitted on 30 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal Smoothing and Gaussian Processes with noisy data under constraints

Laurence Grammont · Hassan Maatouk ·
Xavier Bay

Received: date / Accepted: date

Abstract In this paper, we extend the correspondence between Bayesian estimation and optimal smoothing in a Reproducing Kernel Hilbert Space (RKHS) adding a convex constraint on the solution. Through a sequence of approximating Hilbertian spaces and a discretized model, we prove that the Maximum A Posteriori (MAP) of the posterior distribution is exactly the optimal constrained smoothing function in the RKHS. This paper can be read as a generalization of the paper [11] of Kimeldorf-Wahba where it is proved that the optimal smoothing solution is the mean of the posterior distribution. This is also a generalization of the paper [4] where the case of constrained optimal interpolation is treated. Here we relax the interpolation by introducing noise effect in the data. A numerical example is given to illustrate the theoretical result of this paper.

Keywords Correspondence · Smoothing · Inequality constraints · Reproducing Kernel Hilbert Space · Bayesian estimation

1 Introduction

Consider X a nonempty set of \mathbb{R} and E a Banach space of functions from X to \mathbb{R} . Given data $(x_i, y_i) \in X \times \mathbb{R}$, the smoothing problem is to find a function \hat{u} minimizing

$$\|u\|_H^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (u(x_i) - y_i)^2 \quad (1)$$

on an Hilbert subspace H of E whose injection in E is continuous.

As Kimeldorf and Wahba explained it in [11], the term $\|u\|_H^2$ is the smoothness

L. Grammont

Université de Lyon, Institut Camille Jordan, UMR 5208, 23 rue du Dr Paul Michelon, 42023 Saint-Étienne Cedex 2, France. E-mail: laurence.grammont@univ-st-etienne.fr

H. Maatouk

CY Tech Cergy Paris University, Site du Parc, 95011 Cergy-Pontoise, France.
E-mail: hmk@cy-tech.fr

X. Bay

Mines Saint-Étienne, UMR 6158, LIMOS, F-42023 Saint-Étienne, France. E-mail: bay@emse.fr

criterion for the solution and $\frac{1}{\sigma^2} \sum_{i=1}^n (u(x_i) - y_i)^2$ measures the disparity of u with the data. \hat{u} is a compromise between smoothness and fidelity to the data. In [11], the disparity of the data is measured by $\sum_{i=1}^n \sum_{j=1}^n (u(x_i) - y_i) b^{ij} (u(x_j) - y_j)$. We choose $b^{ij} = \frac{1}{\sigma^2} \delta_{ij}$ with no loss of generality, to facilitate subsequent reading of the paper. In [11], the authors consider $\|u\|_H^2 := \int_{-\infty}^{+\infty} (Lu)^2(t) dt$, where L is a linear differential operator. In that case, under conditions, the solution \hat{u} is an L -spline.

In this paper, H is any Reproducing Kernel Hilbert Space equipped with its associated norm. The notion of Reproducing Kernel Hilbert Space (RKHS) is the material with which one has built a bridge from the deterministic world of optimization to the probabilistic world of estimation. Aronszajn [1] published the theory of reproducing kernel in 1950 and Parzen [12] published Statistical inference on time series by Hilbert space methods in 1959. Later Schwartz [13] extended its formalism to topological spaces. Through its covariance function, $K(x, x') = \text{cov}(U(x), U(x'))$, a Gaussian process (GP) U is represented by the Hilbert space spanned by the kernel K .

The aim of Kimeldorf and Wahba in [11] is to highlight the correspondence between the smoothing by spline and Bayesian estimation. For that purpose, they consider a stochastic model in which the selection of the smoothing criterion corresponds to the specification of a prior distribution on U which is a centered GP with covariance function $K(x, x')$. The disparity of the data corresponds to noisy observations in the uncertainty framework. At points x_1, \dots, x_n the random variables

$$Y_i = U(x_i) + \mathcal{E}_i, \quad (2)$$

are observed, where $\mathcal{E} = (\mathcal{E}_i)_i$ is a centered Gaussian vector $\mathcal{N}(0, \sigma^2 I)$, where I is the $(n \times n)$ identity matrix and σ^2 is the noise variance. They prove that \hat{u} solution of the minimization of (1) is the Bayesian estimation

$$\hat{u}(x) = \mathbb{E}[U(x) | Y_1 = y_1, \dots, Y_n = y_n], \quad (3)$$

where \mathbb{E} denotes expectation. In other terms, it means that we look for u in the RKHS H associated to the centered GP $(U(x))_{x \in X}$ whose covariance is K .

In both framework, one can prove that the solution of the smoothing problem (1) or Bayesian estimation (3) is

$$\hat{u}(x) = \mathbf{y} \left(\mathbf{K} + \sigma^2 I \right)^{-1} \mathbf{k}(x)^\top, \quad (4)$$

where $\mathbf{k}(x) = (K(x_1, x), \dots, K(x_n, x))$, \mathbf{K} is the matrix $(K(x_i, x_j))_{1 \leq i, j \leq n}$ and $\mathbf{y} = (y_1, \dots, y_n)$.

Now, consider that u is known to satisfy some additional constraints given by $u \in C$ where C is a closed convex set. In the present paper, we consider the *constrained* smoothing problem of finding a function \hat{u} , in H and C , minimizing

$$\|u\|_H^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (u(x_i) - y_i)^2, \quad (5)$$

over $H \cap C$. In [10], Theorem 3.1, Micchelli and Utreras proved that the solution exists and is unique under certain conditions. They also give the expression of the solution involving the projection on the convex set C , denoted by P_C , and a nonlinear algebraic system expressed with P_C and approximated usually by some Newton type methods (see [6]). As Andersson and Elfving wrote it in their paper [2], to transform this result into a numerical algorithm, it is necessary to compute the orthogonal projection P_C and the difficulty lies in that calculation. Andersson and Elfving investigated the structure of the projection operator P_C for a particular convex set C representing monotonicity constraints.

The aim of the present paper is to rewrite the constrained smoothing problem (5) with a stochastic model so that the solution can be interpreted as a Bayesian estimation. The difficulty is that the equivalent of the constrained optimal smoothing solution is a truncated GP so that its distribution can not be expressed as a density probability. We found out that we can overcome the difficulty through a discretization of the constrained smoothing problem whose solution \hat{u}_N tends to \hat{u} . The integer N is the discretization parameter. Then we define the equivalent finite-dimensional approximation U_N of the GP U and we prove that the approximate solution \hat{u}_N can be interpreted as a Bayesian estimation. This estimation is the MAP (Maximum A Posteriori) of the posterior distribution of U_N and not the mean like in the smoothing problem without constrained of [11].

In [4], the problem was to minimize a smoothness criterion under constraints, interpolating strictly the observations. It was proved the MAP of the posterior distribution tends uniformly to the optimal constrained interpolation function. Here, relaxing the interpolation of observations through a noise effect, we obtain the same result but its proof is completely different. It is much simpler. The main theoretical difficulty of the interpolation case treated in [4] is that the interpolation equalities are incorporated in the definition of the approximation spaces. Surprisingly, the proof of this paper relies on classical optimization arguments. The only thing to care about is the construction of the finite-dimensional functional to minimize over the right finite-dimensional Reproducing Kernel Hilbert Space. More over, as it is less restrictive for the sample spaces than strict interpolation, it carries cheaper computations.

