A non-linear manifold reduction based on ridge regression

Damiano Lombardi

A non-linear manifold reduction based on ridge regression

Damiano Lombardi[1]

[1]COMMEDIA, Laboratoire Jacques–Louis Lions, Sorbonne Université et Inria Paris, 2, rue Simone Iff, 75012, Paris.

Abstract

The present work deals with non-linear manifold reduction. A database of snapshots is represented by introducing an encoder-decoder pair. In the spirit of quadratic manifold reduction, we consider a linear encoder. The reconstruction of the snapshots (decoder) is then performed by using a dof-wise regression based on ridge functions. The non-linear function for each of the ridge regressions is not fixed a priori. Instead, it is constructively computed by using $\mathbb{P}_3-$Hermite finite elements. In order to promote sparsity and enhance the decoder performances (both in terms of memory burden and number of elementary operations needed to apply it) we define a greedy algorithm which progressively adapt the dof-wise regression until a prescribed accuracy is fulfilled. A theoretical analysis is presented as well as several numerical experiments in order to assess the performances of the proposed reduced representation.

1 Introduction

Classical projection-based Model Order Reduction relies on the construction of a linear (possible small dimensional) subspace, whose basis makes it possible to devise numerical methods which enable fast integration of some problems of interest. An overview of the field is presented in [4]. From an approximation theory point of view, the ability of approximating a set of solutions by means of the elements of a linear subspace is quantified by the notion of Kolmogorov widths [8]. This sequence of non-increasing positive numbers, indexed by $n \in \mathbb{N}^*$, quantifies the worst case scenario error (in a certain norm) when approximating at best the elements of a set by the elements of a linear subspace (the one delivering the best possible approximation) of dimension n. If the Kolmogorov widths of a given set, think for instance to the solutions of a parametrised Partial Differential Equation (PDE) problem, decay slowly, it implies that in order to guarantee a prescribed accuracy, we need the dimension of the subspace n to be large enough. In classical model reduction this translates in a degraded

speedup, as the ability to solve a problem in a fast way relies essentially on the fact that n is significantly smaller than the number of degrees of freedom of the classical discretisation of the problem. The fact that certain sets of parametrised PDE exhibit a slow decay is nowadays referred to as Kolmogorov barrier.

This limitation can be overcome by using notions of non-linear approximation [11, 5] and introducing methods which implement them. Numerous methods of non-linear reduction have been proposed in recent years, tailored to the situations (like advection-dominated phenomena, travelling waves and front propagation) in which the Kolmogorov barrier manifests. In view of describing possible different ways to approximate the elements of a given set, a useful paradigm consists in seeing the approximation as the composition of an encoder and a decoder. The analysis of this setup and the introduction of stable non-linear widths is proposed in [9]. The reduced basis method [4] can be interpreted in this way: the encoder and the decoder are both linear. Examples of non-linear reduction based on autoencoders (both the encoder and the decoder are non-linear) can be found in [15, 12].

In the work [2] a quadratic manifold reduction is proposed. In this, the encoder is linear, and the decoder is non-linear (quadratic). Another contribution on quadratic manifold reduction is proposed in [14]. In quadratic manifold reduction we proceed as follows. First, a set of solutions of the Full Order Model (FOM) is computed for several instances of the parameters, leading to the construction of a database of solutions (the snapshots). Let $N \in \mathbb{N}$ be the number of snapshots of the FOM, denoted by $\left\{u^{(j)}\right\}_{1\leq j\leq N} \in V$. In a continuous setting, V is typically a Sobolev space; when considering standard discretisations of the FOM, $V \subseteq \mathbb{R}^{\mathcal{N}}$, where $\mathcal{N} \in \mathbb{N}$ is the number of the degrees of freedom of the FOM.

After we have computed the snapshots database, we perform the following steps:

1. We compute a reduced basis, for instance by using Proper Orthogonal Decomposition (POD). Let $n \in \mathbb{N}$ and the basis elements be denoted by $\{\varphi_i\}_{1\leq i\leq n} \in V$. They are a basis of a linear subspace $V_n \subset V$. Let the scalar product used to compute the POD be denoted by $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$. For instance, when we consider the snapshots of the discretised FOM, a classical choice consists in using the ℓ^2 scalar product. The modes satisfy: $\langle \varphi_i, \varphi_j \rangle = \delta_{ij}$. The set of reduced coordinates $\left\{a^{(j)}\right\}_{1\leq j\leq N} \in \mathbb{R}^n$ are obtained by:

$$a_i^{(j)} = \langle u^{(j)}, \varphi_i \rangle, \quad i = 1, \ldots, n, \ 1 \leq j \leq N.$$

   This can be interpreted as the application of a linear encoder.

2. A classical linear reconstruction of the $j-$th snapshot would be:

$$\tilde{u}^{(j)} = \sum_{i=1}^{d} a_i^{(j)} \varphi_i.$$

   This can be seen as a linear application $\mathcal{L} : \mathbb{R}^n \to V$, mapping the reduced coordinate to the space V. We introduce then a quadratic correction aiming at

improving the reconstruction. Let $a^{(j)} \otimes a^{(j)} \in \mathbb{R}^{n\times n}$ be the Kronecker product between the reduced coordinates with themselves:

$$[a^{(j)} \otimes a^{(j)}]_{lm} = a_l^{(j)} a_m^{(j)}, \quad 1 \leq l, m \leq n.$$

We define the functions $\{\psi_{lm}\}\, 1 \leq l, m \leq n \in V$ such that the corrected reconstruction reads:

$$\tilde{u}^{(j)} = \sum_{i=1}^{d} a_i^{(j)} \varphi_i + \sum_{l=1}^{n}\sum_{m=1}^{n} \psi_{lm} a_l^{(j)} a_m^{(j)}.$$

For symmetry reasons, we can look for $n(n+1)/2$ functions (instead of n^2) such that the error between the snapshots and their quadratically corrected reconstruction is minimised:

$$\{\psi_{lm}\}_{1\leq l,m\leq n} = \arg \inf_{\{\psi_{lm}\}\in V} \sum_{j=1}^{n} \|u^{(j)} - \tilde{u}^{(j)}(\psi_{lm})\|_V^2.$$

Remark that the quadratic regression problem to compute the modes ψ makes it possible to define a quadratic decoder: $\mathcal{Q} : \mathbb{R}^n \to V$, mapping the reduced coefficients $a^{(j)} \in \mathbb{R}^n$ into an approximation of the snapshots. The composition of encoder and decoder reads:

$$\tilde{u} = \mathcal{Q}(\langle u, \varphi \rangle).$$

A polynomial setting for the decoder is considered in [13], analysed in [7]. A decoder based on neural networks is proposed in [3] and an analysis and several possible ways of defining a non-linear decoder are presented in [10].

In the present work we will focus on the *offline phase* and investigate a method in which the encoder is linear, the decoder is non-linear and it is constructed by considering the family of ridge functions. We introduce a greedy method which constructs a regression on the degrees of freedom of the FOM, resulting in a dof-wise certified projection pursuit.

The structure of the work is as follows. In Section 2 the method is introduced: first, we discuss the methodological choices and then detail the resulting greedy algorithm and the finite elements adaptive regressions. In Section 3 some theoretical results are presented: we will relate the error of the non-linear manifold reduction to the one of a classical linear decoder (Proper Orthogonal Decomposition) and propose some estimation of the memory burden. In Section 4 several numerical experiments are proposed, in 1d, 2d, and 3d settings.

2 The method.

Let $V \subseteq \mathbb{R}^{\mathcal{N}}$. The $k-$th component of the vector $u^{(j)}$, denoted $u_k^{(j)}$, represents the value of the $k-$th degree of freedom of the solution. The main goal is to construct a dof-wise regression and determine a set of functions $\{\sigma_k\}_{1\leq k\leq \mathcal{N}} : \mathbb{R}^n \to \mathbb{R}$ which

approximate the set of the degrees of freedom up to a prescribed (global) accuracy $\varepsilon > 0$. Let $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_{\mathcal{N}}]$ be a vector valued function whose components are the functions σ_k. Given the reduced coordinates of the generic $j-$th snapshot, $a^{(j)}$, the evaluation of $\boldsymbol{\sigma}$ leads to the approximation of the discretised solution:

$$\tilde{u}^{(j)} = \boldsymbol{\sigma}(a^{(j)}).$$

This is the decoder we are going to build. The error in the reconstruction of the $j-th$ snapshot reads:

$$e^{(j)} = u^{(j)} - \tilde{u}^{(j)}.$$

When measuring the norm of the error on the degrees of freedom in $\ell^{q,\mathcal{N}}$ norm, the total error made to reconstruct N snapshots is:

$$e = \sum_{j=1}^{N} \|u^{(j)} - \tilde{u}^{(j)}\|_{\ell^{q,\mathcal{N}}} \leq \varepsilon.$$

In the case in which we use the $\ell^{2,\mathcal{N}}$ norm this reduces to:

$$e^2 = \sum_{j=1}^{N} \sum_{k=1}^{\mathcal{N}} (u_k^{(j)} - \sigma_k(a^{(j)}))^2 \leq \varepsilon^2. \tag{1}$$

In the rest of the paper we will use this norm in order to quantify the reconstruction error. Remark that this is an empirical error based on the database of FOM solutions.

The following notation is introduced:

- $\mathcal{E}_{\mathcal{L},n}$ is the norm of the $\ell^{2,\mathcal{N}}$ error in the snapshots reconstruction, made when using the classical linear reconstruction with $n \in \mathbb{N}$ modes.
- $\mathcal{E}_{\mathcal{R},n}$ is the norm of the $\ell^{2,\mathcal{N}}$ error in the snapshots reconstruction, made when using the method proposed in the present work, based on reduced coordinates $a \in \mathbb{R}^n$.

Two points are crucial in view of defining the method:

(I) **Error distribution.** The problem of finding the functions $\{\sigma_k\}_{1 \leq k \leq \mathcal{N}}$ amounts to solving $\mathcal{N}$ regression problems. The only constraint potentially coupling them is that the global error has to be smaller than a prescribed accuracy. The fact that the problems can be made independent or not depends solely on the way in which we decide to satisfy the constraint on the prescribed accuracy. When deciding to uniformly split the error among all degrees of freedom, or deciding *a priori* how to distribute it according to some importance criterion, we make all the $\mathcal{N}$ regression problems independent. This has the clear advantage of being embarrassingly parallel. However, it is not necessarily optimal in terms of parsimony, *i.e.* the memory that we use to encode the database and to store the decoder. In the present work we try to propose and investigate a parsimonious strategy and henceforth we do not distribute a priori the error in the different

degrees of freedom regressions. Instead, we adopt a greedy strategy, which makes the expressions of the functions $\{\sigma_k\}_{1\leq k\leq \mathcal{N}}$ to progressively evolve until the error criterion is satisfied.

