Preconditioners for multilinear problems arising in parametric Partial Differential Equations

Damiano Lombardi, Sébastien Riffaud

Preconditioners for multilinear problems arising in parametric Partial Differential Equations.

Damiano Lombardi[1] and Sébastien Riffaud[1]

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

Abstract

This work deals with the set up of preconditoning methods for multilinear problems emerging from the discretisation of parametric Partial Differential Equations. We are going to focus, here, on linear problems. We will investigate three different preconditioning strategies. Some theoretical results are presented in order to assess the properties of these preconditioners and understand their limitations. We present several numerical experiments on three-dimensional parametric PDEs.

1 Introduction.

When Partial Differential Equations (PDEs) are discretised following a separation of variable principle, they lead to a discrete multilinear problem to be solved. The most simple way to write the separation of variable principle is the so-called canonical polyadic format, which express a multivariate function as a sum of products of univariate functions. When the number of variables is larger than 2, more stable ways to deal with the separation of variables have been introduced. The use of tensor networks and hierarchical tensor format for the approximation of PDEs solution is discussed in [5, 3] and the references therein. Numerous methods have been proposed and investigated in order to approximate the solution of multilinear problems. In [20] a greedy method to progressively approximate the solution is described, and several methods to solve linear problems when the solution is represented in a canonical polyadic format are detailed in [8]. An energy minimisation method is proposed in [12]. In [9], a Bi-CG method is proposed, in [4] an iterative thresholding method within a Richardson-type iteration is investigated, which proved to be numerically efficient. A method to address a specific class of multilinear problems (whose solutions are third order tensors) was proposed in [22]. Solvers based on a geometrical (Riemannian) description of low-rank format are investigated in [23]. In analogy with Krylov solvers for linear systems, Krylov

methods for multilinear problems are investigated in [18], and in other contributions, [6, 11, 21, 10]. A survey is proposed in [15].

As for linear systems, a preconditioning technique could be introduced to improve the resolution of multilinear systems. For instance, in [19] a SOR preconditioner is investigated. In [17] two methods are investigated: a Riemannian version of the preconditioned Richardson method as well as an approximate Newton scheme based on the Riemannian Hessian. In [13] a semi-definite programming method is describe to approximate operators as Kronecker products. Several contributions in the literature deal with building preconditioners using machine learning methods [14, 1, 2, 16]. In the present work, we try to define a priori the preconditioner, and to compute it without the use of an *offline* phase.

We are going to consider Krylov based solvers, in the spirit of [18], and we are going to describe and compare three possible preconditioning methods. In [24] several preconditioners are studied to enable the efficient resolution of stochastic elliptic PDEs. A strategy which is investigated in this work is based on deflation; the strategy based on the Voronoi quantisation is similar, in the spirit, to one of the strategies we are going to test. A difference is that, in the present work, we consider a general parametric linear problem (not necessary elliptic), and we focus on devising methods dealing with a low-rank approximation of the parametric solution.

The structure of the work is as follows: in Section 2 we introduce the notation and the problem we are going to focus on; in particular, we are going to define the three different preconditioning strategies we are going to study. In Section 3 we propose some theoretical results to get some insight on the different strategies, their properties and their limitations. In Section 4 we propose several numerical experiments on three-dimensional PDEs.

2 Notation and problem setting.

Let $d \in \mathbb{N}^*$ and $\Omega \subseteq \mathbb{R}^d$, where typically $d = 2, 3$. Let $p \in \mathbb{N}^*$ and $\Theta \subset \mathbb{R}^p$ be a compact set, the parametric domain. Let $u \in V \otimes W^{1,k}(\Theta)$ (V being a Hilbert space and $W^{1,k}$ being the standard Sobolev space $(1, k)$) be a real (eventually vector valued) function defined as:

$$u : \left\{ \begin{array}{l} \Omega \times \Theta \ \rightarrow \ \mathbb{R}^m \\ (x, \vartheta) \ \mapsto u(x, \vartheta) \end{array} \right. \tag{1}$$

Let $L \in \mathbb{N}^*$. Let $\{\mathcal{A}_i\}_{1 \leq i \leq L} \in \mathcal{L}(V, W_i)$ be a set of linear operators. In what follows, we are going to assume that there exists a function space W such that $W_i \hookrightarrow W$, $1 \leq i \leq L$. Let $\{g_i\}_{1 \leq i \leq L} \in W$ be a set of real valued measurable functions of the parameters. Let $f \in W \otimes W^{1,k}(\Theta)$. A (eventually semi-discretised in time) parametric PDE can be written as:

$$\big[\sum_{l=1}^{L} g_l(\vartheta)\mathcal{A}_l\big] u = f.$$

Let us remark that numerous systems of parametric PDEs can be written in this form. We provide hereafter some examples.

Example 1: Let us consider a Poisson problem with linear reaction. Let $\Omega \subset \mathbb{R}^3$ and $\Theta \subset \mathbb{R}^2$. Let $u \in H_0^1(\Omega) \otimes L^2(\Theta)$ be the solution of:

$$-\vartheta_1 \Delta_x u + \vartheta_2 u = f(x; \vartheta_1, \vartheta_2), \quad \text{in } \Omega \times \Theta$$

$$u = 0, \quad \text{on } \partial\Omega \times \Theta.$$

Numerous examples arising in all areas of mathematical modelling can be cast in this form. Even when considering time dependent problems, or non-linear problems, a semi-discretisation in time or a fixed point iteration may result in a problem written in this form.

Let us introduce a discretisation of the parametric PDEs written above. Let $\{v_i\}_{1 \leq i \leq N_x} \in V_{N_x} \subset V$ be a basis of V_{N_x}, dense in V. Concerning the discretisation in the parameter space, several options are available. Assuming that the dimension of Θ can be large, and for the sake of simplicity in the notation, we adopt a collocation discretisation in ϑ. This is not necessary, per se, and the use of other forms of discretisation in the parameter direction is straightforward. Let $\left\{\vartheta^{(j)}\right\}_{1 \leq j \leq N_\vartheta} \in \Theta$ be a set of parameter instances. The weak formulation in space of the parametric system consists in finding the solution $u \in V \otimes W^{1,k}$ such that, for all $v \in V$ and all $\vartheta \in \Theta$:

$$\sum_{l=1}^{L} g_l(\vartheta) \langle \mathcal{A}_l u, v \rangle = \langle f, v \rangle. \tag{2}$$

When introducing the discretisation, we particularise the variational formulation of the problem by introducing the discretisation, asking that, for all $v \in V_{N_x}$ and all $\left\{\vartheta^{(j)}\right\}_{1 \leq j \leq N_\vartheta}$:

$$\sum_{i=l}^{L} g_l(\vartheta^{(j)}) \langle \mathcal{A}_l u_N, v \rangle = \langle f, v \rangle.$$

Here, we have introduced the notation u_N to denote the approximation of u. Let $\hat{u} \in \mathbb{R}^{N_x \times N_\vartheta}$, u_N can be written as:

$$u_N(\cdot, \vartheta^{(j)}) = \sum_{i=1}^{N_x} \hat{u}_{ij} v_i.$$

When inserting this into the weak formulation we get:

$$\sum_{i=l}^{L} g_l(\vartheta^{(j)}) \langle \mathcal{A}_l \sum_{i=1}^{N_x} \hat{u}_{ij} v_i, v_m \rangle = \langle f, v_m \rangle, \ 1 \leq m \leq N_x, \ 1 \leq j \leq N_\vartheta. \tag{3}$$