The paper is organized as follows. Section 2 is devoted to the description of the deterministic problem and Section 3 to its discretization. The proof of the convergence of the discretized problem to the constrained optimal smoothing solution is done in Section 4. In Section 5, we prove the correspondence between Constrained Optimal Smoothing and Bayesian estimation on stochastic processes. Section 6 is devoted to the numerical illustration of the theoretical result. In Section 7, we

quote practical algorithms induced by our result in the stochastic processes community. In them, assumption (H3) is not fulfilled. The proof of our convergence result, without this assumption, is done in the appendix.

2 Framework of Constrained Optimal Smoothing

To simplify the paper, we suppose that $X = [0, 1]$ and $E = C([0, 1], \mathbb{R})$ is the linear space of real valued continuous functions on $[0, 1]$ equipped with the supremum norm. Let H be a RKHS of E associated to the symmetric positive definite function K . Then, H is an Hilbertian subspace of E since

$$\|h\|_E = \sup_{x \in X} |(h, K(\cdot, x))_H| \leq c \|h\|_H, \quad (6)$$

where $c = \sup_{x \in X} K(x, x)^{1/2} < +\infty$. Let $(x_i, y_i)_{1 \leq i \leq n}$ be the given data. Let us define the function $J : H \rightarrow \mathbb{R}$ by

$$J(u) := \|u\|_H^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (u(x_i) - y_i)^2. \quad (7)$$

The Constrained Optimal Smoothing (5) can be rewritten as

$$\min_{u \in H \cap C} J(u) \quad (P)$$

where C is a closed convex set of E such that

$$H \cap C \neq \emptyset. \quad (H1)$$

It is easy to see that J is Fréchet differentiable and $\lim_{\|v\|_H \rightarrow +\infty} J(v) = +\infty$. Moreover J is strongly convex: for all $u, v \in H$ and $t \in [0, 1]$,

$$J(tu + (1-t)v) \leq tJ(u) + (1-t)J(v) - t(1-t)\|u - v\|_H^2.$$

The problem (P) has a unique solution, denoted by \hat{u} .

3 Discretization of Constrained Optimal Smoothing

We propose a discretized optimization problem (P_N) of (P) associated with Δ_N a subdivision of $[0, 1]$

$$\Delta_N : \quad 0 = t_0 < t_1 < \dots < t_N = 1, \quad (8)$$

such that $\delta_N = \max\{|t_{i+1} - t_i|, i = 0, \dots, N-1\}$ tends to zero as N tends to infinity. We assume

$$\Delta_N \subset \Delta_{N+1} \quad (H2)$$

and the data points x_i belong to the partition for N large enough

$$\{x_1, \dots, x_n\} \subset \Delta_N. \quad (H3)$$

Let H_N be the classical subspace of piecewise linear continuous functions associated to Δ_N . A basis of H_N is the so-called hat functions denoted by $(\varphi_0, \dots, \varphi_N)$. Next, we define π_N to be the classical piecewise linear interpolation projection defined from E onto H_N by

$$\forall f \in E, \quad \pi_N(f) = \sum_{j=0}^N f(t_j) \varphi_j.$$

We assume that

$$\pi_N(C) \subset C. \quad (H4)$$

According to a classical approximation result

$$\pi_N(f) \xrightarrow{N \rightarrow +\infty} f \quad \text{in } E. \quad (9)$$

Let us define the linear evaluation operator $\mathcal{I}_N : E \mapsto \mathbb{R}^{N+1}$ on the nodes t_i and $\mathcal{I}_n : E \mapsto \mathbb{R}^n$ on the data points x_i :

$$\begin{aligned} \forall f \in E, \quad \mathcal{I}_N(f) &:= (f(t_0), \dots, f(t_N))^\top; \\ \forall f \in E, \quad \mathcal{I}_n(f) &:= (f(x_1), \dots, f(x_n))^\top. \end{aligned}$$

To lighten the notations, c_f denotes $\mathcal{I}_N(f)$. Let us define the matrix of K on the nodes as follows:

$$\Gamma_N = (K(t_i, t_j))_{0 \leq i, j \leq N}. \quad (10)$$

We suppose

$$\Gamma_N \quad \text{is invertible.} \quad (H5)$$

As H is a RKHS associated to the kernel K , we can define a new scalar product on H_N : for all $u_N, v_N \in H_N$

$$(u_N, v_N)_{H_N} := c_{u_N}^\top \Gamma_N^{-1} c_{v_N} \quad (11)$$

which induces a norm on H_N : $\forall u_N \in H_N$

$$\|u_N\|_{H_N}^2 = c_{u_N}^\top \Gamma_N^{-1} c_{u_N}. \quad (12)$$

Let us define the linear operator $\rho_N : H_N \rightarrow H$ defined by

$$\forall v_N \in H_N, \quad \rho_N(v_N) := \sum_{i=0}^N \lambda_i K(\cdot, t_i), \quad (13)$$

where $\Lambda = (\lambda_0, \dots, \lambda_N)^\top$, solves

$$\Gamma_N \Lambda = c_{v_N}.$$

Let us notice that ρ_N has been defined so that $\rho_N \circ \pi_N$ is the orthogonal projection from H onto $H_{1,N}$ where

$$H_{1,N} = \text{Span} \{K(\cdot, t_j), j = 0, \dots, N\}. \quad (14)$$

The following proposition highlights the nature of the finite-dimensional space H_N :

Proposition 1 H_N is a RKHS with kernel K_N given by

$$\forall x', x \in [0, 1], \quad K_N(x', x) = \sum_{i,j=0}^N K(t_i, t_j) \varphi_j(x) \varphi_i(x'). \quad (15)$$

For any $h_N \in H_N$,

$$\|h_N\|_E \leq d \|h_N\|_{H_N}, \quad (16)$$

where d is a constant independent of N .

Proof Clearly, H_N is a finite-dimensional Hilbert space. Let x be in $[0, 1]$. We have

$$K_N(\cdot, x) = \sum_{i=0}^N \lambda_{i,x} \varphi_i \in H_N,$$

where $\lambda_{i,x} = \sum_{j=0}^N K(t_i, t_j) \varphi_j(x) = (\Gamma_N \varphi(x))_i$, with

$$\varphi(x) := (\varphi_0(x), \dots, \varphi_N(x))^\top. \quad (17)$$

Let $h := \sum_{i=0}^N \alpha_i \varphi_i = \alpha^\top \varphi(x) \in H_N$, $\alpha := (\alpha_0, \dots, \alpha_N)^\top$. We obtain

$$(h, K_N(\cdot, x))_{H_N} = \alpha^\top \Gamma_N^{-1} (\Gamma_N \varphi(x)) = \alpha^\top \varphi(x) = h(x),$$

which is the reproducing property in H_N . For $x \in X$, we have

$$|h(x)| = |(h, K_N(\cdot, x))_{H_N}| \leq \|h\|_{H_N} \times \sqrt{K_N(x, x)},$$

where $K_N(x, x) = \sum_{i,j=0}^N K(t_i, t_j) \phi_i(x) \phi_j(x)$. Since $\sum_{i,j=0}^N \phi_i(x) \phi_j(x) = 1$, we obtain

$$0 \leq \sup_{x \in X} K_N(x, x) \leq M = \max_{x, x' \in X} |K(x, x')|.$$

As $\|h_N\|_E = \|h_N\|_\infty$, the proof of the lemma is completed. \square

In the following proposition, one proves that the sequence of projections π_N is stable.