(II) **From linear reconstruction to ridge regression.** Another fundamental point is the following: finding a function $\sigma_k : \mathbb{R}^n \to \mathbb{R}$ is potentially a very hard problem, as the function domain is in $\mathbb{R}^n$. There are several ways in which we can try to mitigate or circumvent this difficulty based on the nature of the regression problem we have to solve. In the present context, we proceed as done in quadratic manifold reduction: we choose a class of functions we are going to build the decoder in. As for the quadratic manifold reduction, this can be interpreted as adding a correction to the linear reconstruction. As we will show hereafter, one of the natural ways to do that is through a ridge regression. Let us consider the classical linear reconstruction by of the $k-$th degree of freedom value:

$$\tilde{u}_k^{(j)} = \sum_{i=1}^{n} [\varphi_i]_k a_i^{(j)}.$$

Let $\left\{\xi^{(k)}\right\}_{1\leq k\leq \mathcal{N}} \in \mathbb{R}^n$ be the set of vectors whose components are defined as:

$$\xi_i^{(k)} = [\varphi_i]_k, \quad 1 \leq i \leq n.$$

The reconstruction of the $k-$th degree of freedom reads henceforth:

$$\tilde{u}_k^{(j)} = [\xi^{(k)}]^T a^{(j)},$$

the ℓ^2 scalar product between the $k-$th vector ξ and the reduced coordinates. We will look for an improvement in the reconstruction in the form:

$$\tilde{u}_k^{(j)} = \sigma_k \left([\xi^{(k)}]^T a^{(j)} \right),$$

which is a ridge regression with ridge $\xi^{(k)}$. The function σ_k will be computed by means of a $\mathbb{P}_3$-Hermite adaptive finite elements method, detailed in Section 2.2. Let us remark that the identity $I_k : [\xi^{(k)}]^T a^{(j)} \mapsto [\xi^{(k)}]^T a^{(j)}$ reduces to the linear reconstruction, and the identity is exactly represented by the $\mathbb{P}_3$-Hermite finite elements. Henceforth, in terms of error, we are guaranteed that we cannot do worse than the linear reconstruction.

These two methodological choices leads to the definition of a greedy method, which is detailed in Section 2.1.

2.1 A greedy method to build a constructive non-linear decoder.

The first stage of the method, as for the quadratic manifold reduction presented in [2], is the computation of the classical POD basis. This makes it possible to compute

the modes $\varphi_i \in V$, the set of vectors $\left\{\xi^{(k)}\right\}_{1\leq k\leq \mathcal{N}} \in \mathbb{R}^n$, and the reduced coordinates associated to the snapshots in the database $\left\{a^{(j)}\right\}_{1\leq j\leq N}$.

We compute $n_* \in \mathbb{N}$, the number of POD modes such that the linear reconstruction satisfies the error criterion:

$$n_* = \arg\min_{n\in\mathbb{N}} \{n \text{ such that } : e_{\mathcal{L},n} \leq \varepsilon\}$$

The method consists in assessing the parsimony of the set of the reconstructions we obtain by ridge regression satisfying $\{\mathcal{E}_{\mathcal{R},n}\}_{1\leq n\leq n_*} \leq \varepsilon$. Let us remark that for $n = n_*$ we can simply use the linear reconstruction to satisfy the error constraint and that the tests varying n are all independent, and, hence, can be performed in parallel.

In the present work, we consider the *best reconstruction* the one that, while respecting the prescribed accuracy, requires the least amount of memory to compute the reconstruction of the whole database. We are going to denote $\mathcal{C}$ the memory cost (the amount of double precision numbers we need to store). The classical linear reconstruction, which is the particular case $n = n_*$ has a cost: $\mathcal{C}_{\mathcal{L},n_*} = (\mathcal{N} + N)n_*$. Here we consider the cost of storing the reconstruction of the database, hence we included the Nn_* reduced coordinates. When considering a generic ridge regression the cost to reconstruct the $k-$th degree of freedom is given by the cost of the ridge $n + Nn$ and the cost of the $\mathbb{P}_3$ finite element representation of the non-linearity σ_k. This is denoted $c_k^{\mathbb{P}_3}$ and depends on the number of finite elements degrees of freedom required. The total cost reads:

$$\mathcal{C}_{\mathcal{R},n} = (\mathcal{N} + N)n + \sum_{k=1}^{\mathcal{N}} c_k^{\mathbb{P}_3}.$$

This is simply the cost a linear reconstruction would have if we used $n < n_*$ modes plus the additional cost of the representation of the non-linearities. It is advantageous if this extra cost is compensated by the fact that we need less modes, *i.e.* if:

$$\sum_{k=1}^{\mathcal{N}} c_k^{\mathbb{P}_3} < (\mathcal{N} + N)(n_* - n).$$

We denote the reconstruction by ridge regression based on $n < n_*$ modes as $\mathcal{R}_n$. This is computed by means of a greedy method. Let the iteration of the greedy be denoted by $l \in \mathbb{N}$. At the beginning we set $l = 0$ and set all the functions σ_k equal to the identity: $\sigma_k^{(0)} = I_k$. This means that all the functions are defined by considering one $\mathbb{P}_3$ Hermite finite element.

The error is clearly not matching the prescribed accuracy as $n < n_*$. We compute k_*, the index of the degree of freedom which is contributing the most to the error reconstruction in ℓ^2 norm:

$$E_k^2 = \sum_{j=1}^{N} \left(u_k^{(j)} - \sigma_k^{(l)}(a^{(j)})\right)^2, \tag{2}$$

$$k_* = \arg\max_{1\leq k\leq \mathcal{N}} E_k^2.$$

To improve the error, we add one finite element to the definition of $\sigma_{k_*}^{(l)}$ and determine the values of the $\mathbb{P}_3$ finite elements degrees of freedom as detailed in Section 2.2. This defines a function $\sigma_{k_*}^{(l+1)}$. We set:

$$\boldsymbol{\sigma}^{(l+1)} = \begin{bmatrix} \sigma_1^{(l)} \\ \dots \\ \sigma_{k_*}^{(l+1)} \\ \dots \\ \sigma_{\mathcal{N}}^{(l)} \end{bmatrix} \tag{3}$$

The $l-$th greedy iteration finishes by recomputing the total error and checking wether it is smaller than the prescribed accuracy. If it is, we stop, otherwise we perform a new iteration. It might happen that, for a given n the reconstruction error stagnates, at a value which is larger than the prescribed accuracy ε. As an illustrative example, consider the case $n = 1$. We could have different snapshots (corresponding to two different values of the parameter μ) with the same value of a_1. The map associating to a_1 the values of the degrees of freedom u_k could be featured by a large error. For this reason it may happen that the regression error $e_{\mathcal{R},1}$ stagnates before reaching the prescribed accuracy. When the approximation stagnates, we set: $\mathcal{C}_{\mathcal{R},n} = +\infty$. Remark that there exists at least one n for which we reach the prescribed accuracy: for $n = n_*$, by construction, the linear reconstruction satisfies the error criterion.

Among the reconstructions which fulfil the error criterion, we are going to choose the one which has the smallest memory burden. The best reconstruction is $\mathcal{R}_{j_*}$, where:

$$j_* = \arg \min_{1 \leq n \leq n_*} \mathcal{C}_{\mathcal{R},n}.$$

This is synthetically described in Algorithm 1.

Algorithm: ***Decoder construction.***
Data: $\left\{u^{(j)}\right\}_{1\leq j\leq N}\in\mathbb{R}^{\mathcal{N}}$, $\varepsilon>0$, n_*
Result: The functions $\{\sigma_k^*\}_{1\leq k\leq\mathcal{N}}$.
Initialise: $\mathcal{C}^*=\mathcal{C}_{\mathcal{L},n_*}$, ;
$\{\sigma_k^*\}_{1\leq k\leq\mathcal{N}}=\eta_k$;
for $n=1$; $n<n_*$ **do**
  Reduced coordinates: $\left\{a^{(j)}\right\}_{1\leq j\leq N}\in\mathbb{R}^n$;
  Ridges: $\left\{\xi^{(k)}\right\}_{1\leq k\leq\mathcal{N}}\in\mathbb{R}^n$;
  while *($\mathcal{E}_{\mathcal{R},n}>\varepsilon$) and (**not** stagnation)* **do**
    Find $k_*=\arg\max_{1\leq k\leq\mathcal{N}}E_k^2$, Eq.(2) ;
    Refine the approximation of σ_{k_*} ;
    Compute $\mathcal{E}_{\mathcal{R},n}$;
  end
  Compute $\mathcal{C}_{\mathcal{R},n}$;
  if *($\mathcal{C}_{\mathcal{R},n}<\mathcal{C}^*$) and (**not** stagnation)* **then**
    Set $\mathcal{C}^*=\mathcal{C}_{\mathcal{R},n}$;
    Set $\{\sigma_k^*\}_{1\leq k\leq\mathcal{N}}=\{\sigma_k\}_{1\leq k\leq\mathcal{N}}$;
  end
end

Algorithm 1: Greedy algorithm detailed in Section 2.1.

2.2 $\mathbb{P}_3$-Hermite adaptive finite element approximation.

In this section we detail the computation of the functions σ_k, the $1-d$ regression we perform dof-wise by means of the $\mathbb{P}_3$ Hermite finite elements. In all this section we present the computation for the generic $k-$th degree of freedom. It is intended that $1\leq k\leq\mathcal{N}$.

Let $\eta_k\in\Omega_k\subset\mathbb{R}$ denote the real coordinate obtained by taking the scalar product between $\xi^{(k)}$ and a:

$$\eta_k=[\xi^{(k)}]^T a.$$

The function σ_k is defined as:

$$\sigma_k:\left\{\begin{array}{rcl}\Omega_k&\to&\mathbb{R}\\ \eta_k&\mapsto&\sigma_k(\eta_k)\end{array}\right. \tag{4}$$

The interval $\Omega^{(k)}$ is non-uniformly subdivided into a certain number of elements, each of which has 4 degrees of freedom. We give a piece-wise cubic representation of the function σ_k. There is a number of advantages motivating the use of this type of finite element. As already stated, the identity is exactly represented, hence making it possible to include the linear reconstruction as a particular case. This is important, as it ensures that, in the worst case scenario, we recover the linear approximation. Moreover, the $\mathbb{P}_3$-Hermite finite element discretisation provides $\sigma_k\in\mathcal{C}^1(\Omega_k)$. This can be useful in

view of using the reduced-order representation of the function in the online phase, or in optimisation problems, or in all tasks involving differentiation.

Let $N_k \in \mathbb{N}$ be the number of points in which we subdivide the interval $\Omega^{(k)}$, and let these points be $\left\{\hat{\eta}_k^{(q)}\right\}_{1\leq q\leq N_k}$. Taking the classical expression for $\mathbb{P}_3$-Hermite elements in $1d$ leads to a number of basis functions equal to $2N_k$. Let the $\mathbb{P}_3$-Hermite finite element basis functions be $\{w_i\}_{1\leq i\leq 2N_k}$ and the $q-$th degree of freedom value be $S_{kq} \in \mathbb{R}$. The function $\sigma^{(k)}$ is defined as:

$$\sigma^{(k)}(\eta_k) = \sum_{q=1}^{2N_k} S_{kq} w_q(\eta_k).$$

Two aspects are detailed in this section: first, how the regression is performed and, second, how the mesh is adapted.

Finite elements regression.