By virtue of the linearity of the operators $\mathcal{A}_l$, the expression can be cast as:

$$\sum_{i=l}^{L} g_l(\vartheta^{(j)}) \sum_{i=1}^{N_x} \langle \mathcal{A}_l v_i, v_m \rangle \hat{u}_{ij} = \langle f, v_m \rangle.$$

Let $A^{(l)} \in \mathbb{R}^{N_x \times N_x}$ be real valued matrices and $D^{(l)} \in \mathbb{R}^{N_\vartheta \times N_\vartheta}$ be diagonal real valued matrices, whose components are defined as:

$$A^{(l)}_{mi} = \langle \mathcal{A}_l v_i, v_m \rangle,$$

$$D^{(l)}_{jj} = g_l(\vartheta^{(j)}).$$

Moreover, let $F \in \mathbb{R}^{N_x \times N_\vartheta}$ be the matrix whose components are defined as:

$$F_{mj} = \langle f(\vartheta^{(j)}), v_m \rangle.$$

The Eq.(3) can be written as:

$$\sum_{i=l}^{L} \sum_{i=1}^{N_x} A^{(l)}_{mi} \hat{u}_{ij} D^{(l)}_{jj} = F_{mj}, \ 1 \leq m \leq N_x, \ 1 \leq j \leq N_\vartheta.$$

We introduce the Kronecker product of matrices and rewrite synthetically this system as:

$$\sum_{l=1}^{L} \left[A^{(l)} \otimes D^{(l)} \right] \hat{u} = F.$$

This is a multi-linear system which, in this case, results from the discretisation of the parametric PDE. In the following, we denote:

$$\mathbf{A} = \sum_{l=1}^{L} \left[A^{(l)} \otimes D^{(l)} \right].$$

This is a tensor operator with L terms also sometimes refereed to as rank-L tensor operator.

Since $\hat{u}$ is a matrix with a large number of entries (making its computation and full storage prohibitive), we are going to approximate it in a low-rank format. We consider in this work the following separated discretisation of the solution. Let $n \in \mathbb{N}^*$. Let $\{\varphi_k\}_{1 \leq k \leq n} \in V_{N_x}$ and $\{\psi_k\}_{1 \leq k \leq n} \in W^{1,k}$. The solution approximation is written as:

$$u_N = \sum_{k=1}^{n} \varphi_k \otimes \psi_k.$$

The single term $\varphi_k \otimes \psi_k$ is also called rank-1 tensor or pure tensor term. When the solution is evaluated in (x, ϑ), it is intended that: $u_N(x, \vartheta) = \sum_{k=1}^{n} \varphi_k(x) \psi_k(\vartheta)$. The functions φ_k can be written as:

$$\varphi_k = \sum_{i=1}^{N_x} \Phi_{ik} v_i,$$

where $\Phi \in \mathbb{R}^{N_x \times n}$ is the matrix containing the values of the degrees of freedom of the functions φ_k. When discretising the functions ψ_k by a collocation approach we get:

$$\Psi_{jk} = \psi_k(\vartheta^{(j)}).$$

And hence:

$$u_N(x, \vartheta^{(j)}) = \sum_{k=1}^{n} \sum_{i=1}^{N_x} \Phi_{ik} v_i(x) \Psi_{jk}.$$

By identification we can write:

$$\hat{u}_{ij} = \sum_{k=1}^{n} \Phi_{ik} \Psi_{kj}^T,$$

which is the low-rank approximation of the degrees of freedom $\hat{u}$. Solving the problem amounts to compute the matrices Φ and Ψ.

If the right-hand side F can be expressed in a low-rank format (which we are going to assume), the problem residual can be written in a low-rank format. This makes it possible to exploit GMRES-like iterations to solve in an iterative way the multi-linear problem. When the operator is in Kronecker format, its action can be evaluated as:

$$\left[A^{(l)} \otimes D^{(l)}\right] \Phi \otimes \Psi = A^{(l)} \Phi \otimes D^{(l)} \Psi,$$

which highlights what is the clear advantage of introducing a low-rank approximation of the degrees of freedom $\hat{u}$.

The problem can be interpreted as follows: given a linear system of parametric PDEs, its discretisation can be written as a linear system of size $N_x \cdot N_\vartheta$, the system matrix being $\mathbf{A} \in \mathbb{R}^{N_x \cdot N_\vartheta \times N_x \cdot N_\vartheta}$. When the PDE has the structure detailed above, the matrix can be factored as a sum of Kronecker products. We are going to leverage this interpretation in order to propose several preconditioning strategies. Classically, we are going to build approximations of $\mathbf{A}^{-1}$, the inverse of the matrix $\mathbf{A}$. Of course, these matrices are not computed nor stored at any stage of the computation. Instead, their Kronecker product structure is exploited.

2.1 Preconditioners

Three kind of preconditioners are described.

Preconditioner type I: the first preconditioner has been used in [7] and it consists in evaluating the operator in one single instance of the parameters, called $\overline{\vartheta} \in \Theta$. Let $P \in \mathbb{R}^{N_x \times N_x}$, defined as:

$$P = \sum_{l=1}^{L} g_l(\overline{\theta}) A^{(l)}.$$

Since the original problem admits a solution for all $\vartheta \in \Theta$, the matrix P is invertible and $P^{-1}v$, for $v \in \mathbb{R}^{N_x}$ costs the resolution of a linear system of size N_x. In order

to solve this system, the classical methods (also adapted to the problem at hand) can be used. Remark that this preconditioner can be written in multilinear format straightforwardly: let $I_{N_\vartheta} \in \mathbb{R}^{N_\vartheta \times N_\vartheta}$ be the identity matrix,

$$\mathbf{P}^{(I)} = P^{-1} \otimes I_{N_\vartheta},$$

Let $v \in \mathbb{R}^{N_x}$ and $s \in \mathbb{R}^{N_\vartheta}$, we have:

$$\left[P^{-1} \otimes I_{N_\vartheta}\right] v \otimes s = P^{-1} v \otimes s.$$

Preconditioner type II: the second type of preconditioner can be seen as a generalisation of the first type of preconditioner. Instead of considering one single parametric instance, we are going to consider a set of instances $\mathcal{Q} = \left\{\vartheta^{(q)}\right\}_{1 \leq q \leq Q} \in \Theta$ and construct:

$$P^{(q)} = \sum_{l=1}^{L} g_l(\theta^{(q)}) A^{(l)}$$

All the matrices $P^{(q)}$ are invertible. Let the $q-$th Voronoi cell be defined as:

$$\mathcal{V}^{(q)} = \left\{\vartheta \in \Theta \text{ such that } |\vartheta^{(q)} - \vartheta| < |\vartheta^{(p)} - \vartheta|, \text{ for all } \vartheta^{(p)} \in \mathcal{Q}\right\}.$$

We are going to denote c_q the characteristic function of the $q-$th voronoi cell, and, henceforth:

$$c_q(\vartheta) = 1, \ \vartheta \in \mathcal{V}^{(q)}.$$

Intuitively, we would like to solve the linear system defined by the $q-$th matrix $P^{(q)}$ in all the parametric instances for which $\vartheta^{(q)}$ is the closest parametric instance.

Let $C^{(q)} \in \mathbb{R}^{N_\vartheta \times N_\vartheta}$, $1 \leq q \leq Q$ the diagonal matrix whose components are defined as:

$$C^{(q)}_{jj} = c_q(\vartheta^{(j)}).$$

The preconditioner type II has the following expression:

$$\mathbf{P}^{(II)} = \sum_{q=1}^{Q} \left[P^{(q)}\right]^{-1} \otimes C^{(q)}$$

Let $v \in \mathbb{R}^{N_x}$ and $s \in \mathbb{R}^{N_\vartheta}$, we have:

$$\sum_{q=1}^{Q} \left[\left[P^{(q)}\right]^{-1} \otimes C^{(q)}\right] v \otimes s = \sum_{q=1}^{Q} \left[P^{(q)}\right]^{-1} v \otimes C^{(q)} s.$$

The computational cost of evaluating this preconditioner on a rank-one term is equivalent to solving Q linear systems of size N_x and applying a diagonal matrix Q times on s.