It is straightforward that for all f in E ,

$$\|\pi_N(f)\|_{H_N}^2 = c_f^\top \Gamma_N^{-1} c_f.$$

Proposition 2 (Stability of π_N) π_N is stable, i.e.

$$\forall h \in H, \quad \|\pi_N(h)\|_{H_N} \leq \|h\|_H. \quad (18)$$

Moreover H is characterized by

$$H = \left\{ h \in E : \sup_N \|\pi_N(h)\|_{H_N} < +\infty \right\} \quad (19)$$

and, for all $h \in H$, by

$$\|f\|_H^2 = \lim_{N \rightarrow +\infty} \|\pi_N(f)\|_{H_N}^2. \quad (20)$$

Proof Consider the usual orthogonal decomposition in the RKHS H : $H = H_{0,N} \oplus^\perp H_{1,N}$, with

$$\begin{aligned} H_{0,N} &= \{h \in H : h(t_j) = 0, j = 0, \dots, N\}, \\ H_{1,N} &= \text{Span}\{K(\cdot, t_j), j = 0, \dots, N\}. \end{aligned}$$

For all $h \in H$, there exists a unique $h_0 \in H_{0,N}$ and $h_1 \in H_{1,N}$ such that $h = h_0 + h_1$. Thus,

$$\|h_1\|_H^2 \leq \|h\|_H^2.$$

Additionally, every $h_1 \in H_{1,N}$ can be expressed as $h_1(\cdot) = \sum_{j=0}^N \alpha_j K(\cdot, t_j)$. From the reproducing property $(K(\cdot, t_j), K(\cdot, t_i))_H = K(t_i, t_j)$, we get

$$\|h_1\|_H^2 = (h_1, h_1)_H = \sum_{i,j=0}^N \alpha_i \alpha_j K(t_i, t_j) = \alpha^\top \Gamma_N \alpha.$$

As $h_1(t_i) = \sum_{j=0}^N \alpha_j K(t_i, t_j)$ for $i = 0, \dots, N$, we have $\alpha = \Gamma_N^{-1} c_{h_1}$ and

$$\|h_1\|_H^2 = c_{h_1}^\top \Gamma_N^{-1} \Gamma_N \Gamma_N^{-1} c_{h_1} = c_{h_1}^\top \Gamma_N^{-1} c_{h_1}.$$

Since $h_0 \in H_{0,N}$, $c_{h_1} = c_h$ and $\|h_1\|_H^2 = c_h^\top \Gamma_N^{-1} c_h = \|\pi_N(h)\|_{H_N}^2$, which completes the proof (18). \square

The characterization (19) of H and the property (20) has been proposed by Parzen in [12]. In [4], Theorem 3.1 p 1587, Bay, Grammont and Maatouk give a proof easy to understand in the framework of this paper.

Proposition 3 (Isometric property of ρ_N) For all $v_N \in H_N$, we have

$$\|\rho_N(v_N)\|_H^2 = c_{v_N}^\top \Gamma_N^{-1} c_{v_N}. \quad (21)$$

The operator ρ_N is an isometry from H_N into H , i.e.

$$\forall v_N \in H_N, \quad \|\rho_N(v_N)\|_H^2 = \|v_N\|_{H_N}^2. \quad (22)$$

For all $h \in H$,

$$\|\rho_N(\pi_N(h)) - h\|_H \xrightarrow{N \rightarrow +\infty} 0. \quad (23)$$

Proof We have

$$\|\rho_N(v_N)\|_H^2 = (\rho_N(v_N), \rho_N(v_N))_H = \sum_{i=0}^N \sum_{j=0}^N \alpha_j \alpha_i (K(\cdot, t_i) K(\cdot, t_j))_H.$$

Since $(K(\cdot, t_i), K(\cdot, t_j))_H = K(t_i, t_j)$, we obtain

$$\|\rho_N(v_N)\|_H^2 = \sum_{i=0}^N \sum_{j=0}^N \alpha_j \alpha_i K(t_{N,i}, t_{N,j}) = \alpha^\top \Gamma_N \alpha.$$

As $\alpha = \Gamma_N^{-1} c_{h_N}$ and Γ_N is symmetric, we obtain (21). We have $v_N = c_{v_N}^\top \varphi(x)$. According to the definition of the inner product in H_N , we have

$$\|v_N\|_{H_N}^2 = (v_N, v_N)_{H_N} = c_{v_N}^\top \Gamma_N^{-1} c_{v_N}.$$

Using (21), we obtain $\|\rho_N(v_N)\|_H^2 = \|v_N\|_{H_N}^2$. We have

$$\begin{aligned} \|\rho_N(\pi_N(h)) - h\|_H &= (\rho_N(\pi_N(h)) - h, \rho_N(\pi_N(h)) - h)_H \\ &= \|\rho_N(\pi_N(h))\|_H^2 + \|h\|_H^2 - 2(h, \rho_N(\pi_N(h))). \end{aligned}$$

Thanks to (5), and according to (13) with $\Lambda = \Gamma_N^{-1}c_h$

$$\begin{aligned} \|\rho_N(\pi_N(h)) - h\|_H &= \|\pi_N(h)\|_{H_N}^2 + \|h\|_H^2 - 2(h, \sum_{i=0}^N \lambda_i K(\cdot, t_i)) \\ &= \|\pi_N(h)\|_{H_N}^2 + \|h\|_H^2 - 2 \sum_{i=0}^N \lambda_i h(t_i) \\ &= c_h^\top \Gamma_N^{-1} c_h + \|h\|_H^2 - 2c_h^\top \Gamma_N^{-1} c_h \\ &= \|h\|_H^2 - c_h^\top \Gamma_N^{-1} c_h \\ &= \|h\|_H^2 - \|\pi_N(h)\|_{H_N}^2. \end{aligned}$$

And (23) comes from (20).

Now, we can formulate the approximation problem : let us define the function $J_N : H_N \rightarrow \mathbb{R}^+$: $\forall v_N \in H_N$,

$$J_N(v_N) := \|v_N\|_{H_N}^2 + \frac{1}{\sigma^2} \|\mathcal{I}_n(v_N) - \mathbf{y}^\top\|_n^2 \quad (24)$$

where $\mathbf{y} = (y_1, \dots, y_n)$ and $\|\cdot\|_n$ the Euclidean norm in \mathbb{R}^n .