In terms of regression, we proceed in a classical way. Let $\left\{a^{(j)}\right\}_{1\leq j\leq N}$ be the reduced coordinates of the database snapshots. These, when taking the scalar product with the vector $\xi^{(k)}$ give us a set of points $\left\{\eta_k^{(j)}\right\}_{1\leq j\leq N} \in \Omega_k$. The residual of the least square problem is then a vector $r^{(k)} \in \mathbb{R}^N$ whose components read:

$$r_j^{(k)} = u_k^{(j)} - \sum_{q=1}^{2N_k} S_{kq} w_q(\eta_k^{(j)}), \quad 1 \leq j \leq N.$$

The least square problem consists in determining the coefficients $\{S_{kq}\}_{1\leq q\leq 2N_k}$ minimising the $\ell^{2,N}$ norm of the residual:

$$S^* = \arg \inf_{s\in\mathbb{R}^{2N_k}} \|r^{(k)}(s)\|^2_{\ell^{2,N}}.$$

To solve the problem we compute the collocation matrix $W^{(k)} \in \mathbb{R}^{N\times 2N_k}$. We have, typically, that $2N_k < N$. The matrix entries are given by:

$$W_{jq}^{(k)} = w_q(\eta_k^{(j)}), \quad 1 \leq j \leq N, \ 1 \leq q \leq 2N_k.$$

Let $\hat{u} \in \mathbb{R}^N$ be the vector whose components are:

$$\hat{u}_j = u_k^{(j)}, \ 1 \leq j \leq N.$$

We perform a QR decomposition of the collocation matrix ($W^{(k)} = QR$) and we solve the least square problem:

$$S^* = R^{-1}Q^T\hat{u}.$$

In terms of implementation, it is important to solve the least square problem in a robust way. This is why we make use of a rank-revealing QR decomposition, to deal with linear dependent (or almost linear-dependent) columns of the collocation matrix.

Mesh adaptation.

In order to perform the mesh adaptation we proceed by splitting in two equal parts the elements in which the residual is the largest. First, given a mesh of Ω_k and the residual vector $r^{(k)} \in \mathbb{R}^N$, we find the element (among the $N_k - 1$ elements) in which the residual is the largest. Let $\left\{\hat{E}_p\right\}_{1\leq p\leq N_k-1}$ be the element-wise errors defined as:

$$\hat{E}_p^2 = \sum_{j=1}^{N} (r_j^{(k)})^2 \mathbb{1}_{jp},\ 1 \leq p \leq N_k - 1$$

where $\mathbb{1}_{jp}$ is defined as:

$$\mathbb{1}_{jp} = \begin{cases} 1 & \text{if } \hat{\eta}_k^{(p)} \leq \eta_k^{(j)} \leq \hat{\eta}_k^{(p+1)} \\ 0 & \text{otherwise} \end{cases}$$

The element in which we are making the largest error in fitting the data is henceforth identified by the index p^* satisfying:

$$p_* = \arg \max_{1\leq p\leq N_k-1} \hat{E}_p^2.$$

To refine the mesh, we split the p_*-th element into two equal in size sub-elements. This is an a priori choice and more sophisticated choices will be the object of further investigations.

We then perform the regression step explained in the first part of this section in order to compute the values of the degrees of freedom and update the error.

The collocation matrix W is updated by adding 2 columns, corresponding to the the degrees of freedom added after the point $\hat{\eta}_k^+$ has splitted the p_* element:

$$\eta_k^+ = \frac{1}{2}\left(\hat{\eta}_k^{p_*} + \hat{\eta}_k^{p_*+1}\right)$$

Only some of the entries of these two columns are non-zero, as the basis functions have compact support. Remark that, in terms of implementation, the update of the QR factorisation can be performed in a fast way, in the spirit of [6]. The cost of the $\mathbb{P}_3$-Hermite representation of the $k-$th function σ_k is given by $3N_k$.

Algorithm: $\mathbb{P}_3$-***Hermite refinement and regression.***
Data: $\left\{\hat{\eta}_k^{(q)}\right\}_{1\leq q\leq N_k}$, $\left\{u_k^{(j)}\right\}_{1\leq j\leq N}$
Result: $\left\{\hat{\eta}_k^{(q)}\right\}_{1\leq q\leq N_k+1}$, $\{S_{kq}\}_{1\leq q\leq 2(N_k+1)}$.
Find the element p^*;
Split the element p^* and update the nodes $\left\{\hat{\eta}_k^{(q)}\right\}_{1\leq q\leq N_k+1}$;
Update the columns of W;
Update the QR decomposition of W;
Compute the dofs $\{S_{kq}\}_{1\leq q\leq 2N_k}$.

Algorithm 2: Mesh refinement and regression, described in Section 2.2.

Remark: in terms of implementation, since the method consists in building the decoder by performing a dof-wise regression, we can proceed as follows. Before starting, we build the set of the degrees of freedom whose value is constant and independent from μ. This is for instance the case when considering certain type of boundary conditions, or if there are regions of the domain in which the solution value does not depend on the parameters. For these dofs, the regression reduces simply to store the value of the degree of freedom (as if we had just one finite element, and all the snapshots would have the same coordinate η).

3 Analysis of the method.

A preliminary analysis is presented hereafter. The goal is to investigate, under certain technical assumptions, the ability to approximate the set of values $u_k^{(j)}$ by using the functions σ_k. Let $\Omega \subset \mathbb{R}^d$ be an open bounded set, the physical domain, and let $\Omega_\mu \subset \mathbb{R}^p$ be the parameter domain.

The solution of the system is a function $u : \Omega \times \Omega_\mu \to \mathbb{R}^m$. We consider here the case in which $m = 1$, the extension to vector valued function being straightforward. The FOM is obtained by discretising the PDE, introducing a space $V \subset \mathcal{V}$. The degrees of freedom of the discretisation can be seen as the application of continuous linear forms to the solutions: let $\{v_k^*\}_{1\leq k\leq \mathcal{N}} \in \mathcal{V}^*$ be the set of bi-orthogonal functionals associated to the basis functions $\{v_k\}_{1\leq k\leq \mathcal{N}} \in \mathcal{V}$, $u_k = \langle u, v_k^*\rangle \in \mathbb{R}$. In this section, the subscript k is the index of the degree of freedom of the FOM discretisation of the solution.

We denote $\{g_k\}_{1\leq k\leq \mathcal{N}}$ the maps associating a parameter instance to the values of the degrees of freedom of the FOM. The $k-$th map is the composition between the $k-$th linear form and the parameter to solution map, and it is defined as:

$$g_k : \left\{ \begin{array}{l} \Omega_\mu \ \to \ \mathbb{R} \\ \mu \ \mapsto u_k = g_k(\mu) = \langle u(\cdot, \mu), v_k^* \rangle_{\mathcal{V},\mathcal{V}^*} \end{array} \right.$$

The POD step provides an orthonormal basis $\{\varphi_i\}_{1\leq i\leq n}$. In what follows we denote the classical $L^2(\Omega)$ scalar product $\langle u, v\rangle = \int_\Omega uv \ dx$, for $u, v \in L^2(\Omega)$.

In this section, in order to highlight the dependence of the reduced coordinates on the parameter values, we denote $a(\mu) \in \mathbb{R}^n$, whose components are defined as $a_i = \langle u(\cdot, \mu), \varphi_i\rangle$. We consider here the $L^2(\Omega \times \Omega_\mu)$ norm and define the best linear reconstruction error with n modes as:

$$\mathcal{E}^2_{\mathcal{L},n} = \|e_{\mathcal{L},n}\|^2_{L^2(\Omega\times\Omega_\mu)} = \int_{\Omega_\mu}\int_\Omega \left(u - \sum_{i=1}^n \langle u, \varphi_i\rangle u\right)^2 \ dxd\mu.$$

The functions σ_k are ridge functions, where the ridge is built by exploiting the POD modes. Let us introduce a function f_k, which is the scalar product between the $k-$th ridge vector and the reduced coefficients $a(\mu)$:

$$f_k : \left\{ \begin{array}{l} \Omega_\mu \ \to \ \mathbb{R} \\ \mu \ \mapsto f_k(\mu) = [\xi^{(k)}]^T a(\mu) \end{array} \right.$$

Let the image of f_k be $[a_k, b_k] \in \mathbb{R}$, and let $\eta_k \in [a_k, b_k]$. We introduce the following set:

$$\Omega_{\eta_k} = \{\mu \in \Omega_\mu \text{ such that : } f_k(\mu) = \eta_k\}.$$

The set Ω_{η_k} is a level set, implicitly defined by the equation $f_k(\mu) = \eta_k$. Its characteristic function will be denoted by $\mathbb{1}_{\eta_k} = \mathbb{1}_{\{f(\mu)=\eta_k\}}$.

The linear reconstruction errors of the degrees of freedom as function of the parameters values is a set of functions $\{e_k\}_{1\leq k\leq \mathcal{N}}$ defined as:

$$e_k(\mu) = g_k(\mu) - f_k(\mu).$$

We introduce a last function which will be useful in what follows:

$$\overline{u}_k : \begin{cases} [a_k, b_k] \ \rightarrow \ \mathbb{R} \\ \eta_k \ \mapsto \overline{u}_k(\eta_k) = \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_{\eta_k}} g_k(\mu) \ ds \end{cases} \tag{5}$$

We present the hypotheses of the current analysis:

(H_1) The solution u belongs to $\mathcal{V} \times W^{1,\infty}(\Omega_\mu) \cap H^2(\Omega_\mu)$. Moreover, $\nabla_\mu u$ vanishes on a zero measure set. The regularity of $\mathcal{V}$ depends essentially on the differential operators involved in the PDE and $\mathcal{V} \subseteq L^2(\Omega)$. The regularity in the parametric domain states that in the present analysis we consider parametric domains in which the solution is differentiable in μ almost everywhere and the second derivatives are square integrable. Otherwise stated, we consider the situation in which the model sensitivity functions are bounded and do not vanish, excepted possibly on a zero measure set and their derivatives are square integrable.

Before stating the main result of the present analysis, we introduce some Lemmas.

1. *Lemma:* *Let the hypotheses* (H_1) *hold. The function* $\overline{u}_k$ *defined in Eq.*(5) *is in* $W^{1,\infty}([a_k, b_k])$.

Proof. The function $\overline{u}_k$ evaluated in a point η_k is defined as:

$$\overline{u}_k(\eta_k) = \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_\mu} g_k(\mu) \mathbb{1}_{\eta_k} \ d\mu$$

The level set, whose characteristic function is $\mathbb{1}_{\eta_k}$, is defined implicitly by:

$$f_k(\mu) = \sum_{i=1}^{n} \langle u(\cdot, \mu), \varphi_i \rangle \langle \varphi_i, v_k^* \rangle_{\mathcal{V},\mathcal{V}^*} = \eta_k.$$

The function f_k inherits the regularity of u for what concerns the parametric dependence. Henceforth: $f_k \in W^{1,\infty}(\Omega_\mu)$. In a level set which is differentiable almost everywhere, the gradient of f_k, denoted by $\nabla_\mu f_k \in L^\infty(\Omega_\mu)$ is orthogonal almost everywhere to the level set. The differential of the level set with respect to the value η_k

is related to the gradient of f_k and the normal to the level set, which is the gradient itself, renormalised. This makes it possible to write:

$$\partial_{\eta_k} \left(|\Omega_{\eta_k}| \overline{u}_k \right) = \partial_{\eta_k} \int_{\Omega_{\eta_k}} g_k(\mu) \, d\mu.$$

where the domain Ω_{η_k} depends on the value of η_k. The value of the integral of g_k at the right-hand side changes because of the change in the level set. The field making the level set change is simply $\nabla_\mu f_k$, and the normal field to the level set is defined as $\mathbf{n} = \frac{\nabla_\mu f_k}{|\nabla_\mu f_k|}$. We get:

$$\partial_{\eta_k} \int_{\Omega_{\eta_k}} g_k(\mu) \, ds = \int_{\Omega_{\eta_k}} \nabla \cdot (g_k \nabla_\mu f_k) - g_k \left(\nabla_\mu \nabla_\mu f_k \cdot \frac{\nabla_\mu f_k}{|\nabla_\mu f_k|} \cdot \frac{\nabla_\mu f_k}{|\nabla_\mu f_k|} \right) ds.$$