Preconditioner type III: the third type of preconditioner is based on a compositional approach, in which a sequence of rank-one operators is applied to the input tensor. We observe that building and applying rank-one operators to tensors is equivalent to solving linear systems in each of the directions separately. This is relatively advantageous from a computational point of view. Let $Q \in \mathbb{N}^*$ be the number of preconditioning steps. Let us describe the first step of the procedure. The input tensor to which we would like to apply the preconditioner is $R = \sum_{k=1}^{n_R} r_k^{(x)} \otimes r_k^{(\vartheta)}$.

We are going to build a sequence of single term tensor operators which approximates the action of A on R. Let $\{\beta_l\}_{1 \leq l \leq L} \in \mathbb{R}$ and $\{\gamma_l\}_{1 \leq l \leq L} \in \mathbb{R}$ be two sets of coefficients. We define:

$$P^{(x)} = \sum_{l=1}^{L} \beta_l A^{(l)}$$

and

$$P^{(\vartheta)} = \sum_{l=1}^{L} \gamma_l D^{(l)}.$$

The rank-one operator takes the form:

$$\mathbf{P}^{(q)} = [P^{(x),q}]^{-1} \otimes [P^{(\vartheta),q}]^{-1},$$

where the superscript $q \in \mathbb{N}, q \geq 0$ denotes the iteration. The method reads as follows:

$$Z^{(q)} = R - \mathbf{A} R^{(q)},$$

$$R^{(q+1)} = R^{(q)} + \mathbf{P}^{(q)} Z^{(q)},$$

and $\mathbf{P}^{(q)}$ is computed in such a way that the application on the operator $\mathbf{A}$ to the residuals $Z^{(q)}$ is approximated at best. Remark that, at the first iteration: $Z^{(0)} = R$.

Putting these together, we obtain:

$$R^{(q+1)} = \left[\mathbf{I} - \mathbf{P}^{(q)} \mathbf{A}\right] R^{(q)} + \mathbf{P}^{(q)} R.$$

Formally, this relation makes it possible to define, by composition, the preconditioner:

$$\mathbf{P}^{(III),Q} = \left[\mathbf{I} - \mathbf{P}^{(Q-1)} \mathbf{A}\right] \mathbf{P}^{(III),Q-1} + \mathbf{P}^{Q-1}.$$

Remark that this preconditioner is not assembled in this form. Instead, the sequence of rank-one operators $\mathbf{P}^{(q)}$ is identified and applied to the input tensor and to the elements $R^{(q)}$.

Let us detail a possible way to compute the sequence of preconditioning steps, leading to a simplification, *i.e.* matrices $P^{(x)}$ and $P^{(\vartheta)}$ which do not depend on the iteration q.

Let $\mathbf{v}_x \in \mathbb{R}^{N_x}$ and $\mathbf{v}_\vartheta \in \mathbb{R}^{N_\vartheta}$ be two vectors whose and

$$R^{(x)} = \sum_{k=1}^{n_R} r_k^{(x)} \otimes \mathbf{v}_\vartheta,$$

$$R^{(\vartheta)} = \sum_{k=1}^{n_R} \mathbf{v}_x \otimes r_k^{(\vartheta)}.$$

Let $T^{(x)} = \mathbf{A}R^{(x)}$ and $T^{(\vartheta)} = \mathbf{A}R^{\vartheta}$. Their detailed expression reads:

$$T^{(x)} = \sum_{l=1}^{L}\sum_{k=1}^{n_R} A^{(l)} r_k^{(x)} \otimes D^{(l)}\mathbf{v}_\vartheta = \sum_{l=1}^{L}\left(\sum_{k=1}^{n_R} A^{(l)} r_k^{(x)}\right) \otimes D^{(l)}\mathbf{v}_\vartheta,$$

$$T^{(\vartheta)} = \sum_{l=1}^{L}\sum_{k=1}^{n_R} A^{(l)}\mathbf{v}_x \otimes D^{(l)} r_k^{(\vartheta)} = \sum_{l=1}^{L} A^{(l)}\mathbf{v}_x \otimes \left(\sum_{k=1}^{n_R} D^{(l)} r_k^{(\vartheta)}\right).$$

Let $\left\{U_l^{(x)}\right\}_{1\le l\le L}$ and $\left\{U_l^{(\vartheta)}\right\}_{1\le l\le L}$ be defined as:

$$U_l^{(x)} = \sum_{k=1}^{n_R} A^{(l)} r_k^{(x)} \otimes \mathbf{v}_\vartheta,$$

$$U_l^{(\vartheta)} = \sum_{k=1}^{n_R} \mathbf{v}_x \otimes D^{(l)} r_k^{(\vartheta)}.$$

The coefficients $\beta_l,\ 1 \le l \le L$ are computed as the solution of:

$$\{\beta_l^*\}_{1\le l\le L} = \arg\inf_{\beta_l} \|T^{(x)} - \sum_{l=1}^{L} \beta_l U_l^{(x)}\|_F^2,$$

and, analogously, the coefficients $\gamma_l,\ 1 \le l \le L$ are the solution of:

$$\{\gamma_l^*\}_{1\le l\le L} = \arg\inf_{\gamma_l} \|T^{(\vartheta)} - \sum_{l=1}^{L} \gamma_l U_l^{(\vartheta)}\|_F^2.$$

This defines one single step of the preconditioner:

$$\mathbf{P}^{(0)} = P^{(x)} \otimes P^{(\vartheta)}.$$

3 Analysis of the preconditioners.

In this section we analyse the preconditioning strategies detailed in the previous section. In particular, we are going to consider the following hypotheses:

1. The parameter set $\Theta \subset \mathbb{R}^p$ is a compact set. This entails that there exists a constant $\varrho_* > 0$ (the Chebyshev radius) and a point $\vartheta^{(0)}$ (the Chebyshev centre) such that for all $\vartheta \in \Theta$, $|\vartheta - \vartheta^{(0)}| \le \varrho_*$.

2. For the present analysis we start from the semi-discretised in x formulation and consider the matrix field: $A : \Theta \to \mathbb{R}^{N_x \times N_x}$ which is defined by:

$$A(\vartheta) = \sum_{l=1}^{L} g_l(\vartheta) A_l.$$

We make the hypothesis that, for all $\vartheta \in \Theta$, the matrix $A(\vartheta)$ is invertible.

3. The functions g_l are Hölder continuous, and, in particular $g_l \in \mathcal{C}^{0,\alpha}(\Theta)$, $1 \leq l \leq L$, $0 < \alpha < 1$, where we denote $|\cdot|$ the Euclidean norm.

3.1 Analysis of the preconditioners (I) and (II).

Let us first consider the Preconditioners type (I) and (II). We are going to analyse the first kind of preconditioner, and the result of the second will follow. This is also compatible to the fact that the first kind of preconditioner can be interpreted as a particular case of the second one.