Proposition 4 For all u_N and v_N in H_N , we have

$$J_N(tu_N + (1-t)v_N) = tJ_N(u_N) + (1-t)J_N(v_N) - t(1-t) \left(\|u_N - v_N\|_{H_N}^2 + \frac{1}{\sigma^2} \|\mathcal{I}_n(u_N) - \mathcal{I}_n(v_N)\|_n^2 \right). \quad (25)$$

So that J_N is strongly convex :

$$J_N(tu_N + (1-t)v_N) \leq tJ_N(u_N) + (1-t)J_N(v_N) - t(1-t)\|u_N - v_N\|_{H_N}^2. \quad (26)$$

Moreover, J_N is Fréchet differentiable and

$$\lim_{\|v_N\|_{H_N} \rightarrow +\infty} J_N(v_N) = +\infty.$$

Additionally,

$$J_N(v_N) = J(\rho_N(v_N)). \quad (27)$$

Proof Inequality (25) is straightforward.

It is easy to notice that $\mathcal{I}_N(\rho_N(v_N)) = \mathcal{I}_n(v_N)$, so that, as (H3) is satisfied, then

$$\mathcal{I}_n(\rho_N(v_N)) = \mathcal{I}_n(v_N)$$

so that, thanks to (5)

$$\begin{aligned} J(\rho_N(v_N)) &= \|\rho_N(v_N)\|_H^2 + \frac{1}{\sigma^2} \|\mathcal{I}_n(\rho_N(v_N)) - \mathbf{y}^\top\|_n^2 \\ &= \|v_N\|_{H_N}^2 + \frac{1}{\sigma^2} \|\mathcal{I}_n(v_N) - \mathbf{y}^\top\|_n^2 \\ &= J_N(v_N). \end{aligned}$$

□

The constraints space is simply defined as

$$C_N = H_N \cap C. \quad (28)$$

Then the discretized problem is defined by

$$\min_{u_N \in C_N} J_N(u_N). \quad (P_N)$$

Theorem 1 *Under the assumptions (H1) to (H5), (P_N) has a unique solution \hat{u}_N .*

Proof Let $g \in H \cap C$, then, thanks to hypothesis (H4), $\pi_N(g) \in C_N = H_N \cap C$. So that C_N is a nonempty closed convex of H_N . According to the properties of J_N (Proposition 4), we have the conclusion. \square

4 Convergence result

The aim of this paragraph is to prove that, if \hat{u}_N is the solution of (P_N) and \hat{u} the solution of (P), then in E ,

$$\lim_{N \rightarrow +\infty} \hat{u}_N = \hat{u}.$$

We will prove two intermediate results leading to the convergence result.

Proposition 5 *Under (H1) (H2) and (H3)*

$$\lim_{N \rightarrow +\infty} J_N(\pi_N(\hat{u})) = J(\hat{u}); \quad (29)$$

$$\lim_{N \rightarrow +\infty} J_N(\hat{u}_N) = J(\hat{u}). \quad (30)$$

Proof Let us set

$$h^N := \rho_N(\hat{u}_N) \in H.$$

Using (27), as $\pi_N(\hat{u}) \in H_N \cap C$, we have

$$J(h^N) = J(\rho_N(\hat{u}_N)) = J_N(\hat{u}_N) \leq J_N(\pi_N(\hat{u})) \leq J(\hat{u}). \quad (31)$$

Let us prove the last inequality:

$$J_N(\pi_N(\hat{u})) = \|\pi_N(\hat{u})\|_{H_N} + \frac{1}{\sigma^2} \|\mathcal{I}_n(\pi_N(\hat{u})) - y\|_n^2.$$

Thanks to (H3), $\mathcal{I}_n(\pi_N(\hat{u})) = \mathcal{I}_n(\hat{u})$ so that, thanks to (18),

$$J_N(\pi_N(\hat{u})) \leq \|\hat{u}\|_H + \frac{1}{\sigma^2} \|\mathcal{I}_n(\hat{u}) - y\|_n^2 = J(\hat{u}).$$

As $\|h^N\|_H \leq J(h^N) \leq J(\hat{u})$, then the sequence $(h^N)_{N \in \mathbb{N}}$ is bounded in H so that, by weak compactness in Hilbert space, there exists a sub-sequence $(h^{N_k})_{k \in \mathbb{N}}$ and $h^* \in H$ such that

$$h^{N_k} \rightharpoonup_{k \rightarrow +\infty} h^* \in H, \quad (\text{weak convergence}). \quad (32)$$

As H is a RKHS with kernel K , for all $t_i \in \Delta_N$, $K(\cdot, t_i) \in H$ and

$$(h^{N_k}, K(\cdot, t_i))_H = h^{N_k}(t_i) \xrightarrow{k \rightarrow +\infty} (h^*, K(\cdot, t_i))_H = h^*(t_i).$$

Therefore for all $N \geq 1$, $\pi_N(h^{N_k}) \xrightarrow{k \rightarrow +\infty} \pi_N(h^*)$ in the finite-dimensional space H_N . According to assumption (H2), as far as $N_k \geq N$, $\mathcal{I}_N(h^{N_k}) = \mathcal{I}_N(\rho_{N_k}(\hat{u}_{N_k})) = \mathcal{I}_N(\hat{u}_{N_k})$ and

$$\pi_N(h^{N_k}) = \pi_N(\rho_{N_k}(\hat{u}_{N_k})) = \pi_N(\hat{u}_{N_k}),$$

so that

$$\pi_N(\hat{u}_{N_k}) \xrightarrow{k \rightarrow +\infty} \pi_N(h^*) \quad \text{in } H_N.$$

As H_N is an Hilbertian subspace of E (Inequality (16) of Proposition 1.),

$$\pi_N(\hat{u}_{N_k}) \xrightarrow{k \rightarrow +\infty} \pi_N(h^*) \quad \text{in } E.$$

Under (H4), $\pi_N(\hat{u}_{N_k}) \in C$ and C is closed in E , so that for all N ,

$$\pi_N(h^*) \in C.$$

C is closed in E and $\pi_N(h^*) \xrightarrow{N \rightarrow +\infty} h^*$ in E , then

$$h^* \in C \quad \text{and} \quad J(\hat{u}) \leq J(h^*).$$

Then, as J is convex and lower semi continuous and $h^{N_k} \xrightarrow{k \rightarrow +\infty} h^* \in H$, using (31),

$$J(\hat{u}) \leq J(h^*) \leq \varliminf_k J(h^{N_k}) = \varliminf_k J_{N_k}(\hat{u}_{N_k}) \leq \overline{\varliminf}_k J_{N_k}(\hat{u}_{N_k}) \leq J(\hat{u})$$

so that

$$J_{N_k}(\hat{u}_{N_k}) \xrightarrow{k \rightarrow +\infty} J(\hat{u}).$$

(31) implies that

$$J(\hat{u}) \leq J(h^*) \leq \varliminf_k J(h^{N_k}) = \varliminf_k J_{N_k}(\hat{u}_{N_k}) \leq \varliminf_k J_{N_k}(\pi_{N_k}(\hat{u})) \leq \overline{\varliminf}_k J_{N_k}(\pi_{N_k}(\hat{u})) \leq J(\hat{u})$$

so that

$$J_{N_k}(\pi_{N_k}(\hat{u})) \xrightarrow{k \rightarrow +\infty} J(\hat{u}).$$

As the sequences $(J_N(\hat{u}_N))_{N \in \mathbb{N}}$ and $(J_N(\pi_N \hat{u}))_{N \in \mathbb{N}}$ are in the compact set $[0, J(\hat{u})]$, then the results hold. \square