This gives:

$$\partial_{\eta_k} \int_{\Omega_{\eta_k}} g_k(\mu) \, ds = \int_{\Omega_{\eta_k}} \nabla \cdot (g_k \nabla_\mu f_k) - \frac{g_k}{|\nabla_\mu f_k|^2} \left(\nabla_\mu \nabla_\mu f_k \cdot \nabla_\mu f_k \cdot \nabla_\mu f_k \right) ds.$$

On the left hand side, we have the term $\partial_{\eta_k} |\Omega_{\eta_k}|$ which reads:

$$\partial_{\eta_k} \int_{\Omega_{\eta_k}} 1 \, ds = \int_{\Omega_{\eta_k}} \Delta f_k - \left(\nabla_\mu \nabla_\mu f_k \cdot \frac{\nabla_\mu f_k}{|\nabla_\mu f_k|} \cdot \frac{\nabla_\mu f_k}{|\nabla_\mu f_k|} \right) ds.$$

We put the equations together and get:

$$|\Omega_{\eta_k}| \partial_{\eta_k} \overline{u}_k = -\partial_{\eta_k} |\Omega_k| \overline{u}_k + \int_{\Omega_{\eta_k}} \nabla \cdot (g_k \nabla_\mu f_k) - \frac{g_k}{|\nabla_\mu f_k|^2} \left(\nabla_\mu \nabla_\mu f_k \cdot \nabla_\mu f_k \cdot \nabla_\mu f_k \right) ds.$$

We observe that $\overline{u}_k$ depends on η_k and η_k is constant by construction on Ω_{η_k}. Thus:

$$|\Omega_{\eta_k}| \partial_{\eta_k} \overline{u}_k = \int_{\Omega_{\eta_k}} \nabla \cdot ((g_k - \overline{u}_k) \nabla_\mu f_k) - \frac{(g_k - \overline{u}_k)}{|\nabla_\mu f_k|^2} \left(\nabla_\mu \nabla_\mu f_k \cdot \nabla_\mu f_k \cdot \nabla_\mu f_k \right) ds.$$

Since f_k, g_k inherit the regularity properties of the solution u, we see that the derivative $\partial_{\eta_k} \overline{u}_k$ is bounded almost everywhere on the domain, for all values of η_k. □

2. *Lemma:* *Under the hypothesis* (H_1) *it holds:*

$$\int_{\Omega_\mu} \left(g_k(\mu) - \overline{u}_k(\mu) \right)^2 \, d\mu = \int_{a_k}^{b_k} \int_{\Omega_{\eta_k}} \left(e_k - \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_{\eta_k}} e_k \, ds \right)^2 ds \, d\eta_k \leq \|e_k\|^2_{L^2(\Omega_\mu)}.$$

Proof. Under the hypothesis (H_1) we can see the volume measure $d\mu$ as the product between the measure ds of the hypersurfaces Ω_{η_k} and $d\eta_k$. It holds:

$$\int_{\Omega_\mu} \left(g_k(\mu) - \overline{u}_k(\mu) \right)^2 \, d\mu = \int_{a_k}^{b_k} \int_{\Omega_{\eta_k}} \left(g_k(\mu) - \overline{u}_k(\mu) \right)^2 \, ds \, d\eta_k$$

Before developing the integral, let us observe the following:

$$g_k = f_k + e_k,$$

$$\overline{u}_k = \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} f\ ds + \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds = \eta_k + \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds.$$

We insert these expressions in the integral and obtain:

$$\int_{a_k}^{b_k}\int_{\Omega_{\eta_k}} \left(g_k(\mu) - \overline{u}_k(\mu)\right)^2\ ds\ d\eta_k = \int_{a_k}^{b_k}\int_{\Omega_{\eta_k}} \left(e_k - \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds + f_k - \eta_k\right)^2 ds\ d\eta_k$$

On every hypersurface Ω_{η_k}, by construction, $f_k = \eta_k$, leading to:

$$\int_{a_k}^{b_k}\int_{\Omega_{\eta_k}} \left(g_k(\mu) - \overline{u}_k(\mu)\right)^2\ ds\ d\eta_k = \int_{a_k}^{b_k}\int_{\Omega_{\eta_k}} \left(e_k - \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds\right)^2 ds\ d\eta_k.$$

This quantity is henceforth bounded by the linear reconstruction error:

$$\int_{\Omega_\mu} \left(g_k(\mu) - \overline{u}_k(\mu)\right)^2\ d\mu \leq \|e_k\|^2_{L^2(\Omega_\mu)}.$$

□

3. *Lemma:* *Let us consider the error when trying to approximate the function $\overline{u}_k$ with the function σ_k. The function σ_k is decomposed into the following sum: $\sigma_k = \eta_k + \zeta_k$, where ζ_k is built by using the $\mathbb{P}_3$-Hermite finite element space. Let the error of the linear reconstruction be $e_k : \Omega_\mu \to \mathbb{R}$ defined as $e_k(\mu) = g_k(\mu) - f_k(\mu)$. Its average on Ω_{η_k}, denoted $\overline{e}_k$, is defined as:*

$$\overline{e}_k(\eta_k) = \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds.$$

Let $M_k \in \mathbb{N}^*$ be the number of degrees of freedom used in the finite element approximation. There exists two constants $\omega_k > 0$ and $0 \leq \beta_k \leq 1$, such that:

$$2\int_{a_k}^{b_k} \zeta_k \overline{e}_k |\Omega_{\eta_k}| d\eta_k - \int_{a_k}^{b_k} \zeta_k^2\ |\Omega_{\eta_k}| d\eta_k \geq \left[1 - \frac{\omega_k^2}{M_k^2}\right] \beta_k^2 \|e_k\|^2_{L^2(\Omega_\mu)}.$$

Proof. Let us consider the term: $\int_\Omega |\overline{u}_k(f_k(\mu)) - \sigma(f_k(\mu))|^2\ d\mu$.

Under the hypotheses (H_1) we can write:

$$\int_{a_k}^{b_k}\int_{\Omega_{\eta_k}} |\overline{u}_k(\eta_k) - \sigma(\eta_k)|^2 ds\ d\eta_k = \int_{a_k}^{b_k} |\overline{u}_k(\eta_k) - \sigma(\eta_k)|^2 |\Omega_{\eta_k}|\ d\eta_k.$$

The term $\overline{u}_k$ can be written as:

$$\overline{u}_k = \eta_k + \frac{1}{|\Omega_{\eta_k}|}\int_{\Omega_{\eta_k}} e_k\ ds = \eta_k + \overline{e}_k.$$

We have:

$$\int_{a_k}^{b_k} |\overline{u}_k(\eta_k) - \sigma(\eta_k)|^2 |\Omega_{\eta_k}|\ d\eta_k = \int_{a_k}^{b_k} |\overline{e}_k(\eta_k) - (\sigma(\eta_k) - \eta_k)|^2 |\Omega_{\eta_k}|\ d\eta_k,$$

where we put in evidence the update term $\zeta_k = \sigma(\eta_k) - \eta_k$, which represent the difference between the linear reconstruction and the non-linear one. When considering $M_k = 2$ degrees of freedom, the linear reconstruction can be recovered exactly (the identity being exactly represented). The expression leads to:

$$\int_{a_k}^{b_k} |\overline{e}_k(\eta_k) - \zeta_k|^2 |\Omega_{\eta_k}|\ d\eta_k = \int_{a_k}^{b_k} \overline{e}_k^2\ |\Omega_{\eta_k}| d\eta_k - 2\int_{a_k}^{b_k} \zeta_k \overline{e}_k |\Omega_{\eta_k}| d\eta_k + \int_{a_k}^{b_k} \zeta_k^2\ |\Omega_{\eta_k}| d\eta_k.$$

By computing the best $\mathbb{P}_3$-Hermite approximation of $\overline{u}_k$ we cannot do worse, by construction, than when using the linear approximation. It holds:

$$\int_{a_k}^{b_k} |\overline{e}_k(\eta_k) - \zeta_k|^2 |\Omega_{\eta_k}|\ d\eta_k \leq \int_{a_k}^{b_k} \overline{e}_k^2\ |\Omega_{\eta_k}| d\eta_k \ \Rightarrow\ -2\int_{a_k}^{b_k} \zeta_k \overline{e}_k |\Omega_{\eta_k}| d\eta_k + \int_{a_k}^{b_k} \zeta_k^2\ |\Omega_{\eta_k}| d\eta_k \leq 0.$$

Let us consider the error that P_3 Hermite finite elements have when approximating a function belonging to $H^1([a_k, b_k])$, which is true by virtue of the result of Lemma 1. By making use of Aubin-Nitsche Lemma, we can claim that there exists a constant $C_F > 0$, such that:

$$\int_{a_k}^{b_k} |\overline{e}_k(\eta_k) - \zeta_k|^2 |\Omega_{\eta_k}|\ d\eta_k \leq \frac{C_F^2}{M_k^2} |\overline{e}_k|^2_{H^1([a_k,b_k])}, \tag{6}$$

where $M_k \in \mathbb{N}$ is the number of degrees of freedom used to compute ζ_k. Remark that the constant C_F accounts also for the measure $|\Omega_{\eta_k}|$ in the present case. Let us distinguish two cases:

(i) $\overline{e}_k = 0$: in this case, the linear approximation, which provides the values η_k is the best approximation. Henceforth $M_k = 0$ in view of promoting parsimony.

(ii) $\overline{e}_k \neq 0$. In this case, $\int_{a_k}^{b_k} \overline{e}_k^2 |\Omega_k|\ d\eta_k > 0$. There exists a constant $0 < \beta_k \leq 1$ such that:

$$\int_{a_k}^{b_k} \overline{e}_k^2 |\Omega_{\eta_k}|\ d\eta_k \geq \beta_k^2 \|e_k\|^2_{L^2(\Omega_\mu)}.$$

Let us consider the second case, and develop the expression of Eq.(6):

$$-2\int_{a_k}^{b_k} \zeta_k \overline{e}_k |\Omega_{\eta_k}| d\eta_k + \int_{a_k}^{b_k} \zeta_k^2\ |\Omega_{\eta_k}| d\eta_k \leq \frac{C_F^2}{M_k^2} |\overline{e}_k|^2_{H^1([a_k,b_k])} - \int_{a_k}^{b_k} \overline{e}_k^2 |\Omega_{\eta_k}|\ d\eta_k.$$

There exists a constant, $c_{10} > 0$ such that: $|\overline{e}_k|^2_{H^1([a_k,b_k])} \leq c_{10}^2 \int_{a_k}^{b_k} \overline{e}_k^2 |\Omega_{\eta_k}|\ d\eta_k$. It holds:

$$-2\int_{a_k}^{b_k} \zeta_k \overline{e}_k |\Omega_{\eta_k}| d\eta_k + \int_{a_k}^{b_k} \zeta_k^2\ |\Omega_{\eta_k}| d\eta_k \leq \left[\frac{C_F^2 c_{10}^2}{M_k^2} - 1\right] \int_{a_k}^{b_k} \overline{e}_k^2 |\Omega_{\eta_k}|\ d\eta_k$$

It is clear that, in this case, the term at the left hand side is negative. We introduce the constant β_k and state that:

$$-2\int_{a_k}^{b_k}\zeta_k\overline{e}_k|\Omega_{\eta_k}|d\eta_k+\int_{a_k}^{b_k}\zeta_k^2\ |\Omega_{\eta_k}|d\eta_k\leq-\left[1-\frac{C_F^2c_{10}^2}{M_k^2}\right]\beta_k^2\|e_k\|^2_{L^2(\Omega_\mu)}$$

In more general terms, we can consider, from now on: $0\leq\beta_k\leq 1$ and, if $\beta_k=0$, then $M_k=0$. Furthermore, there exists a constant $\omega_k>0$ such that:

$$-2\int_{a_k}^{b_k}\zeta_k\overline{e}_k|\Omega_{\eta_k}|d\eta_k+\int_{a_k}^{b_k}\zeta_k^2\ |\Omega_{\eta_k}|d\eta_k\leq-\left[1-\frac{\omega_k^2}{M_k^2}\right]\beta_k^2\|e_k\|^2_{L^2(\Omega_\mu)},$$

and this concludes the proof. □

We have now all the elements to prove the main result concerning the analysis of the method.