To analyse the effect of the preconditioners (I) on the problem we are going to consider the matrix field as function of the parameters and analyse how close is P to the inverse $A(\vartheta)^{-1}$. Let us recall that $A(\vartheta) \in \mathbb{R}^{N_x \times N_x}$ is the matrix corresponding to the discretisation of the problem for the parametric instance ϑ:

$$A(\vartheta) = \sum_{l=1}^{L} g_l(\vartheta) A^{(l)}.$$

Analogously:

$$P = \sum_{l=1}^{L} g_l(\overline{\vartheta}) A^{(l)}$$

We present a first Lemma bounding the difference of the matrix $A(\vartheta)$ and P.

1. *Lemma:* *Let $\alpha \in (0,1)$ and $g_l \in C^{0,\alpha}(\Theta)$, $1 \leq l \leq L$ be Hölder continuous, with $\{\kappa_l\}_{1 \leq l \leq L}$ being the Hölder continuity constants. Then, there exists a constant b_0 such that, for any matrix norm $\|\cdot\|$ it holds:*

$$\|A(\vartheta) - P\| \leq b_0 \sup_{\vartheta \in \Theta} |\vartheta - \overline{\vartheta}|^{\alpha}$$

where

$$b_0 = \sum_{l=1}^{L} \kappa_l \|A^{(l)}\|.$$

Proof. Let us consider:

$$\|A(\vartheta) - P\| = \|\sum_{l=1}^{L} (g(\vartheta) - g(\overline{\vartheta})) A^{(l)}\|.$$

By virtue of the properties of the norms and of the Hölder continuity it holds:

$$\|\sum_{l=1}^{L}(g_l(\vartheta)-g_l(\overline{\vartheta}))A^{(l)}\| \leq \sum_{l=1}^{L}|g_l(\vartheta)-g_l(\overline{\vartheta})|\|A^{(l)}\| \leq \sum_{l=1}^{L}\kappa_l|\vartheta-\overline{\vartheta}|^{\alpha}\|A^{(l)}\|.$$

From this, we conclude:

$$\|A(\vartheta)-P\| \leq \sup_{\vartheta\in\Theta}|\vartheta-\overline{\vartheta}|^{\alpha}\left(\sum_{l=1}^{L}\kappa_l\|A^{(l)}\|\right).$$

□

From this Lemma it is clear that it is advantageous to take $\overline{\vartheta}$ in such a way as to minimize the maximum distance with respect to all the elements of the set Θ. The optimal choice for $\overline{\vartheta}$ is represented by the Chebyshev centre $\vartheta^{(0)}$ of the set Θ, and the maximum distance would be, in that case, the Chebyshev radius. We are going to denote $\varrho = \sup_{\vartheta\in\Theta}|\vartheta-\overline{\vartheta}|$ and it holds $\varrho \geq \varrho_*$.

Let us consider the singular value decomposition of the matrices $A(\vartheta)$ and P:

$$A(\vartheta) = U(\vartheta)S(\vartheta)V(\vartheta)^T,$$

$$P = \overline{U}\ \overline{S}\ \overline{V}^T.$$

The fact that, for all $\vartheta \in \Theta$ the matrix $A(\vartheta)$ is invertible, implies that, for all ϑ:

$$0 < \sigma_{N_x}(\vartheta) \leq \ldots \leq \sigma_1(\vartheta),$$

and U, V are elements of the unitary group $\mathbf{U}(N_x)$.

2. *Lemma:* *Let the result of Lemma 1 holds. There exists two constants $b_0^{(2)}, b_1 > 0$ such that:*

$$|\sigma_i(\vartheta)-\overline{\sigma}_i| \leq b_0^{(2)}\varrho^{\alpha}+b_1\varrho, \quad 1 \leq i \leq N_x.$$

Proof. For a generic triplet $u(\vartheta), \sigma(\vartheta), v(\vartheta)$ it holds:

$$A(\vartheta)v(\vartheta) = \sigma(\vartheta)u(\vartheta).$$

Analogously,

$$P\overline{v} = \overline{\sigma}\ \overline{u}.$$

We take the difference of the two problems:

$$A(\vartheta)v(\vartheta)-P\overline{v} = \sigma(\vartheta)u(\vartheta)-\overline{\sigma}\ \overline{u}.$$

We add and subtract $Pv(\vartheta)$ and $\overline{\sigma}u(\vartheta)$ on the left and right hand side respectively:

$$(A-P)v(\vartheta)+P(v(\vartheta)-\overline{v}) = (\sigma(\vartheta)-\overline{\sigma})u(\vartheta)+\overline{\sigma}(u(\vartheta)-\overline{u}).$$

We take on both side the Euclidean scalar product with $u(\vartheta)$:

$$u(\vartheta)^T(A(\vartheta)-P)v(\vartheta)+u(\vartheta)^TP(v(\vartheta)-\overline{v})=(\sigma(\vartheta)-\overline{\sigma})+\overline{\sigma}u(\vartheta)^T(u(\vartheta)-\overline{u}).$$

We get:

$$(\sigma(\vartheta)-\overline{\sigma})=u(\vartheta)^T(A(\vartheta)-P)v(\vartheta)+u(\vartheta)^T\left[P(v(\vartheta)-\overline{v})-\overline{\sigma}(u(\vartheta)-\overline{u})\right].$$

Aiming at bounding $|\sigma(\vartheta)-\overline{\sigma}|$ we consider the two terms on the right-hand side separately. Concerning the first one, we exploit the result of Lemma 1 particularising it for the $2-$norm $\|\cdot\|_2$ and write:

$$|u(\vartheta)^T(A(\vartheta)-P)v(\vartheta)|\leq\|A(\vartheta)-P\|_2\leq b_0^{(2)}\varrho^{\alpha},$$

where: $b_0^2=\sum_{l=1}^{L}\kappa_l\|A^{(l)}\|_2$.

Concerning the second term on the right-hand side, let us first recall that the set Θ is compact. Moreover, we observe that $u(\vartheta),v(\vartheta)$ as well as $\overline{u},\overline{v}$ are all elements of the compact Stiefel manifold $\mathcal{M}^{N_x,1}$ (when considering the whole set of left and right singular vectors, we can consider them as elements of the unitary group). The set being simply connected, we can move away from the identity with continuity. Furthermore, when $\vartheta=\overline{\vartheta}$ the second term on the right-hand side vanishes. It is possible to introduce a constant $b_1>0$ such that, for all $\vartheta\in\Theta$:

$$|u(\vartheta)^T\left[P(v(\vartheta)-\overline{v})-\overline{\sigma}(u(\vartheta)-\overline{u})\right]|\leq b_1\varrho.$$

The constant b_1 depends on the singular values $\overline{\sigma}$ of P.