Proposition 6 *We have the following result*

$$\lim_{N \rightarrow +\infty} \|\pi_N(\hat{u}) - \hat{u}_N\|_{H_N} = 0. \quad (33)$$

Proof As J_N is strongly convex (26) and differentiable, then

$$J_N(\pi_N(\hat{u})) - J_N(\hat{u}_N) \geq (J'_N(\hat{u}_N), \pi_N(\hat{u}) - \hat{u}_N)_{H_N} + \|\pi_N(\hat{u}) - \hat{u}_N\|_{H_N}^2,$$

where J'_N denotes the derivative of J_N . As $\pi_N(\hat{u}) \in H_N \cap C$ and \hat{u}_N solves (P_N) ,

$$(J'_N(\hat{u}_N), \pi_N(\hat{u}) - \hat{u}_N)_{H_N} \geq 0,$$

so that

$$\|\pi_N(\hat{u}) - \hat{u}_N\|_{H_N}^2 \leq J_N(\pi_N(\hat{u})) - J_N(\hat{u}_N).$$

(33) comes from the application of Proposition 5. \square

Theorem 2 Under (H1), (H2), (H3) (H4) and (H5)

$$\hat{u}_N \xrightarrow{N \rightarrow +\infty} \hat{u} \quad \text{in } E.$$

Proof

$$\|\hat{u}_N - \hat{u}\|_E \leq \|\hat{u}_N - \pi_N(\hat{u})\|_E + \|\pi_N(\hat{u}) - \hat{u}\|_E.$$

We know from approximation theory in the Banach space E that

$$\|\pi_N(\hat{u}) - \hat{u}\|_E \xrightarrow{N \rightarrow +\infty} 0. \quad (34)$$

As H_N is an Hilbertian subspace of E (see (16)),

$$\|\pi_N(\hat{u}) - \hat{u}\|_E \leq c \|\pi_N(\hat{u}) - \hat{u}\|_{H_N}.$$

Proposition 6 gives the result. \square

5 Stochastic Correspondence of Constrained Optimal Smoothing

The overall goal of the paper is to find a correspondence between the solution \hat{u} of (P) and the posterior distribution $\{U(x) \mid U \in C, Y_i = y_i\}$, $(U(x))_{x \in [0,1]}$ being the GP associated to the covariance function K , the kernel of the RKHS H . The observations are written as

$$Y_i = U(x_i) + \mathcal{E}_i, \quad 1 \leq i \leq n,$$

with $\mathcal{E} = (\mathcal{E}_i)_i$ is a centered Gaussian vector $\mathcal{N}(0, \sigma^2 I)$.

As we defined a finite-dimensional approximation space H_N of H , it is quite natural to construct a finite-dimensional GP U_N to approach U . Using the subdivision (8), we approximate the GP U by the following finite-dimensional GP (see [9] for more details):

$$U_N(x) := \sum_{j=0}^N U(t_j) \varphi_j(x), \quad x \in X. \quad (35)$$

Note that $\xi := (U(t_0), \dots, U(t_N))^T$ is a zero-mean Gaussian vector with covariance matrix $\Gamma_N = (K(t_i, t_j))_{0 \leq i, j \leq N}$, where K is the covariance function of U .

In the following proposition, we prove that the GP U_N is associated with the Hilbertian space H_N defined in Section 3.

Proposition 7 *The process U_N defined by (35) is a centered GP with covariance function :*

$$\forall x', x \in [0, 1], \quad K_N(x, x') = \sum_{i,j=0}^N K(t_i, t_j) \varphi_j(x) \varphi_i(x') = \varphi(x)^\top \Gamma_N \varphi(x'), \quad (36)$$

where $\varphi = (\varphi_0, \dots, \varphi_N)^\top$. If (H5) is satisfied, then the RKHS associated to U_N with the reproducing kernel K_N is

$$H_N := \text{Span}\{\varphi_j, j = 0, \dots, N\}$$

with the scalar product (11).

Proof Obvious.

Theorem 3 *The posterior likelihood function of $\{U_N \mid U_N \in C, Y_i = y_i\}$ is of the form*

$$L_{pos}^N(u) = k_N^{-1} \mathbf{1}_{u \in C \cap H_N} \exp\left(-\frac{1}{2} J_N(u)\right), \quad (37)$$

where k_N is a normalizing constant. Then, the MAP estimator \hat{v}_N as the mode of the posterior distribution $\{U_N \mid U_N \in C, Y_i = y_i\}$ is well defined and is equal to \hat{u}_N solution of (P_N) .

Proof First, remark that the sample paths of U_N are in H_N by construction. Hence, it makes sense to define the density of U_N with respect to the uniform reference measure λ_N on H_N . The density is defined up to a multiplicative constant and to give it an explicit expression, we consider the following linear isomorphism :

$$i : c \in \mathbb{R}^{N+1} \mapsto u := \sum_{j=0}^N c_j \varphi_j \in H_N.$$

We can define the measure λ_N on H_N as the image measure $\lambda_N := i(dc)$, where $dc = dc_0 \times \dots \times dc_N$ is the volume measure in \mathbb{R}^{N+1} . So, if $B \in \mathcal{B}(H_N)$ is a Borelian subset of H_N , we have

$$\lambda_N(B) = \int_{\mathbb{R}^N} \mathbf{1}_{i^{-1}(B)}(c) dc_1 \times \dots \times dc_N.$$

To calculate the probability density function of the GP $U_N = \sum_{j=0}^N U(t_j) \varphi_j$, we write

$$P(U_N \in B) = P\left(\xi \in i^{-1}(B)\right).$$

Using the fact that $\xi = (U(t_0), \dots, U(t_N))^\top$ is a zero-mean Gaussian vector $\mathcal{N}(0, \Gamma_N)$, we have

$$\begin{aligned} P(U_N \in B) &= \int_{\mathbb{R}^N} \mathbf{1}_{i^{-1}(B)}(c) \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}} \exp\left(-\frac{1}{2} c^\top \Gamma_N^{-1} c\right) dc \\ &= \int_{\mathbb{R}^N} \mathbf{1}_B(i(c)) \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}} \exp\left(-\frac{1}{2} \|i(c)\|_N^2\right) dc. \end{aligned}$$

By the transfer formula, we get

$$P(U_N \in B) = \int_{H_N} \mathbf{1}_B(u) \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}} \exp\left(-\frac{1}{2}\|u\|_N^2\right) d\lambda_N(u).$$

Hence, the density of U_N with respect to λ_N is the function

$$f_{\{U_N\}} : u \in H_N \mapsto \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}} \exp\left(-\frac{1}{2}\|u\|_N^2\right).$$

Now let us find the density of the conditional distribution $\{U_N \mid Y_i = y_i\}$.