1. *Proposition:* *Let the hypotheses of Lemma 1-3 hold. There exists positive constants* $\{z_k\}_{1\leq k\leq\mathcal{N}}\leq 1$, *such that:*

$$\mathcal{E}^2_{\mathcal{R},n}\leq\sum_{k=1}^{\mathcal{N}}z_k\|e_k\|^2_{L^2(\Omega_\mu)}\leq\mathcal{E}^2_{\mathcal{L},n},$$

and:

$$z_k=\left[1-\left(1-\frac{\omega_k^2}{M_k^2}\right)\beta_k^2\right].$$

Proof. Let us consider the $k-$th degree of freedom. The regression error we make when considering the ridge function σ_k to approximate the function g_k reads:

$$\|E_k\|^2_{L^2(\Omega_\mu)}=\int_{\Omega_\mu}|g_k(\mu)-\sigma_k(f_k(\mu))|^2\ d\mu.$$

First, remark that, by construction, we have:

$$\sum_{k=1}^{\mathcal{N}}\|E_k\|^2_{L^2(\Omega_\mu)}\leq\sum_{k=1}^{\mathcal{N}}\|e_k\|^2_{L^2(\Omega_\mu)}\leq\mathcal{E}^2_{\mathcal{L},n}.$$

We consider the k-th term of the sum:

$$\|E_k\|^2_{L^2(\Omega_\mu)}=\int_{\Omega_\mu}|g_k(\mu)-\sigma_k(f_k(\mu))|^2\ d\mu.$$

We add and subtract the function $\overline{u}_k$, and obtain:

$$\|E_k\|^2_{L^2(\Omega_\mu)}=\int_{\Omega_\mu}|g_k(\mu)-\overline{u}_k(f_k(\mu))+\overline{u}_k(f_k(\mu))-\sigma_k(f_k(\mu))|^2\ d\mu.$$

We develop the integral and get:

$$\|E_k\|^2_{L^2(\Omega_\mu)} = \int_{\Omega_\mu} |g_k(\mu) - \overline{u}_k|^2 \, d\mu + 2\int_{\Omega_\mu} (g_k(\mu) - \overline{u}_k)(\overline{u}_k - \sigma_k(f_k(\mu))) \, d\mu + \int_{\Omega_\mu} |\overline{u}_k - \sigma_k(f_k(\mu))|^2 \, d\mu$$

Let us focus on the second term and prove that it vanishes. Let us consider the volume measure as the product of the hypersurface measure and the measure of η_k:

$$\int_{\Omega_\mu} (g_k(\mu) - \overline{u}_k)(\overline{u}_k - \sigma_k(f_k(\mu))) \, d\mu = \int_{a_k}^{b_k} \int_{\Omega_{\eta_k}} (g_k - \overline{u}_k)(\overline{u}_k - \sigma_k) \, ds \, d\eta_k.$$

The term $(\overline{u}_k - \sigma_k)$ is constant on the hypersurfaces and $\int_{\Omega_{\eta_k}} (g_k - \overline{u}_k) \, ds = 0$ by definition of $\overline{u}_k$. This term vanishes.

By making use of Lemma 2 we get:

$$\|E_k\|^2_{L^2(\Omega_\mu)} = \int_{a_k}^{b_k} \int_{\Omega_{\eta_k}} \left(e_k - \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_{\eta_k}} e_k \, ds \right)^2 ds \, d\eta_k + \int_{a_k}^{b_k} (\overline{u}_k - \sigma_k)^2 |\Omega_{\eta_k}| \, d\eta_k.$$

Since we can write $\overline{u}_k = \eta_k + \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_{\eta_k}} e_k \, ds$ and $\sigma_k = \eta_k + \zeta_k$:

$$\|E_k\|^2_{L^2(\Omega_\mu)} = \int_{a_k}^{b_k} \int_{\Omega_{\eta_k}} \left(e_k - \frac{1}{|\Omega_{\eta_k}|} \int_{\Omega_{\eta_k}} e_k \, ds' \right)^2 ds d\eta_k + \int_{a_k}^{b_k} \left(\frac{1}{|\Omega_k|} \int_{\Omega_{\eta_k}} e_k ds' - \zeta_k \right)^2 |\Omega_{\eta_k}| d\eta_k.$$

For the sake of compactness in the notation, we recall that:

$$\overline{e}_k = \frac{1}{|\Omega_k|} \int_{\Omega_{\eta_k}} e_k \, ds$$

And get:

$$\|E_k\|^2_{L^2(\Omega_\mu)} = \int_{a_k}^{b_k} \left(\int_{\Omega_{\eta_k}} e_k^2 \, ds \right) - 2|\Omega_{\eta_k}| \overline{e}_k^2 + |\Omega_{\eta_k}| \overline{e}_k^2 + |\Omega_{\eta_k}| \overline{e}_k^2 - 2|\Omega_{\eta_k}| \overline{e}_k \zeta_k + |\Omega_{\eta_k}| \zeta_k^2 \, d\eta_k$$

which leads to:

$$\|E_k\|^2_{L^2(\Omega_\mu)} = \int_{a_k}^{b_k} \left(\int_{\Omega_{\eta_k}} e_k^2 \, ds \right) - 2|\Omega_{\eta_k}| \overline{e}_k \zeta_k + |\Omega_{\eta_k}| \zeta_k^2 \, d\eta_k.$$

Let us remark that the second term is negative, by virtue of Lemma 3. This is consistent with the fact that the function σ_k helps in reducing the error made by the linear reconstruction performed with $n \leq n_*$ modes. We make use of the result of the Lemma 3 and write:

$$\|E_k\|^2_{L^2(\Omega_\mu)} \leq \|e_k\|^2_{\Omega_\mu} - \left[1 - \frac{\omega_k^2}{M_k^2} \right] \beta_k^2 \|e_k\|^2_{L^2(\Omega_\mu)}.$$

The general expression of the error reads:

$$\mathcal{E}^2_{\mathcal{R},n} = \sum_{k=1}^{\mathcal{N}} \|E_k\|^2_{L^2(\Omega_\mu)} \leq \sum_{k=1}^{\mathcal{N}} \left[1 - \left(1 - \frac{\omega_k^2}{M_k^2}\right)\beta_k^2\right] \|e_k\|^2_{L^2(\Omega_\mu)} \leq \mathcal{E}^2_{\mathcal{L},n}. \tag{7}$$

□

We see from the expression in Eq. (7) that the worst case scenario is given by $\beta_k = 0$, for all $k = 1, \ldots \mathcal{N}$, and in such a case we obtain the result of the linear approximation. The best possible scenario is the one in which the linear information is such that there exists a one-to-one mapping between the variables η_k (the linear reconstruction of the degrees of freedom) and the degrees of freedom values u_k, and in this case it would lead to $\beta_k = 1,\ 1 \leq k \leq \mathcal{N}$. The error bound would become:

$$\mathcal{E}^2_{\mathcal{R},n}\big|_{\beta_k=1} = \sum_{k=1}^{\mathcal{N}} \left[\frac{\omega_k^2}{M_k^2}\right] \|e_k\|^2_{L^2(\Omega_\mu)}.$$

In this case, the error can be made arbitrarily small by increasing the number of finite elements degrees of freedom of the dof-wise regressions.

Let us remark that β_k, ω_k are intrinsic properties of the set of solutions we are considering, and they depend on n, the number of POD modes we used to get the linear information about this set. The number of finite element degrees of freedom used in the dof-wise regressions, M_k, helps in reducing the error. In the most general case, the best possible error we can achieve is:

$$\lim_{\{M_k\}_{1\leq k\leq \mathcal{N}} \to \infty} \mathcal{E}^2_{\mathcal{R},n} = \sum_{k=1}^{\mathcal{N}} \left[1 - \beta_k^2\right] \|e_k\|^2_{L^2(\Omega_\mu)}.$$

From this expression we can deduce that, in order to reach the prescribed accuracy, a necessary condition is:

$$\sum_{k=1}^{\mathcal{N}} \left[1 - \beta_k^2\right] \|e_k\|^2_{L^2(\Omega_\mu)} \leq \varepsilon^2.$$

This tells us that, if the value of n is too low, we do not have enough information about the solution set to be able to reach the prescribed accuracy.

An important question concerns the computational cost, in terms of the number of finite element degrees of freedom to be used in the dof-wise regression. The error expression makes it possible to investigate the best possible computational cost; this is the object of the next Proposition.

2. *Proposition:* *Let the hypotheses of Proposition 1 hold. The memory burden, consisting in the number of quantities to be stored, denoted $\mathcal{C}_{\mathcal{R},n}$, is not smaller than:*

$$\mathcal{C}^*_{\mathcal{R},n} = (\mathcal{N} + N)n + \frac{3}{2} \frac{\left[\sum_{k=1}^{\mathcal{N}} (\beta_k \omega_k \|e_k\|_{L^2(\Omega_\mu)})^{2/3}\right]^{3/2}}{\left[\varepsilon^2 - \sum_{k=1}^{\mathcal{N}} (1 - \beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)}\right]^{1/2}}.$$

Proof. Let us focus on the number of finite element degrees of freedom we have to use in the dof-wise regression. Based on the error expression found in Eq.(7), we can try to optimise the number of degrees of freedom under the constrain that we satisfy the prescribed accuracy. To this end, we introduce the Lagrangian:

$$L = \sum_{k=1}^{\mathcal{N}} M_k + \lambda \left[\sum_{k=1}^{\mathcal{N}} (1-\beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)} + \frac{(\beta_k \omega_k \|e_k\|_{L^2(\Omega_\mu)})^2}{M_k^2} - \varepsilon^2 \right],$$

and we solve the following problem:

$$(\{M_k^*\}_{1\leq k\leq \mathcal{N}}, \lambda_*) = \arg \inf_{\{M_k\}_{1\leq k\leq \mathcal{N}}} \sup_{\lambda} L(M_k, \lambda).$$

First, this Lagrangian is linear with convex constraint: if the solution exists, it is unique. The problem is solved in a classical way: we derive the Euler-Lagrange equations, express the primal variables M_k as function of the dual variable λ and then inject the expression into the constraint in order to solve for the dual variable. We have:

$$M_r^3 = (\beta_r \omega_r \|e_r\|_{L^2(\Omega_\mu)})^2 \lambda \;\Rightarrow\; M_r^2 = (\beta_r \omega_r \|e_r\|_{L^2(\Omega_\mu)})^{4/3} \lambda^{2/3}, \quad 1 \leq r \leq \mathcal{N}.$$

When this expression is injected into the constraint, we have:

$$\lambda^{2/3} = \frac{\sum_{k=1}^{\mathcal{N}} (\beta_k \omega_k \|e_k\|_{L^2(\Omega_\mu)})^{2/3}}{\varepsilon^2 - \sum_{k=1}^{\mathcal{N}} (1-\beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)}}.$$

The denominator of the above written expression highlights the fact that, if the quantities β_k, as defined in Lemma 3, are too small, there might not be a solution to the problem, *i.e.* we cannot satisfy the error constraint. The necessary condition for this to be possible reads:

$$\varepsilon^2 - \sum_{k=1}^{\mathcal{N}} (1-\beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)} > 0.$$

We can now use the expression of the Lagrange multiplier in view of computing the best possible computational cost (the memory to store the finite element approximation of the non-linear functions σ_k). After some algebra, we get:

$$\mathcal{C}^*_{\mathcal{R},n} = (\mathcal{N}+N)n + \frac{3}{2} \sum_{k=1}^{\mathcal{N}} M_k = (\mathcal{N}+N)n + \frac{3}{2} \frac{\left[\sum_{k=1}^{\mathcal{N}} (\beta_k \omega_k \|e_k\|_{L^2(\Omega_\mu)})^{2/3} \right]^{3/2}}{\left[\varepsilon^2 - \sum_{k=1}^{\mathcal{N}} (1-\beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)} \right]^{1/2}}.$$

and this concludes the proof. □

Remark that the expression found is compatible with the fact that it is necessary that $\varepsilon^2 - \sum_{k=1}^{\mathcal{N}} (1-\beta_k^2) \|e_k\|^2_{L^2(\Omega_\mu)} > 0$ in order for a dof-wise regression to be able to fulfil the prescribed accuracy. The quantities β_k, ω_k depend on the kind of solution sets we are considering, and have to be studied case by case.