It holds:

$$|\sigma(\vartheta)-\overline{\sigma}|\leq b_0^{(2)}\varrho^{\alpha}+b_1\varrho.$$

□

In view of proving the main results concerning the preconditioners of first and second type, we are going to introduce a further assumption. Let $u_i(\vartheta),\overline{u}$ be the left singular vectors of $A(\vartheta)$ and P respectively. There exist $\{\eta_i\}_{1\leq i\leq N_x}\in\mathbb{R}^+$ such that

$$\overline{u}_i^Tu_i(\vartheta)\geq\exp\left(-\eta_i|\vartheta-\overline{\vartheta}|\right),\quad 1\leq i\leq N_x.$$

Remark that this implies that there exists a $\eta>0$ such that:

$$\overline{u}_i^Tu_i(\vartheta)\geq\exp\left(-\eta_i\varrho\right)\geq\exp\left(-\eta\varrho\right),\quad 1\leq i\leq N_x.$$

1. *Proposition:* *Let Θ be compact, such that, for all $\vartheta\in\Theta$ the matrix $A(\vartheta)$ is invertible. Moreover, let $g_l\in C^{0,\alpha}(\Theta)$, $1\leq l\leq L$ be Hölder continuous. Then, the condition number with respect to the $2-$norm of $P^{-1}A(\vartheta)$ satisfies:*

$$\mathrm{cond}_2\left(P^{-1}A(\vartheta)\right)\leq c_3(\varrho)+c_2(\varrho)(1-e^{-\eta\varrho})^{1/2}+c_0(\varrho)c_1(\varrho)(1-e^{-\eta\varrho}),$$

for some positive constants $c_1,c_2,c_3>0$ depending upon ϱ, such that:

$$\lim_{\varrho\to 0}\mathrm{cond}_2\left(P^{-1}A(\vartheta)\right)=1.$$

Proof. It holds, by definition:

$$\mathrm{cond}_2\left(P^{-1}A(\vartheta)\right) = \|P^{-1}A(\vartheta)\|_2 \|A(\vartheta)^{-1}P\|_2.$$

Let us consider the singular value decomposition of the matrices P^{-1} and $A(\vartheta)$. We have:

$$P^{-1}A(\vartheta) = \overline{V}\ \overline{S}^{-1}\overline{U}^T U S V^T.$$

When considering the $2-$norm, by virtue of the fact that $\overline{V}$ and V are unitary matrices, we can write:

$$\|P^{-1}A(\vartheta)\|_2 = \|\overline{S}^{-1}\overline{U}^T U S\|_2.$$

Analogously:

$$\|A(\vartheta)^{-1}P\|_2 = \|S^{-1}U^T\overline{U}\ \overline{S}\|_2.$$

We are going to use the Gershgorin theorem to provide bounds on the absolute value of the eigenvalues. Let's start by the matrix $\overline{S}^{-1}\overline{U}^T U S$. Its components read:

$$[\overline{S}^{-1}\overline{U}^T U S]_{ij} = \frac{1}{\overline{\sigma}_i}[\overline{U}^T U]_{ij}\sigma_j.$$

The diagonal entries are:

$$[\overline{S}^{-1}\overline{U}^T U S]_{ii} = \frac{\sigma_i}{\overline{\sigma}_i}[\overline{U}^T U]_{ii}.$$

By using the result of Lemma 2 and the hypothesis (H1) we get:

$$|[\overline{S}^{-1}\overline{U}^T U S]_{ii}| \leq \left(1 + \frac{b_0^{(2)}\varrho^\alpha + b_1\varrho}{\overline{\sigma}_i}\right), \quad 1 \leq i \leq N_x.$$

The radius of the Gershgorin circle is:

$$r_i = \sum_{j\neq i}^{N_x} \frac{\sigma_j}{\overline{\sigma}_i}|[\overline{U}^T U]_{ij}|.$$

The upper bound of the radius is obtained by applying the Cauchy-Schwarz inequality and the hypothesis (H1) on the extra-diagonal entries of $[\overline{U}^T U]$:

$$r_i \leq \frac{\left(\sum_{j\neq i}^{N_x}\sigma_j^2\right)^{1/2}}{\overline{\sigma}_i}\left(1 - e^{-\eta\varrho}\right)^{1/2}.$$

Remark that:

$$\sum_{j\neq i}^{N_x}\sigma_j^2 = \|A\|_F^2 - \sigma_i^2.$$

By virtue of Lemma 1, we can write:

$$\|A\|_F^2 \leq \|P\|_F^2 \left[1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right]^2.$$

Moreover, Lemma 2 makes it possible to write:

$$\sigma_i^2 \geq \overline{\sigma}_i^2 \left[1 - \frac{b_0^{(2)} \varrho^\alpha + b_1 \varrho}{\overline{\sigma}_i}\right]^2.$$

By combining these two results we get:

$$\sum_{j \neq i}^{N_x} \sigma_j^2 \leq \|P\|_F^2 \left[1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right]^2 - \overline{\sigma}_i^2 \left[1 - \frac{b_0^{(2)} \varrho^\alpha + b_1 \varrho}{\overline{\sigma}_i}\right]^2.$$

Let us consider that $\|P\|_F^2 = \sum_{i=1}^{N_x} \overline{\sigma}_i^2$:

$$\sum_{j \neq i}^{N_x} \sigma_j^2 \leq \overline{\sigma}_i^2 \left[\left(1 + \sum_{j \neq i}^{N_x} \overline{\sigma}_j^2 / \overline{\sigma}_i^2\right) \left(1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right)^2 - \left(1 - \frac{b_0^{(2)} \varrho^\alpha + b_1 \varrho}{\overline{\sigma}_i}\right)^2\right].$$

The bound on the radius has the following expression:

$$r_i \leq \left[\left(1 + \sum_{j \neq i}^{N_x} \overline{\sigma}_j^2 / \overline{\sigma}_i^2\right) \left(1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right)^2 - \left(1 - \frac{b_0^{(2)} \varrho^\alpha + b_1 \varrho}{\overline{\sigma}_i}\right)^2\right]^{1/2} \left(1 - e^{-\eta \varrho}\right)^{1/2}.$$

For the sake of compactness in the notation we denote: $\xi_i = \frac{b_0^{(2)} \varrho^\alpha + b_1 \varrho}{\overline{\sigma}_i}$, $1 \leq i \leq N_x$. The upper bound on the $i-$th eigenvalue is:

$$|\lambda_i| \leq (1 + \xi_i) + \left(1 - e^{-\eta \varrho}\right)^{1/2} \left[\left(1 + \sum_{j \neq i}^{N_x} \frac{\overline{\sigma}_j^2}{\overline{\sigma}_i^2}\right) \left(1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right)^2 - (1 - \xi_i)^2\right]^{1/2}.$$

Remark that the lowest part of the spectrum of P is the most critical one. Let us particularise the expression of the bound by considering $i = N_x$. We get:

$$\|P^{-1} A(\vartheta)\|_2 \leq (1 + \xi_{N_x}) + \left(1 - e^{-\eta \varrho}\right)^{1/2} \left[\frac{\|P\|_F^2}{\overline{\sigma}_{N_x}^2} \left(1 + \frac{b_0^{(F)} \varrho^\alpha}{\|P\|_F}\right)^2 - (1 - \xi_{N_x})^2\right]^{1/2}.$$

Let us remark that $\lim_{\varrho \to 0} \|P^{-1} A(\vartheta)\|_2 = 1$.

We develop an analogous computation for the matrix $S^{-1} U^T \overline{U}\ \overline{S}$. Concerning its diagonal elements we can write:

$$|[S^{-1} U^T \overline{U}\ \overline{S}]_{ii}| \leq \frac{1}{|1 - \xi_i|},\ 1 \leq i \leq N_x.$$

The Gershgorin radius is bounded by:

$$r_i \leq \frac{1}{|1-\xi_i|}\left(\frac{\|P\|_F^2}{\overline{\sigma}_i}-1\right)^{1/2}\left(1-e^{-\eta\varrho}\right)^{1/2}.$$

This is particularised for $i=i_*$, the index which is maximising the bound expression over all $1\leq i\leq N_x$ and makes it possible to write the following:

$$\|A(\vartheta)^{-1}P\|_2 \leq \frac{1}{|1-\xi_{i_*}|}+\frac{(1-e^{-\eta\varrho})^{1/2}}{|1-\xi_{i_*}|}\left(\frac{\|P\|_F^2}{\overline{\sigma}_{i_*}}-1\right)^{1/2}.$$