As $Y_i = U_N(x_i) + \mathcal{E}_i$, where $\mathcal{E} = (\mathcal{E}_i)_i$ is a centered Gaussian vector $\mathcal{N}(0, \sigma^2 I)$, the density of $\{(Y_1, \dots, Y_n) \mid U_N = u\}$ is given by

$$f_{\{(Y_1, \dots, Y_n) \mid U_N = u\}} : (y_1, \dots, y_n) \mapsto \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - u(x_i))^2\right).$$

Let us apply the Bayes principle: the density of the distribution $\{U_N \mid Y_i = y_i\}$ is given by

$$f_{\{U_N \mid Y_i = y_i\}}(u) = \frac{f_{\{(Y_1, \dots, Y_n) \mid U_N = u\}}(y_1, \dots, y_n) \times f_{\{U_N\}}(u)}{f_{\{(Y_1, \dots, Y_n)\}}(y_1, \dots, y_n)}.$$

We consider $f_{\{(Y_1, \dots, Y_n)\}}(y_1, \dots, y_n)$ as a constant of normalization k . Then

$$f_{\{U_N \mid Y_i = y_i\}}(u) = k^{-1} \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - u(x_i))^2 - \frac{1}{2}\|u\|_N^2\right).$$

Setting $k_{N,n} = k^{-1} \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sqrt{2\pi}^N |\Gamma_N|^{1/2}}$,

$$f_{\{U_N \mid Y_i = y_i\}}(u) = k_{N,n} \exp\left(-\frac{1}{2}J_N(u)\right). \quad (38)$$

Let us introduce the inequality constraints described by the set C .

If $B \in \mathcal{B}(H_N)$,

$$P(U_N \in B \mid U_N \in C, Y_i = y_i) = P_{Y_i = y_i}(U_N \in B \mid U_N \in C) = \frac{P_{Y_i = y_i}(U_N \in B \cap C)}{P_{Y_i = y_i}(U_N \in C)},$$

so that

$$P(U_N \in B \mid U_N \in C, Y_i = y_i) = \frac{1}{P_{Y_i = y_i}(U_N \in C)} \int_{H_N} \mathbf{1}_{B \cap C}(u) f_{\{U_N \mid Y_i = y_i\}}(u) d\lambda_N(u)$$

is equivalent to

$$P(U_N \in B \mid U_N \in C, Y_i = y_i) = \frac{k_{N,n}}{P_{Y_i = y_i}(U_N \in C)} \int_{H_N \cap C} \mathbf{1}_B(u) \exp\left(-\frac{1}{2}J_N(u)\right) d\lambda_N(u).$$

In the Bayesian framework, as far as $P_{Y_i=y_i}(U_N \in C) \neq 0$, the density of the posterior conditional distribution $\{U_N \mid U_N \in C, Y_i = y_i\}$ is the following truncated probability density function with respect to λ_N :

$$f_{\{U_N \in B \mid U_N \in C, Y_i = y_i\}}(u) = c \mathbf{1}_{\{u \in H_N \cap C\}} \exp\left(-\frac{1}{2}J_N(u)\right). \quad (39)$$

The density $f_{\{U_N \in B \mid U_N \in C, Y_i = y_i\}}$ is also called the posterior likelihood function denoted by L_{pos}^N . By definition, the MAP estimator \hat{u}_N is the solution of the following optimization problem

$$\arg \max L_{pos}^N(u) = \arg \min \left(-2 \log L_{pos}^N(u)\right).$$

From expression (39), the MAP estimate \hat{v}_N is \hat{u}_N solution of (P_N)

6 Numerical illustration

The aim of this section is to illustrate the correspondence established in previous sections between the MAP estimator and the optimal *constrained* smoothing function solution of problem (P) . We consider the cases where the real function f respects boundedness or monotonicity constraints. The associated convex sets are denoted respectively C_B and C_M and are equal to :

$$C_B = \left\{ f \in \mathcal{C}^0([0, 1]) : -\infty \leq a \leq f(x) \leq b \leq +\infty, x \in [0, 1] \right\}$$

$$C_M = \left\{ f \in \mathcal{C}^0([0, 1]) : f(x) \leq f(x'), \forall x \leq x' \right\}.$$

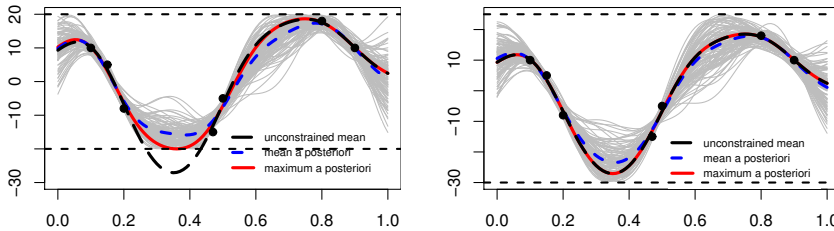


Fig. 1: Unconstrained and constrained mean together with the maximum *a posteriori* (MAP) estimator using the *constrained* model. The lower and upper bounds are equal to -20 and 20 (Left) and equal to -30 and 25 (Right). The black points represent the observations where the noise variance $\sigma^2 = 1.5^2$.

In Figure 1, the constrained data are of size $n = 7$ (black points). The noise variance is fixed at $\sigma^2 = 1.5^2$. The Squared Exponential (or Gaussian) covariance function is used

$$K(x, x') := \nu^2 \exp\left(-\frac{(x - x')^2}{2\theta^2}\right),$$

where the hyper-parameters (θ, ν) are fixed at $(0.12, 10)$. In the left panel of Figure 1, we choose $N = 100$ and generate 100 sample paths (gray solid lines) taken from the finite-dimensional approximation of GPs (35) conditionally to boundedness constraints and noisy data, where the lower and upper bounds are respectively -20 and 20 . The sample paths of the conditional GP (gray solid lines) respect the boundedness constraints in the entire domain unlike the unconstrained mean. In the right panel of Figure 1, we just relax the boundedness constraints such that the unconstrained mean respects it. The lower and upper bounds become -30 and 25 . In that case, the unconstrained mean coincides with the MAP estimator but not with the posterior mean (i.e., the mean of the paths). Hence, in the *constrained* case, the mean of the posterior distribution does not correspond to the optimal *constrained* smoothing function.

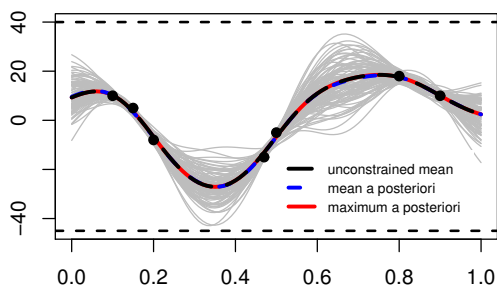


Fig. 2: 100 sample paths taken from the GP (gray solid line) respecting boundedness constraints between -45 and 40 . The unconstrained mean, the mean and the maximum *a posteriori* coincide as expected

In Figure 2, we also relax the boundedness constraints (lower and upper bounds are equal -45 and 40 respectively) such that they do not have an impact on the model. In that case, the unconstrained mean, the mean and the maximum of the posterior distribution coincide as expected.

This numerical example shows that the Maximum A Posteriori estimator is smoother and much more likely than the posterior mean. This is coherent with the theoretical result of this paper since the MAP is the solution of a regularization problem in the RKHS associated to the covariance kernel of the process.