Let us focus on the most favorable case, namely, the one in which $\beta_k = 1,\ 1 \leq k \leq \mathcal{N}$. In this case the memory burden reduces to:

$$\mathcal{C}^*_{\mathcal{R},n}\big|_{\beta_k=1} = (\mathcal{N}+N)n + \frac{3}{2\varepsilon}\left[\sum_{k=1}^{\mathcal{N}}(\omega_k\|e_k\|_{L^2(\Omega_\mu)})^{2/3}\right]^{3/2}. \tag{8}$$

Let us consider two scenarios and try to understand in which situations the method could provide an advantage over the linear reconstruction.

(H_2) We consider two distinct cases concerning the way the POD performs. Let us assume that there exist two constants $C_P, \gamma > 0$, such that:

(a) The error of the linear reconstruction satisfies: $e_{\mathcal{L},n} \leq C_P n^{-\gamma}$, for $\gamma > 1/2$.

(b) The error of the linear reconstruction satisfies: $e_{\mathcal{L},n} \leq C_P e^{-\gamma(n+1)}$.

3. *Proposition.* *Let the hypotheses (H_2) hold and let us assume that, for $n_* \in \mathbb{N}^*$ we have:*

$$C_P n_*^{-\gamma} = \varepsilon,$$

or

$$C_P n_* e^{-\gamma n} = \varepsilon,$$

for the hypotheses (a) and (b) respectively. Then, a necessary condition for the non-linear reduction to outperform the linear one is:

$$\max_{1\leq k\leq\mathcal{N}}\omega_k < \frac{2}{3}(\mathcal{N}+N)(n_*-n)\left(\frac{n_*}{n}\right)^{-\gamma}.$$

if the hypothesis $(H_2).(a)$ holds and:

$$\max_{1\leq k\leq\mathcal{N}}\omega_k < \frac{2}{3}(\mathcal{N}+N)(n_*-n)e^{-\gamma(n_*-n)}.$$

if the hypothesis $(H_2).(b)$ holds.

Proof. The proof is done by direct computation. The best case scenario for the non-linear reconstruction is described in Eq. (8). For the best non-linear reconstruction to outperform the linear one, the following has to be true:

$$(\mathcal{N}+N)n + \frac{3}{2\varepsilon}\left[\sum_{k=1}^{\mathcal{N}}(\omega_k\|e_k\|_{L^2(\Omega_\mu)})^{2/3}\right]^{3/2} < (\mathcal{N}+N)n_*,$$

which leads to:

$$\frac{3}{2\varepsilon}\left[\sum_{k=1}^{\mathcal{N}}(\omega_k\|e_k\|_{L^2(\Omega_\mu)})^{2/3}\right]^{3/2} < (\mathcal{N}+N)(n_*-n).$$

Let $v \in \mathbb{R}^{\mathcal{N}}$ be the vector whose components are: $v_k = \omega_k \|e_k\|_{L^2(\Omega_\mu)}$. The cost of the dof-wise regression is the quasi-norm $\ell^{2/3,\mathcal{N}}$ of v. The smallest possible value is achieved when only one term of the vector is different from zero (whereas the maximum is achieved when all the components are equal). We assume, without loss of generality, that this is achieved for $k = 1$.

Let us consider the case $(H_2).(a)$:

$$\frac{3\omega_1}{2\varepsilon} C_P n^{-\gamma} < (\mathcal{N} + N)(n_* - n),$$

which leads to:

$$\omega_1 < \frac{2}{3}(\mathcal{N} + N)(n_* - n)\frac{\varepsilon}{C_P} n^{\gamma}.$$

by hypothesis $\varepsilon / C_P = n_*^{-\gamma}$, which entails:

$$\omega_1 < \frac{2}{3}(\mathcal{N} + N)(n_* - n)\left(\frac{n_*}{n}\right)^{-\gamma}.$$

Let us now consider the hypothesis $(H_2).(b)$.

$$\frac{3\omega_1}{2\varepsilon} C_P e^{-\gamma n} < (\mathcal{N} + N)(n_* - n),$$

which entails:

$$\omega_1 < \frac{2}{3}(\mathcal{N} + N)(n_* - n)e^{-\gamma(n_* - n)}.$$

and this concludes the proof. □

The results of the Propositions 2 and 3 concern the best possible performances the non-linear reconstruction based on one ridge regression per degree of freedom can achieve. A more detailed analysis on the approximation performances for particular sets of functions will be the object of further investigations. Far from being exhaustive, the results of these propositions allow, nonetheless, to get some insight into the behaviour of the method. The non-linear reconstruction will be advantageous in all situations in which the linear approximation provided by the POD is poor (the constant γ is not large) and the error of the linear reconstruction tends to be concentrated in sparse subsets of the degrees of freedom.

4 Numerical experiments.

In this section, several numerical experiments are proposed to assess the properties of the proposed method. In all the tests proposed, we proceed as follows. We make the tolerance ε vary. For all the values tested, we report the value of n_*, the number of POD modes needed to reach the target accuracy, n, the number of modes used to build the ridge regression, leading to the maximal memory compression of the database. For this value, we report the memory compression ratio $\kappa \in \mathbb{R}$ defined as the ratio between

ε	n_*	n	κ	χ	n_0
10^{-1}	17	3	1.759	1.546	1
10^{-2}	739	2	38.62	63.45	1
10^{-3}	1021	2	47.90	85.66	1

Table 1: Values of the number of modes $n*, n, n_0$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the translating heaviside presented in Section 4.1.1.

the memory used to compress the database by using POD and the memory used by the proposed approach. Furthermore, for this n, we report the ratio between the number of operations we need to perform with POD and with the proposed method in order to, given a single snapshot, compute the reduced coordinates and reconstruct it. We denote it χ.

4.1 1-D test cases.

The first tests are related to 1-D setups, which are known to be challenging when we try to perform the reduction by using classical POD or the reduced basis method.

4.1.1 Translating Heaviside function.

Let $\overline{\Omega} = [0, 1]$. Let the parameter $\mu \in [0.1, 0.9]$. We consider the function u defined as:

$$u(x; \mu) = \begin{cases} 1 \text{ if } x \leq \mu \\ 0 \text{ if } x > \mu \end{cases} \tag{9}$$

In particular, we discretise the space by using classical finite differences (the degrees of freedom of the discretised solution are the values on the solution in the grid points) with $\mathcal{N} = 2048$ points. We consider $N = 1024$ snapshots, obtained by uniformly subdividing the parametric interval. We apply the method by considering three different tolerances $\varepsilon = \left\{10^{-1}, 10^{-2}, 10^{-3}\right\}$.

The results are reported in Table 1. The translating heaviside function is particularly critical for POD as the Kolmogorov widths are decreasing slowly (the decay goes as $n^{-1/2}$). This has been shown theoretically in [4], Chapter 3, and in [16]. The best results are obtained by adaptively computing a non-linearity based on ridges which exploits just two or three reduced coordinates. This leads to a significant compression ratio κ and reconstruction speed-up χ, especially when increasing the accuracy. When asking for $\varepsilon = 10^2$, this is already more than one order of magnitude.

The behaviour observed experimentally can be understood from a theoretical point of view. Let us remark that, given a certain degree of freedom in space (in this case the value of the solution in a grid point $\{x_i\}_{1 \leq i \leq \mathcal{N}}$), the solution to be approximated is a heaviside, function of μ. The approximation is built by means of $\mathbb{P}_3-$Hermite finite elements. It is exact in all the elements but the one containing the discontinuity.

Let the element size be $h > 0$. In the element the discontinuity falls in, say the $\bar{j}-$th one, we are trying to approximate the heaviside with a third order polynomial. At the boundary of the element the degrees of freedom are $(0, 1)$ for what concerns the function values and $(0, 0)$ for what concerns the derivatives. The local representation of the Heaviside function is the polynomial: $u_h^{(\bar{j})} = \frac{3}{h^2}(\mu - \mu_{\bar{j}})^2 - \frac{2}{h^3}(\mu - \mu_{\bar{j}})^3$. The error in $L^2(\Omega_\mu)$ norm, denoted by $e_{\bar{j}}$ is henceforth bounded by:

$$e_{\bar{j}}^2 \leq \int_0^h (1 - u_h(\mu))^2 \, d\mu.$$

This gives us that there exists a constant $c_{\bar{j}}$ such that $|e_{\bar{j}}| \leq c_{\bar{j}} h^{1/2}$. If, in order to obtain the prescribed accuracy, we could distribute the error squared uniformly, we would get:

$$h \leq \left(\frac{\varepsilon^2}{c_{\bar{j}}^2 \mathcal{N}} \right).$$

Let $L_{\bar{j}} \subset \mathbb{R}$ be the size of the 1-d domain in μ. Since the adaptive algorithm proposed splits the elements in half, progressively, based on the error, we would reach the size h after $p \in \mathbb{N}^*$ steps such that:

$$2^{-p} L_{\bar{j}} \leq h \leq \left(\frac{\varepsilon^2}{c_{\bar{j}}^2 \mathcal{N}} \right)$$

This gives us:

$$p \geq 2 \log_2 \left(\frac{c_{\bar{j}} L_{\bar{j}}^{1/2} \mathcal{N}^{1/2}}{\varepsilon} \right)$$

The number of elements corresponding to this equals $p+1$ leading to a memory of the order of $3 \left[1 + 2 \log_2 \left(\frac{c_{\bar{j}} L_{\bar{j}}^{1/2} \mathcal{N}^{1/2}}{\varepsilon} \right) \right]$ per degree of freedom in x and a total storage of size:

$$\mathcal{M} \approx 3 \left[1 + 2 \log_2 \left(\frac{c_{\bar{j}} L_{\bar{j}}^{1/2} \mathcal{N}^{1/2}}{\varepsilon} \right) \right] \mathcal{N}.$$

If we use the POD, the storage is: $\mathcal{M}_{\text{POD}} = n_*(\mathcal{N} + N)$. The rank n_* depends on the accuracy. In particular, for the translating Heaviside there exists a constant $C_{\text{POD}} > 0$ such that: $n_* \leq C_{\text{POD}} \varepsilon^{-2}$.