Let:

$$c_0(\varrho)=\left[\frac{\|P\|_F^2}{\overline{\sigma}_{N_x}^2}\left(1+\frac{b_0^{(F)}\varrho^\alpha}{\|P\|_F}\right)^2-(1-\xi_{N_x})^2\right]^{1/2},$$

$$c_1(\varrho)=\frac{1}{|1-\xi_{i_*}|}\left(\frac{\|P\|_F^2}{\overline{\sigma}_{i_*}}-1\right)^{1/2},$$

$$c_2(\varrho)=\frac{c_0(\varrho)}{|1-\xi_{i_*}|}+(1+\xi_{N_x})c_1(\varrho).$$

The bound on the condition number can be expressed as:

$$\mathrm{cond}_2\left(P^{-1}A(\vartheta)\right)\leq\frac{1+\xi_{N_x}}{|1-\xi_{i_*}|}+c_2(\varrho)(1-e^{-\eta\varrho})^{1/2}+c_0(\varrho)c_1(\varrho)(1-e^{-\eta\varrho}).$$

And this concludes the proof. □

Let us remark that:

$$\lim_{\varrho\to 0}\mathrm{cond}_2\left(P^{-1}A(\vartheta)\right)=1.$$

The result of the Proposition 1 highlights the intuitive fact that the smaller the set Θ is, the better the performance of the preconditioner is. The different terms of the expression shed some light on the amplification factors acting on the matrix condition number. In particular, if the eigenvectors of the matrix A change significantly, or the matrix have a very small singular value, the amplification of the condition number might be significant even for sets Θ characterised by a relatively small ϱ. This motivates the development of other types of preconditioner. This analysis holds true also for the second type of preconditioner. By construction, the Chebyshev radius of the Voronoi cells of the set Θ, which we denote $\left\{\varrho^{(q)}\right\}_{1\leq q\leq Q}$ satisfy: $\varrho^{(q)}\leq\varrho$. It follows that the condition number cannot be worse than when using the first kind of preconditioner.

3.2 Analysis of the preconditioner (III).

The third type of preconditioner is based on a different principle. Given a residual, it constructs a sequence of approximations of the inverse of $\mathbf{A}$ applied to the residual.

When considering the semi-discretised in x formulation, the preconditioner $\mathbf{P}$ takes the form:

$$P(\vartheta) = \frac{1}{\sum_{m=1}^{L} \gamma_m g_m(\vartheta)} [P^{(x)}]^{-1}.$$

First, let us discuss the action of this rank one operator onto the matrix $A(\vartheta)$.

$$\left[\big(\sum_{m=1}^{L} \gamma_m g_m(\vartheta) \big) P^{(x)} \right]^{-1} \left(\sum_{l=1}^{L} g_l(\vartheta) A^{(l)} \right) = [P^{(x)}]^{-1} \left(\sum_{l=1}^{L} \frac{g_l(\vartheta)}{\sum_{m=1}^{L} \gamma_m g_m(\vartheta)} A^{(l)} \right).$$

Let:

$$y_l(\vartheta) = \frac{g_l(\vartheta)}{\sum_{m=1}^{L} \gamma_m g_m(\vartheta)}, \quad 1 \leq l \leq L.$$

We define:

$$\tilde{A}(\vartheta) = \sum_{l=1}^{L} y_l(\vartheta) A^{(l)}.$$

From this simple computation we see that applying a rank-one operator onto the matrix field is equivalent to construct a preconditioner $P^{(x)}$ acting on a modified matrix $\tilde{A}(\vartheta)$.

3. *Lemma:* *Let $r > p$ and $g_l \in W^{1,r}(\Theta)$, $1 \leq l \leq L$. Let $\left| \sum_{l=1}^{L} \gamma_l g_l(\vartheta) \right| > \alpha_0 > 0$ for all $\vartheta \in \Theta$. For any matrix norm, there exists a positive constant $\kappa > 0$ and $\alpha = 1 - p/r$ such that:*

$$\|\tilde{A}(\vartheta) - P^{(x)}\| \leq \kappa \varrho_*^{\alpha},$$

Proof. Let us start the proof by observing that, for $r > p$ we have $W^{1,r}(\Theta) \hookrightarrow \mathcal{C}^0(\Theta)$. We can verify that the functions:

$$y_l = \frac{g_l}{\sum_{m=1}^{L} \gamma_m g_m}, \quad 1 \leq l \leq L,$$

are such that: $y_l \in W^{1,r} \hookrightarrow \mathcal{C}^{0,\alpha}(\Theta)$, with $\alpha = 1 - \frac{p}{r}$. We shall write:

$$\sup_{\vartheta \in \Theta} |y_l - \beta_l| \leq C_v^{(l)} \left[\frac{\pi^{\frac{p}{2}} \varrho_*^p}{\Gamma(p/2+1)} \right]^{\frac{1}{p} - \frac{1}{r}} \|\nabla g_l\|_{L^p(\Theta)}, \ 1 \leq l \leq L,$$

for positive constants $C_v^{(l)}$, Γ being the Euler gamma function. Let us set $C_v = \max_{1 \leq l \leq L} \left\{ C_v^{(l)} \right\}$.

Any matrix norm of the difference between $\tilde{A}(\vartheta)$ and $P^{(x)}$ satisfies:

$$\|\tilde{A} - P^{(x)}\| = \left\| \sum_{l=1}^{L} (y_l(\vartheta) - \beta_l) A_l \right\| \leq \sum_{l=1}^{L} \sup_{\vartheta} |y_l - \beta_l| \|A_l\|.$$

Let us use the result for the bound of the L^∞ norm of the functions $(y_l - \beta_l)$ and define:

$$\kappa = \frac{\pi^{\frac{1}{2}(1-p/r)} C_v}{[\Gamma(p/2+1)]^{\frac{1}{p}-\frac{1}{r}}} \sum_{l=1}^{L} \|\nabla g_l\|_{L^p(\Theta)} \|A_l\|.$$

It holds:

$$\|\tilde{A}(\vartheta) - P^{(x)}\| \leq \kappa \varrho_*^{1-p/r},$$

which concludes the proof. □

Remark that the use of the embedding for the L^∞ norm made it possible to get the expression of the constants. The conclusion, per se, could be reached by observing that $W^{p,r}(\Theta) \hookrightarrow \mathcal{C}^{0,\alpha}(\Theta)$ with $\alpha = 1 - p/r$.

This Lemma makes it possible to prove a sufficient condition for the preconditioner to get closer and closer to the application of the inverse to the residual. This is the object of the next Proposition. We are going to particularise the result to the case of constant type (III) preconditioner detailed in the last part of the previous section.

Let us prove the following Lemma, based on the results in [25].