In Figure 3, the monotonicity constraint is considered. We choose $N = 100$ and generate fifty sample paths (gray solid lines) taken from the finite-dimensional approximation of GPs (35) conditionally to monotonicity constraints and noisy data. The observations are generated from the monotone function $x \mapsto e^{x^2}$, where the noise variance is fixed at $\sigma^2 = 0.2^2$. The monotone function is supposed to be evaluated at $(0.05, 0.1, 0.15, 0.75)$ on the left panel and at $(0.05, 0.1, 0.15, 0.75, 1)$ on the right panel. The Matérn $5/2$ covariance function has been used

$$K(x, x') = \nu^2 \left(1 + \frac{\sqrt{5}|x - x'|}{\theta} + \frac{5(x - x')^2}{3\theta^2} \right) \exp \left(-\frac{\sqrt{5}|x - x'|}{\theta} \right),$$

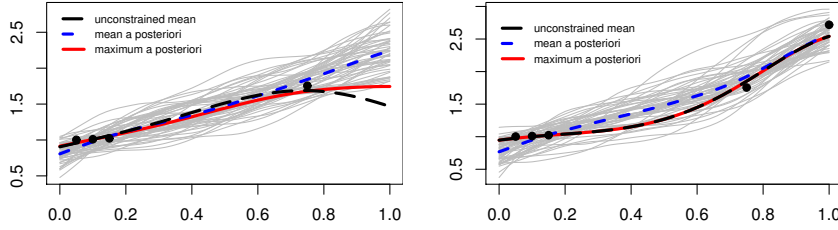


Fig. 3: Unconstrained and constrained mean together with the maximum *a posteriori* (MAP) estimator using the *constrained* model. The unconstrained mean (blue dashed line) respects monotonicity constraints (right panel), contrarily to the left panel. The Matérn 5/2 covariance function has been used. The noise variance is fixed at $\sigma^2 = 0.2^2$

where the hyper-parameters (θ, ν) are fixed at 0.6 and 1 respectively. In the left panel of Figure 3, the GP sample paths (gray lines) respect monotonicity constraints in the entire domain contrarily to the unconstrained mean (blue dashed line). In the right panel of this figure, we just add one observation at $x = 1$. In that case, the unconstrained mean respects monotonicity constraints in the entire domain. We remark that it coincides with the MAP and not with the mean of the posterior distribution. Again, in the *monotonicity* case, the posterior mean (mean of the sample paths) does not correspond to the optimal *constrained* smoothing function as expected.

7 Conclusion and applications

In this paper, we considered the constrained optimal smoothing problem. We approach it by the usual piecewise linear projection to obtain an approximate solution. We proved the convergence of these approximations to the optimal solution. If we rewrite the problem in a GP model, we realized that the obtained approximate solution is also the Maximum A Posteriori of the posterior distribution of the approximate GP. We find the same Bayesian estimation as in the case of strict interpolation treated in [4], but the proof is much easier.

These results are theoretical but they can lead to practical applications. In applications, the constraints are generally positivity, monotonicity or convexity. Numerical illustration are provided with boundedness and monotonicity constraints. Accounting for these type of inequality constraints in GP leads to smaller prediction error and to more realistic uncertainties. The main benefit of our finite-dimensional approximation approach is that it guarantees that the constraints are satisfied everywhere (see [9]). It can be quite easily extended to multidimensional situations. In [7], the constrained GP emulator is implemented, relaxing the interpolation of observations through noise effect. They show, through numerical experiments, that these emulators are faster than in the case of interpolation.

Nevertheless, these models are appropriate to small dimensional case (up to dimension 5). In [3], the authors overcome this limitation with the introduction of the MaxMod algorithm, performing at the same time dimension reduction and efficient knot allocation. With simulated and real data, this algorithm remains efficient in high dimension (at least dimension 20). In [3], the data points x_i are not in the partition Δ_N . Our results are also valid in that case. Its proof is done in the appendix here below.

8 Appendix

If assumption (H3) is not fulfilled, then $\rho_N(v_N)$ and v_N are not the same at the data points, so that $J(\rho_N(v_N)) \neq J_N(v_N)$. Hence Proposition 4, except (27) and Theorem 1 are valid. However Proposition 5 needs to be proved when (H3) is not fulfilled. The proof is just a bit more technical.

As the norm on E is the infinity norm and π_N is stable, according to (6) and (16)

$$\begin{aligned} |\pi_N(\hat{u})(x_i)| &\leq \|\pi_N(\hat{u})\|_E \leq d\|\pi_N(\hat{u})\|_{H_N} \leq d\|\hat{u}\|_H \\ |\hat{u}(x_i)| &\leq \|\hat{u}\|_E \leq c\|\hat{u}\|_H. \end{aligned}$$

Hence

$$\begin{aligned} J_N(\pi_N(\hat{u})) &= \|\pi_N(\hat{u})\|_{H_N}^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\pi_N(\hat{u})(x_i) - y_i)^2 \\ &\leq \|\hat{u}\|_H^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\pi_N(\hat{u})(x_i) - y_i)^2 \\ &\leq J(\hat{u}) + \frac{1}{\sigma^2} \sum_{i=1}^n \left[(\pi_N(\hat{u})(x_i) - y_i)^2 - (\hat{u}(x_i) - y_i)^2 \right] \\ &\leq J(\hat{u}) + \frac{1}{\sigma^2} \sum_{i=1}^n [\pi_N(\hat{u})(x_i) - \hat{u}(x_i)] [\hat{u}(x_i) + \pi_N(\hat{u})(x_i) - 2y_i] \\ J_N(\pi_N(\hat{u})) &\leq J(\hat{u}) + \underbrace{\frac{n((c+d)\|\hat{u}\|_E + 2\max|y_i|)}{\sigma^2}}_{c_1} \|\pi_N(\hat{u}) - \hat{u}\|_E. \end{aligned}$$

Thus, setting $\eta_N = \|\pi_N(\hat{u}) - \hat{u}\|_E$, according to (34), we have $\eta_N \rightarrow 0$ as $N \rightarrow +\infty$.

$$J_N(\pi_N(\hat{u})) \leq J(\hat{u}) + c_1\eta_N,$$

where the constant c_1 is independent of N . By definition of \hat{u}_N , $J_N(\hat{u}_N) \leq J_N(\pi_N(\hat{u}))$ so that

$$J_N(\hat{u}_N) \leq J(\hat{u}) + c_1\eta_N.$$

We have

$$|\rho_N(\hat{u}_N)(x_i)| \leq \|\rho_N(\hat{u}_N)\|_E \leq c\|\rho_N(\hat{u}_N)\|_H = c\|\hat{u}_N\|_{H_N} \leq cJ_N(\hat{u}_N) \leq cJ(\hat{u}) + cc_1\eta_N$$

$$|\hat{u}_N(x_i)| \leq \|\hat{u}_N\|_E \leq d\|\hat{u}_N\|_{H_N} \leq dJ_N(\hat{u}_N) \leq dJ(\hat{u}) + dc_1\eta_N.$$