Let us remark that a scaling $\propto \log(1/\varepsilon)$ is optimal in the sense of the entropy numbers of the set of Heaviside functions in the unit interval.

4.1.2 Parametric family of Poisson solutions in varying domain.

Let $\mu_1 \in [1, 2]$, $\mu_2 \in [0, 8]$ and $\Omega(\mu) = [\mu_2, \mu_2 + \mu_1]$. We consider the solutions of the following problem:

$$-\partial_x^2 u(x) = \frac{1}{\mu_1^2}, \quad \text{in } \Omega(\mu),$$

ε	n_*	n	κ	χ	n_0
10^{-1}	18	13	1.10	1.14	2
10^{-2}	70	17	1.33	3.13	2
10^{-3}	237	14	3.31	9.84	2

Table 2: Values of the number of modes $n*, n, n_0$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the parametric 1d Poisson problem presented in Section 4.1.2.

with:

$$u(\mu_2) = u(\mu_2 + \mu_1) = 0$$

We can see that, for all μ_1, μ_2 there exists a unique solution $u \in H_0^1(\Omega(\mu))$, and it is:

$$u(x; \mu_1, \mu_2) = \frac{(x - \mu_2)}{2\mu_1} \left(1 - \frac{(x - \mu_2)}{\mu_1}\right).$$

Given the ranges in which μ_1, μ_2 varies, we consider $\Omega_0 = [0, 10]$. We can verify that $\Omega(\mu) \subset \Omega_0$ for all μ. The database is built by considering a uniform sampling of both parameters: we considered $N_1 = 10, N_2 = 40$ values for μ_1, μ_2 respectively. The space resolution on Ω_0 was fixed to $N_x = 512$.

The results are reported in Table 2. In this case, the solution is less critical for POD, and we see that the benefit, when the tolerance is 10^{-1} or 10^{-2} is limited. However, when asking for a better accuracy, we see that the scaling of the memory compression ratio κ and the reconstruction speed-up χ with the accuracy is significant. When considering $\varepsilon = 10^{-4}$, we can reconstruct the snapshots 15 times faster than by using POD.

4.2 2d test cases.

4.2.1 Parametric deformation of a circular characteristic function.

In this section we consider the following problem. Let $\Omega = [0, 1]^2$, $(x_1^{(c)}, x_2^{(c)}) = (1/2, 1/2)$, $r = 1/4$, $\mu_1 \in [0.01, 0.04], \mu_2 \in \{1, \ldots, 6\}$ and $\mu_3 \in [0, \pi/2]$. The set of functions to be approximated, parametrised by μ is the following. Let $f : \mathbb{R}^2 \times \Omega_\mu \to \mathbb{R}$ be defined as:

$$q(x; \mu) = \left(x_1 - x_1^{(c)}\right)^2 + \left(x_2 - x_2^{(c)}\right)^2 - r^2 - \mu_1 \sin(\mu_2 \vartheta + \mu_3),$$

where $\theta = \arctan((x_2 - x_2^{(c)})/(x_1 - x_1^{(c)}))$. The characteristic functions are defined as:

$$u(x; \mu) = \begin{cases} 1 \text{ if } q(x, \mu) \leq 0 \\ 0 \text{ if } q(x, \mu) > 0 \end{cases}$$

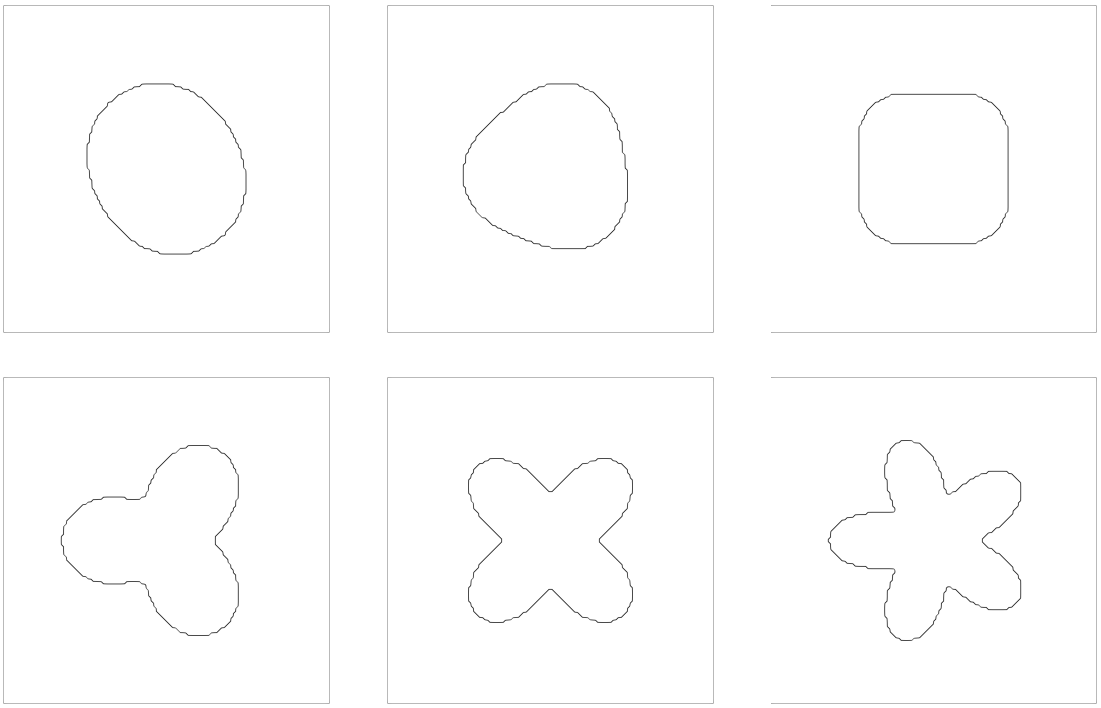

Figure 1: Example of six instances of the domain shapes for the parametric problem detailed in Section 4.2.1, for different values of the parameters. The solutions are the characteristic functions of these domains.

ε	n_*	n	κ	χ	n_0
10^{-1}	122	31	9.10	19.86	3
10^{-2}	1277	30	38.82	163.20	2
10^{-3}	1574	30	44.87	196.15	2

Table 3: Values of the number of modes $n*, n, n_0$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the parametric map of a circular characteristic function presented in Section 4.2.1.

In Fig. 1 six instances of the domain shapes are shown, for different values of the parameters. The result of the reduction are shown in Table 3. We can see that the non-linear reduction is beneficial both in terms of the memory used and in terms of the reconstruction speed-up. In particular, we can see that the application of the pair encoder-decoder is almost 200 times faster when the non-linear reconstruction is used. In the cases considered, the best performances are obtained for roughly $n = 30$. This is much smaller than n_* in this case. Indeed, for $\varepsilon = 10^{-2}$, in order to get the prescribed accuracy, we would need more than 1000 POD modes.

4.2.2 Parametric family of Poisson solutions in varying domain.

Let $r \in [0, R]$, $\alpha \in \mathbb{R}^+$. We consider the solution of the following Poisson problem in cylindrical coordinates:

$$\frac{1}{r}\partial_r(r\partial_r\tilde{u}(r)) = \alpha(1-r),$$

with the condition: $\tilde{u}(R) = 0$. The analytical solution reads:

$$\tilde{u}(r) = \alpha\left[\frac{(R^3 - r^3)}{9} - \frac{(R^2 - r^2)}{4}\right].$$

In order to generate the set of parametric solutions, we proceed in two steps. First, we consider, in cartesian coordinates, simple translation of the above-written solution:

$$r^2(\tilde{x}_1, \tilde{x}_2) = \left(\tilde{x}_1 - x_1^{(c)}\right)^2 + \left(\tilde{x}_2 - x_2^{(c)}\right)^2,$$

Let $\mu = (x^{(c)}, R)$, with $x^{(c)} \in [-0.05, 0.05]^2$ and $R \in [0.15, 0.5]$, $\alpha = 1$. This gives us:

$$\tilde{u}(x;\mu) = \alpha\left[\frac{R^3 - \left(\left(\tilde{x}_1 - x_1^{(c)}\right)^2 + \left(\tilde{x}_2 - x_2^{(c)}\right)^2\right)^{3/2}}{9} - \frac{R^2 - \left(\left(\tilde{x}_1 - x_1^{(c)}\right)^2 + \left(\tilde{x}_2 - x_2^{(c)}\right)^2\right)}{4}\right].$$

The second step consists in using a Joukowski conformal transformation. In particular, let $w, z \in \mathbb{C}$, the Joukowski transformation reads: $w = \frac{1}{2}\big(z + 1/z\big)$. This classical transformation maps circles centred around the origin into ellipses. Circles centred around different points results in droplets like shapes. This has been used in aerodynamics to compute the solution of ideal flows around wing profiles. We will denote the inverse of the Joukowski map $J^{-1} : \mathbb{C} \to \mathbb{C}$. We set: $z = \tilde{x}_1 + i\tilde{x}_2$ and $w = x_1 + ix_2$. We have:

$$w = J^{-1}(z), \quad x_1 = \mathrm{Re}(w), x_2 = \mathrm{Im}(w).$$

With this notation, let $\Omega = [-3, 3]^2$, the solution $u(x;\mu)$ has the following expression:

$$r^2(x) = \left(\mathrm{Re}(J^{-1}(x_1 + ix_2)) - x_1^{(c)}\right)^2 + \left(\mathrm{Im}(J^{-1}(x_1 + ix_2)) - x_2^{(c)}\right)^2$$

$$u(x;\mu) = \alpha\left[\frac{R^3 - r(x)^3}{9} - \frac{R^2 - r(x)^2}{4}\right].$$

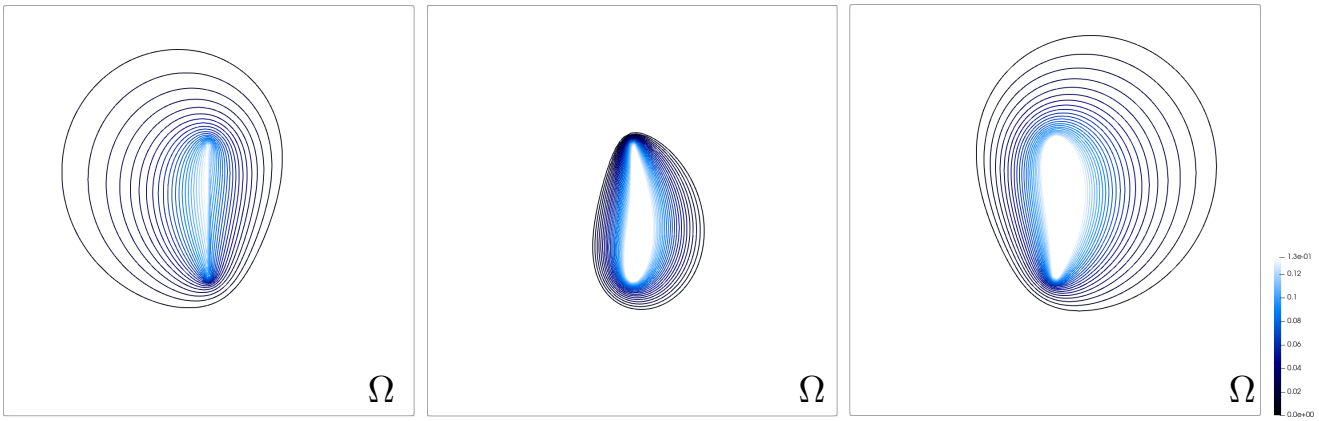

Figure 2: Example of three instances of the solution of the parametric problem detailed in Section 4.2.2, for different values of the parameters.