4. *Lemma:* *Let a matrix $B \in \mathbb{R}^{N_x \times N_x}$ be a square real valued matrix, and we denote its trace $\mathrm{tr}(B)$, its Frobenius norm $\|B\|_F^2 = \mathrm{tr}(B^T B)$. We consider the square matrix $I - B$, where I is the identity. The eigenvalues of $I - B$ are denoted $\{\lambda_i\}_{1 \leq i \leq N_x} \in \mathbb{C}$. The largest in modulus eigenvalue is denoted λ_1. If:*

$$0 < \frac{\mathrm{tr}(B)}{N_x} < 2,$$

and

$$\|B\|_F < \frac{|\mathrm{tr}(B)|}{(N_x - 1)^{1/2}},$$

then:

$$|\lambda_1| < 1.$$

Proof. We are going to apply the main result of Theorem 3.1 in [25], by taking into account that the matrices are real valued. Let:

$$\mu = \frac{\mathrm{tr}(I - B)}{N_x} = 1 - \frac{\mathrm{tr}(B)}{N_x},$$

$$s_a^2 = \frac{\mathrm{tr}\left((I - (B + B^T) + B^T B)\right)}{N_x} - \mu^2.$$

We can simplify the expression for s_a^2:

$$s_a^2 = \frac{\mathrm{tr}(B^T B)}{N_x} - \frac{\left(\mathrm{tr}(B)\right)^2}{N_x^2}.$$

The upper bound being given by $|\lambda_1| = |\mu| + (N_x - 1)^{1/2} s_a$ we obtain:

$$|\lambda_1| \leq \left|1 - \frac{\mathrm{tr}(B)}{N_x}\right| + (N_x - 1)^{1/2} \left(\frac{\mathrm{tr}(B^T B)}{N_x} - \frac{\big(\mathrm{tr}(B)\big)^2}{N_x^2}\right)^{1/2},$$

We now show that the two hypotheses made on the trace and Frobenius norm are sufficient to have an upper bound which is strictly smaller than 1:

$$\left|1 - \frac{\mathrm{tr}(B)}{N_x}\right| + (N_x - 1)^{1/2} \left(\frac{\mathrm{tr}(B^T B)}{N_x} - \frac{\big(\mathrm{tr}(B)\big)^2}{N_x^2}\right)^{1/2} < 1.$$

We reorder and get:

$$(N_x - 1)^{1/2} \left(\frac{\mathrm{tr}(B^T B)}{N_x} - \frac{\big(\mathrm{tr}(B)\big)^2}{N_x^2}\right)^{1/2} < 1 - \left|1 - \frac{\mathrm{tr}(B)}{N_x}\right|.$$

The first hypothesis

$$0 < \frac{\mathrm{tr}(B)}{N_x} < 2.$$

ensures that the right-hand side is positive. We develop the absolute value and get two cases. The most ristrictive one corresponds to the case $1 - \frac{\mathrm{tr}(B)}{N_x} > 0$. We take the square on both sides when considering this case and get:

$$(N_x - 1) \left[\frac{\|B\|_F^2}{N_x} - \frac{\big(\mathrm{tr}(B)\big)^2}{N_x^2}\right] < \frac{\big(\mathrm{tr}(B)\big)^2}{N_x^2}.$$

The assumption:

$$\|B\|_F^2 < \frac{\mathrm{tr}(B)^2}{(N_x - 1)},$$

makes it possible to verify the inequality and this concludes the proof. □

2. *Proposition:* *Let the hypotheses of Lemma 3 hold. We define:*

$$B(\vartheta) = \sum_{l=1}^{L} y_l(\vartheta) A_l [P^{(x)}]^{-1}.$$

If, for all $\vartheta \in \Theta$*:*

$$0 < \mathrm{tr}\big(B(\vartheta)\big) < 2N_x,$$

and

$$\|B(\vartheta)\|_F < \frac{\mathrm{tr}\big(B(\vartheta)\big)}{(N_x - 1)^{1/2}},$$

then,

$$\lim_{q \to \infty} \|Z^{(q)}\| = 0.$$

Proof. Let us recall the iterations defining the third type of preconditioning strategy with constant rank one operator $\mathbf{P}$:

$$Z^{(q)} = R - \mathbf{A}R^{(q)},$$

$$R^{(q+1)} = R^{(q)} + \mathbf{P}Z^{(q)}.$$

First, let us remark that $Z^{(q)} \to 0$ implies $R^{(q)} \to \mathbf{A}^{-1}R$, which is the sought result. Let us show that the sequence of $Z^{(q)}$ is determined by the matrix $\mathbf{I} - \mathbf{AP}$.

$$Z^{(q+1)} = R - \mathbf{A}R^{(q+1)} = R - \mathbf{A}\left[R^{(q)} + \mathbf{P}Z^{(q)}\right].$$

This gives us:

$$Z^{(q+1)} = R - \mathbf{A}R^{(q)} - \mathbf{AP}Z^{(q)} = Z^{(q)} - \mathbf{AP}Z^{(q)} = [\mathbf{I} - \mathbf{AP}]Z^{(q)}.$$

Let us remark that this implies:

$$Z^{(q)} = [\mathbf{I} - \mathbf{AP}]^q Z^{(0)} = [\mathbf{I} - \mathbf{AP}]^q R.$$

A sufficient condition for the sequence $Z^{(q)}$ to converge to 0 is given by:

$$\|I - \mathbf{AP}\| < 1.$$

We introduce the semi-discretised in x formulation and obtain:

$$A(\vartheta)P(\vartheta) = \sum_{l=1}^{L} \frac{g_l(\vartheta)}{\sum_{m=1}^{L} \gamma_m g_m(\vartheta)} A_l [P^{(x)}]^{-1},$$

which gives us:

$$A(\vartheta)P(\vartheta) = \left[\sum_{l=1}^{L} y_l(\vartheta) A_l\right] [P^{(x)}]^{-1} = \tilde{A}(\vartheta)[P^{(x)}]^{-1} = B(\vartheta).$$

We make use of the result of Lemma 4, which guarantees that the spectral radius of $I - B(\vartheta)$ is smaller than 1, and this concludes the proof. □

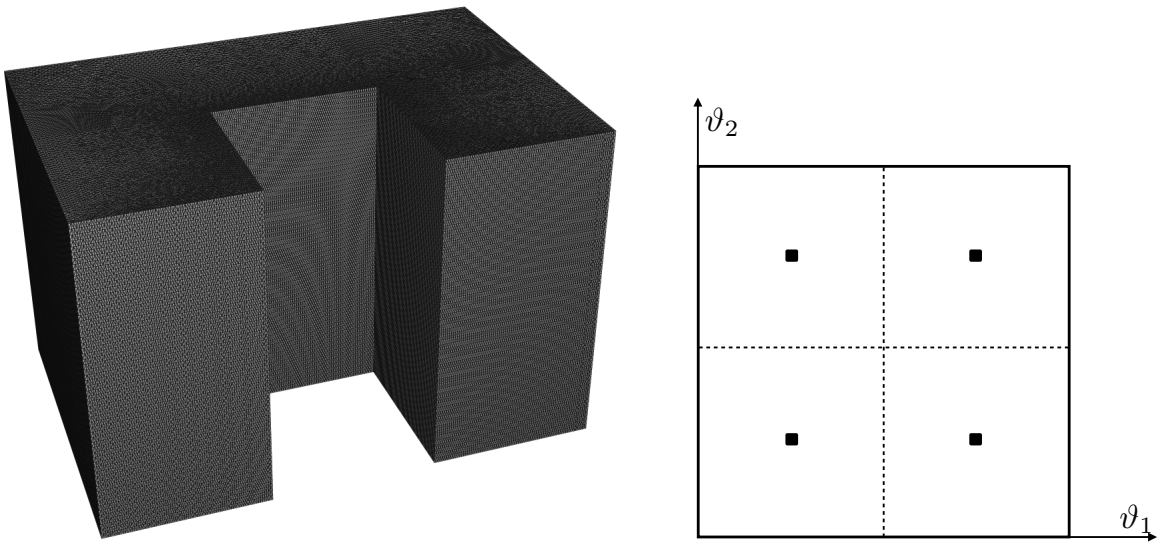
The result of the Proposition highlights that there are sufficient conditions for approximating the action of $[A(\vartheta)]^{-1}$ by the repeated application (eventually an infinite number of times) of a matrix field which is a product of a function depending upon the parameters and a constant matrix.