Let c_2 be such that, for all N ,

$$(c + d)(J(\hat{u}) + c_1\eta_N) \leq c_2.$$

We have, according to the isometric property of ρ_N ,

$$\begin{aligned} J(\rho_N(\hat{u}_N)) &= \|\rho_N(\hat{u}_N)\|_H^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\rho_N(\hat{u}_N)(x_i) - y_i)^2 \\ &= \|\hat{u}_N\|_{H_N}^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\rho_N(\hat{u}_N)(x_i) - y_i)^2 \\ &= J_N(\hat{u}_N) - \frac{1}{\sigma^2} \sum_{i=1}^n ((\hat{u}_N)(x_i) - y_i)^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (\rho_N(\hat{u}_N)(x_i) - y_i)^2 \\ &= J_N(\hat{u}_N) + \frac{1}{\sigma^2} \sum_{i=1}^n [(\rho_N(\hat{u}_N)(x_i) - y_i)^2 - (\hat{u}_N(x_i) - y_i)^2] \\ &= J_N(\hat{u}_N) + \frac{1}{\sigma^2} \sum_{i=1}^n [\rho_N(\hat{u}_N)(x_i) + \hat{u}_N(x_i) - 2y_i] [\rho_N(\hat{u}_N)(x_i) - \hat{u}_N(x_i)] \\ &\leq J_N(\hat{u}_N) + \frac{c_2 + 2 \max |y_i|}{\sigma^2} \sum_{i=1}^n [\rho_N(\hat{u}_N)(x_i) - \hat{u}_N(x_i)]. \end{aligned}$$

Thanks to the reproducing properties of H and H_N and to the isometric property of ρ_N :

$$\begin{aligned} |\rho_N(\hat{u}_N)(x_i) - \hat{u}_N(x_i)| &= (\rho_N(\hat{u}_N), K(\cdot, x_i))_H - (\hat{u}_N, K_N(\cdot, x_i))_{H_N} = (\rho_N(\hat{u}_N), K(\cdot, x_i))_H - (\rho_N(\hat{u}_N), \rho_N(K_N(\cdot, x_i)))_H \\ &= (\rho_N(\hat{u}_N), K(\cdot, x_i) - \rho_N(K_N(\cdot, x_i)))_H \\ &\leq \|\rho_N(\hat{u}_N)\|_H \|K(\cdot, x_i) - \rho_N(K_N(\cdot, x_i))\|_H \\ &\leq (J(\hat{u}) + c_1\eta_N) \max_i \|K(\cdot, x_i) - \rho_N(K_N(\cdot, x_i))\|_H. \end{aligned}$$

Lemma 1 *We have*

$$\sup_{x \in X} \|\rho_N(K_N(\cdot, x)) - K(\cdot, x)\|_H \xrightarrow{N \rightarrow +\infty} 0.$$

Proof Thanks to the isometric property of ρ_N , we have

$$\begin{aligned} \|\rho_N(K_N(\cdot, x)) - K(\cdot, x)\|_H^2 &= \|\rho_N(K_N(\cdot, x))\|_H^2 + \|K(\cdot, x)\|_H^2 - 2(\rho_N(K_N(\cdot, x)), K(\cdot, x))_H \\ &= \|K_N(\cdot, x)\|_{H_N}^2 + \|K(\cdot, x)\|_H^2 - 2 \sum_{j=0}^N \varphi_j(x) K(x, t_j) \\ &= K_N(x, x) + K(x, x) - 2 \sum_{j=0}^N \varphi_j(x) K(x, t_j). \end{aligned}$$

By uniform continuity of $K(\cdot, \cdot)$ on the compact set $X \times X$, we deduce that both $K_N(x, x) = \sum_{i,j=0}^N K(t_i, t_j) \varphi_i(x) \varphi_j(x)$ and $\sum_{j=0}^N \varphi_j(x) K(x, t_j)$ are uniformly convergent to the function $K(x, x)$, which completes the proof of the lemma.

Let us set

$$\varepsilon_N = \max_i \|K(\cdot, x_i) - \rho_N(K_N(\cdot, x_i))\|_H.$$

According to the lemma 1,

$$\varepsilon_N \rightarrow 0 \text{ as } N \rightarrow +\infty$$

and

$$J(\rho_N(\hat{u}_N)) \leq J_N(\hat{u}_N) + c_3 \varepsilon_N$$

with

$$c_3 = \frac{nc_2 + 2n \max y_i}{\sigma^2}.$$

Without assumption (H3), setting $h^N = \rho_N(\hat{u}_N)$ (31) is replaced by

$$J(h^N) \leq J_N(\hat{u}_N) + c_3 \varepsilon_N \leq J_N(\pi_N(\hat{u})) + c_3 \varepsilon_N \leq J(\hat{u}) + c_1 \eta_N + c_3 \varepsilon_N, \quad (40)$$

where $\varepsilon_N \rightarrow 0$ and $\eta_N \rightarrow 0$ as $N \rightarrow +\infty$. Then, the proof is almost the same than in Section 4: (40) is used instead of (31) and the compact set is $[0, J(\hat{u}) + \sup_N(c_1 \eta_N + c_3 \varepsilon_N)]$.

References

1. Aronszajn, N., Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68, (1950)
2. Andersson, L.E. and Elfving, T., Interpolation and approximation by monotone cubic splines, *Journal of Approximation Theory* vol 66, number 3, pages 302-333, (1991)
3. Bachoc, F., López Lopera, A.F. and Roustant, O., Sequential construction and dimension reduction of Gaussian processes under inequality constraints *arXiv:2009.04188v1 [math.ST] 9Sep 2020*
4. Bay, X., Grammont, L. and Maatouk, H., Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation *Electronic Journal of Statistics* Vol 10, pages 1580-1595, (2016)
5. Bay, X., Grammont, L. and Maatouk, H., A new method for interpolating in a convex subset of a Hilbert space *Computational Optimization and Applications* Vol 68 (1), pages 95-120 (2017)
6. Dontchev, A. L and Qi, H. and Qi, L., Convergence of Newton's method for convex best interpolation, *Numerische Mathematik*, vol 87, number 3, pages 435-456, (2001)
7. López-Lopera, A.F.; Bachoc, F.; Durrande, N.; Rohmer, J.; Idier, D. and Roustant, O., Approximating Gaussian Process emulators with linear inequality constraints and noisy observations via MC and MCMC, *arXiv:1901.04827v2 [stat.ML] 21 Jun (2019)*
8. Maatouk, H., Finite-dimensional approximation of Gaussian processes with linear inequality constraints and noisy observations, *Communications in Statistics - Theory and Methods*, doi=10.1080/03610926.2022.2055768 (2022)
9. Maatouk, H. and Bay, X., Gaussian process emulators for computer experiments with inequality constraints, *Mathematical Geosciences* vol 49, pages 557582, (2017)
10. Micchelli, C. and Utreras, F., Smoothing and Interpolation in a Convex Subset of a Hilbert Space, *SIAM Journal on Scientific and Statistical Computing*, vol 9, number 4, pages 728-746, (1988)
11. Kimeldorf, G. S and Wahba, G., A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics*, pages 495-502, (1970)
12. Parzen, E., Statistical inference on time series by Hilbert space methods. *Stanford University*, (1959)
13. Schwartz, L., Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (Noyaux reproduisants), *Journal d'Analyse Mathématique* volume 13, pages 115256 (1964)