ε	n_*	n	κ	χ
10^{-1}	3	3	1.0	1.0
10^{-2}	15	10	2.83	3.81
10^{-3}	173	74	5.63	13.49

Table 4: Values of the number of modes $n*, n$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the parametric Poisson problem presented in Section 4.2.2.

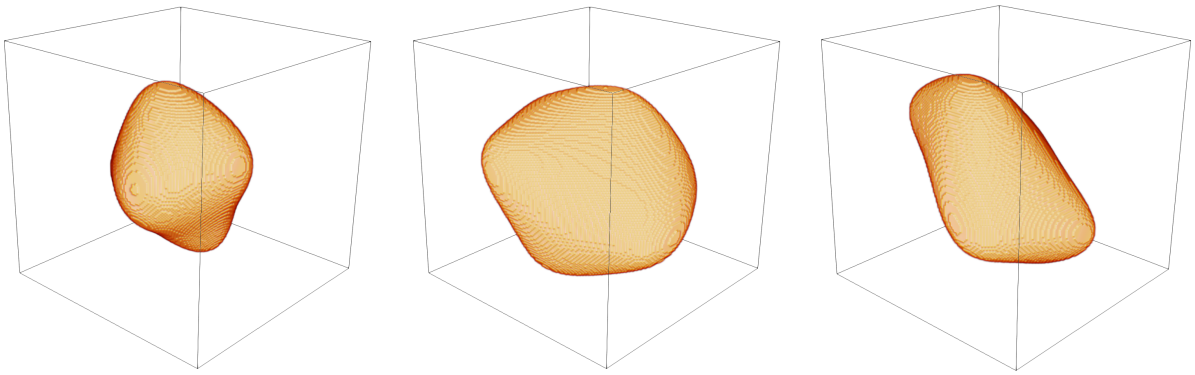

Figure 3: Example of three instances of the solution of the parametric problem detailed in Section 4.3.1, for different values of the parameters. The deformation has been magnified by a factor 10 to better visualise the kind of deformation imposed.

Three examples of solution, for different values of the parameters, are shown in Figure 2.

The results are reported in table 4. We can see that, when asking for a tolerance of $\varepsilon = 10^{-3}$, we get a compression ratio of more than 5 with respect to the POD, and a reconstruction which is one order of magnitude faster.

4.3 3d test cases.

4.3.1 Parametric mapping of a 3d sphere.

In this section we propose a 3d experiment. The set of characteristic functions is generated as follows. Let $\Omega = [0,1]^3$ and $\mu \in [-0.1, 0.1]^3$. We set:

$$\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} x + \mu_3 \sin(2\pi z) \\ y - 0.1z + \mu_1 \sin(2\pi x) \\ z + 0.1x + \mu_2 \sin(2\pi y) \end{bmatrix}$$

Synthetically, we denote this mapping action $\tilde{x} = \xi(x;\mu)$. We then introduce the function q defined as:

$$q(\tilde{x}) = (\tilde{x_1} - 0.5)^2 + (\tilde{x_2} - 0.5)^2 + (\tilde{x_3} - 0.5)^2.$$

The characteristic function is defined as:

$$u(x;\mu) = \begin{cases} 1 \text{ if } q(\xi(x;\mu)) \leq 0 \\ 0 \text{ if } q(\xi(x,\mu)) > 0 \end{cases}$$

We consider $N_x = N_y = N_z = 128$. Three instances of characteristic functions are shown in Figure 3.

ε	n_*	n	κ	χ
10^{-1}	4	4	1.0	1.0
10^{-2}	458	24	162.4	479.0
10^{-3}	512	24	170.8	527.3

Table 5: Values of the number of modes $n*, n, n_0$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the parametric map of a circular characteristic function presented in Section 4.3.1.

The results of the tests are reported in Table 5. As we can see, already when asking for $\varepsilon = 10^{-2}$ the POD does not offer any advantage with respect to storing the full solution, as we need $n_* = 458$ modes. The compression ratio with respect to the POD in this case is more than two orders of magnitude, and the speed-up of the reconstruction of the database snapshots reaches roughly 500.

4.4 Parametric solution of a Poisson problem in a 3d varying domain.

In this last test case, we consider the solution of a Poisson problem in a moving 3d domain. Let the reference domain be $\overline{\Omega}_\xi = [-1, 1]^3$. Let $\left\{q^{(i)}\right\}_{1\leq i\leq 3} : \Omega_\xi \to \mathbb{R}$ be defined as: $q^{(i)}(\xi_i) = 1 - \xi_i^2$. Let the function $q : \Omega_\xi \to \mathbb{R}$ be defined as:

$$q(\xi) = 2\left[q^{(2)}(\xi_2)q^{(3)}(\xi_3) + q^{(1)}(\xi_1)q^{(3)}(\xi_3) + q^{(1)}(\xi_1)q^{(2)}(\xi_2)\right].$$

We consider the following Poisson problem: find $\tilde{u} \in H_0^1(\Omega_\xi)$ such that:

$$-\Delta_\xi \tilde{u} = q \ \text{ in } \Omega_\xi,$$

$$\tilde{u} = 0 \ \text{ on } \ \partial_{\Omega_\xi}.$$

The reference solution of the problem is the following:

$$\tilde{u}(\xi) = q^{(1)}(\xi_1)q^{(2)}(\xi_2)q^{(3)}(\xi_3).$$

Let $\Omega_\mu = [0, \pi/4]^3$, we introduce the following parametric rotation matrix $R : \Omega_\mu \to \mathbb{R}^{3\times 3}$, defined as a composition of 3 elementary rotations:

$$R_1(\mu_1) = \begin{bmatrix} \cos(\mu_1) & -\sin(\mu_1) & 0 \\ \sin(\mu_1) & \cos(\mu_1) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$R_2(\mu_2) = \begin{bmatrix} \cos(\mu_2) & 0 & \sin(\mu_2) \\ 0 & 1 & 0 \\ -\sin(\mu_2) & 0 & \cos(\mu_2) \end{bmatrix},$$

ε	n_*	n	κ	χ
10^{-1}	3	3	1.0	1.0
10^{-2}	65	38	3.09	5.35
10^{-3}	391	330	2.67	5.30

Table 6: Values of the number of modes $n*, n$, the corresponding compression ratio κ and the reconstruction speed-up χ as function of the prescribed accuracy for the 3d Poisson parametric solution presented in Section 4.4.

$$R_3(\mu_3) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\mu_3) & -\sin(\mu_3) \\ 0 & \sin(\mu_3) & \cos(\mu_3) \end{bmatrix}.$$

The parametric matrix R is defined as: $R(\mu) = R_1(\mu_1)R_2(\mu_2)R_3(\mu_3)$. We introduce the following parametric mapping of coordinates:

$$x = R(\mu)\xi, \quad \xi = R^T(\mu)x.$$

The parametric Poisson problem is defined as follows. Let $\Omega_x(\mu) = R(\mu)(\Omega_\xi) \subset \Omega$, with $\Omega = [-\sqrt{3}, \sqrt{3}]^3$, where with a slight abuse of notation we denoted the transformation with the matrix for a given instance of the parameter. Let the function $q_x : \Omega \times \Omega_\mu \to \mathbb{R}$ be defined as:

$$q_x(x; \mu) = q(R^T(\mu)x).$$

We want to find $u \in H_0^1(\Omega) \otimes L^2(\Omega_\mu)$ such that:

$$-\Delta_x u = q_x, \quad \text{in } \Omega_x(\mu),$$

$$u = 0 \text{ on } \partial\Omega_x(\mu).$$

The solution of the problem is the function u defined as:

$$u(x; \mu) = \tilde{u}([R(\mu)]^T x).$$

Remark that this is true because the rotation matrix defines a unitary transformation, and the Laplacian operator is invariant under unitary transformations.

The results are reported in Table 4.4. Let us remark that, in this case, the gain is roughly constant at $\varepsilon = 10^{-2}, 10^{-3}$ both in terms of memory compression and in terms of number of elementary operations to perform the reconstruction.

5 Conclusion and perspectives.

In the present work a non-linear manifold reduction method is presented, which consists of computing a dof-wise ridge regression to improve the performances of the linear reduction obtained by means of the classical POD method. In this method, the

non-linearity of the ridge function is not fixed a priori. Instead, a greedy method is introduced, such that, for every degree of freedom, the non-linear function is built progressively by means of adaptive $\mathbb{P}_3$-Hermite finite elements (hence it belongs to the space of piece-wise third order polynomials). Some theoretical results are presented, in which we could characterise the error with respect to the one obtained by the linear reconstruction and, depending on the properties of the set of functions to be represented, achieve in some cases a significant improvement of the performances (in terms of memory burden at fixed tolerance). The numerical tests showed that this is true for the class of characteristic functions of sets, which are changing in a neighbourhood of a reference configuration. This opens interesting perspectives for the use of the present method for the reduction of fluid-structure interaction problems. Another perspective concerns the investigation of a more general way of constructing a composition starting from the linear reconstruction, in order to better adapt to different classes of parametric problems.

References

[1] Mark J Ablowitz and Harvey Segur. *Solitons and the inverse scattering transform*. SIAM, 1981.

[2] Joshua Barnett and Charbel Farhat. Quadratic approximation manifold for mitigating the kolmogorov barrier in nonlinear projection-based model order reduction. *Journal of Computational Physics*, 464:111348, 2022.

[3] Joshua Barnett, Charbel Farhat, and Yvon Maday. Neural-network-augmented projection-based model order reduction for mitigating the kolmogorov barrier to reducibility. *Journal of Computational Physics*, 492:112420, 2023.

[4] Peter Benner, Mario Ohlberger, Albert Cohen, and Karen Willcox. *Model reduction and approximation: theory and algorithms*. SIAM, 2017.

[5] Dietrich Braess. *Nonlinear approximation theory*, volume 7. Springer Science & Business Media, 2012.

[6] Matthew Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1):20–30, 2006.

[7] Patrick Buchfink, Silke Glas, and Bernard Haasdonk. Approximation bounds for model reduction on polynomially mapped manifolds. *arXiv preprint arXiv:2312.00724*, 2023.

[8] Albert Cohen and Ronald DeVore. Kolmogorov widths under holomorphic mappings. *IMA Journal of Numerical Analysis*, 36(1):1–12, 2016.

[9] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

[10] Albert Cohen, Charbel Farhat, Agustín Somacal, and Yvon Maday. Nonlinear compressive reduced basis approximation for pde's. 2023.

[11] Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

[12] Stefania Fresca, Luca Dede', and Andrea Manzoni. A comprehensive deep learning-based approach to reduced order modeling of nonlinear time-dependent parametrized pdes. *Journal of Scientific Computing*, 87:1–36, 2021.

[13] Rudy Geelen, Laura Balzano, and Karen Willcox. Learning latent representations in high-dimensional state spaces using polynomial manifold constructions. *arXiv preprint arXiv:2306.13748*, 2023.

[14] Rudy Geelen, Stephen Wright, and Karen Willcox. Operator inference for non-intrusive model reduction with quadratic manifolds. *Computer Methods in Applied Mechanics and Engineering*, 403:115717, 2023.

[15] Kookjin Lee and Kevin T Carlberg. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *Journal of Computational Physics*, 404:108973, 2020.

[16] Mario Ohlberger and Stephan Rave. Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing. *Comptes Rendus Mathematique*, 351(23-24):901–906, 2013.

[17] Tohru Yoneyama. The korteweg-de vries two-soliton solution as interacting two single solitons. *Progress of theoretical physics*, 71(4):843–846, 1984.