Remark: An alternative way to get a sufficient condition is given by the following consideration:

$$\|I - \tilde{A}(\vartheta)[P^{(x)}]^{-1}\| = \|(P^{(x)} - A(\vartheta))[P^{(x)}]^{-1}\| \leq \|P^{(x)} - A(\vartheta)\| \|[P^{(x)}]^{-1}\| < 1.$$

We use the result of Lemma 3 and write:

$$\kappa \varrho_*^{\alpha} \|[P^{(x)}]^{-1}\| < 1.$$



(a) (b)

Figure 1: Mesh of the space domain Ω (a) and parameter box with $Q = 4$ centres for Preconditioner (II) (b).

4 Numerical experiments.

We are going to present, in this section, some numerical experiments to assess the properties of the proposed preconditioners. We consider $\Omega \subset \mathbb{R}^3$ the spacial $3d$ domain represented in Figure 4.

The parameter domain is $\Theta = [0.1, 20]^2$. The problem we are going to solve is a linear elliptic PDE of the form:

$$-\vartheta_1 \Delta u + \vartheta_2 u = \frac{\vartheta_1 + \vartheta_2}{2}, \quad \text{on } \Omega,$$

$$u = 0, \quad \text{on } \partial\Omega.$$

In this case the solution $u(\cdot, \vartheta) \in H_0^1(\Omega)$ exists and it is unique for all $\vartheta \in \Theta$. Concerning the space discretisation, we are going to consider P1 finite elements in space and a collocation method for the parameters. The computational mesh is shown in Figure 4.(a), the number of degrees of freedom is roughly 10^6. Concerning the parameters we consider a uniform grid with $N_{\vartheta_1} = N_{\vartheta_2} = 10^2$, hence corresponding to 10^4 collocation points. The total number of degrees of freedom, if we would write the global system would amount to roughly 10^{10} degrees of freedom.

For all the tests we are going to show we are going to consider a TT-GMRES as it described in [18]. Concerning the rounding procedure for the tensors, we use a standard POD with truncation tolerance $\varepsilon_{POD} = 10^{-8}$. For all the tests, we are going

to perform $N_{GMRES} = 40$ iterations of TT-GMRES with truncation. For all the tests we performed, we recorded the Frobenius norm of the non-preconditioned residual at each GMRES iteration, the solution rank and the computational time per iteration. Concerning the Frobenius residual norm, it is defined classically, by considering the matrix whose columns are the residual of the 10^4 parametric problems we solve. To estimate the average residual per problem, the number reported in the next sections can be roughly divided by $\sqrt{N_\vartheta} = 10^2$.

Without preconditioner, after 40 GMRES iteration, the Frobenius norm of the residual is around $2 \cdot 10^{-1}$.

4.1 Preconditioners type (I) and (II).

We start by looking at the results for the preconditioners type (I) and (II). As already stated, the preconditioner type (I) can be seen as a particular case of the second type of preconditioner, when taking just one Voronoi cell. In this case we take $\overline{\vartheta}$ as the barycenter of Θ. In this case, when considering more cells, we decided to uniformly subdivide the square, and place the centres as follows. We recall that $Q \in \mathbb{N}^*$ is the total number of cells, and we tried $Q = \{1, 4, 9, 16\}$.

$$\vartheta_{1,2}^{(j)} = \frac{L(2j-1)}{2\sqrt{Q}}, \quad 1 \leq j \leq \sqrt{Q},$$

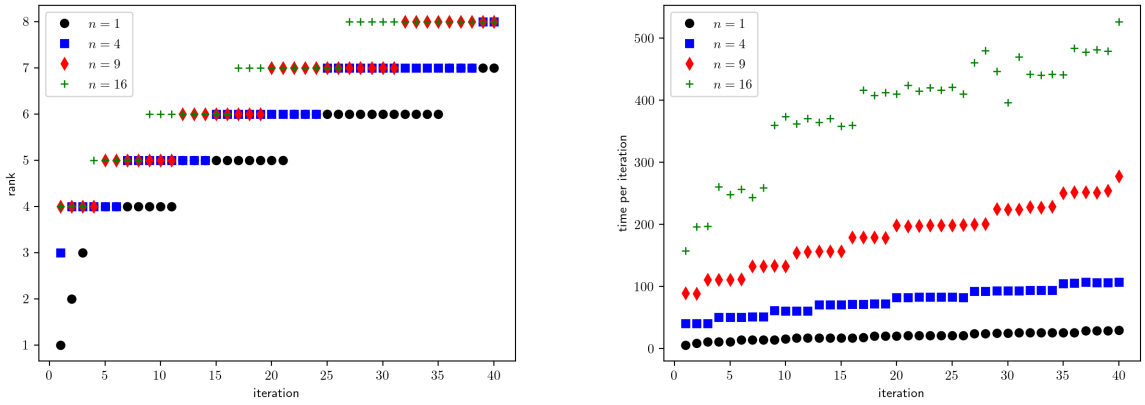
and L for the present problem, represents the lenght of the parameters box sides. An exemple of center placement is represented in Figure 4.(b).

In Figure 4.1.(a) we show the solution rank as function of the iteration, for the 4 different values of Q we tested. We can see that, for the present problem, the solution is relatively low-rank, leading to a significant compression ratio. We see that using a more refined in ϑ subdivision leads sooner to higher ranks, but this will have to be interpreted by looking at the residual norm as function of the iteration. In Figure 4.1.(b) we show the computational time per iteration as function of the iteration. Two factors have an impact of the computational cost: the solution rank, and the number Q which determines the number of linear problem resolution we have to perform. In the plot we clearly see that these effects.

In Figure 4.1.(a) we show the residual Frobenius norm as function of the GMRES iteration. We can clearly see the benefit of introducing a finer subdivision. In particular, for $n = 16$, for the same number of iterations, we can gain almost two orders of magnitude on the residual. Lastly, we plot in Figure 4.1.(b) the residual as function of the computational time (cumulated), in logarithmic scale. We can see that, for the present problem, the finer subdivision is able to reach lower values of the residual, but at a larger cost.

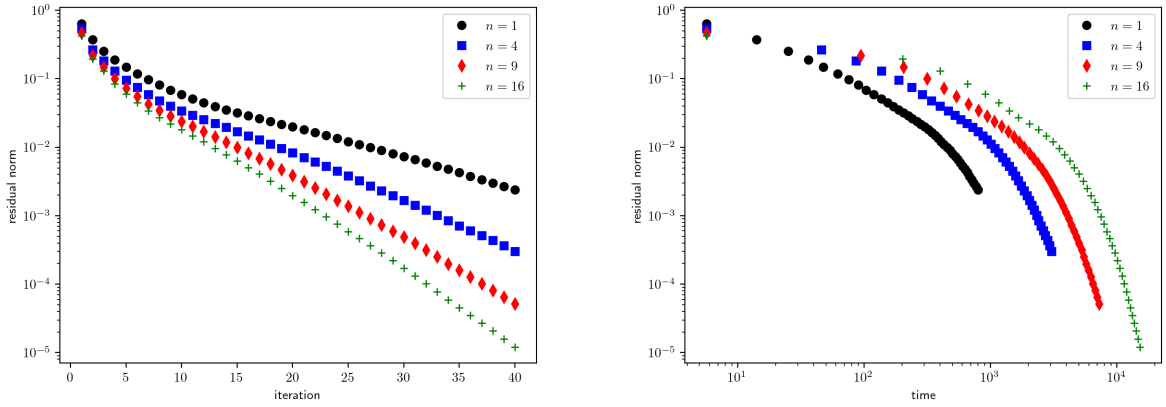
4.2 Preconditioners type (III).

We show in this section the results for the third type of Preconditioner. In particular, we tested extensively the preconditioner we obtain by considering, as constant vectors



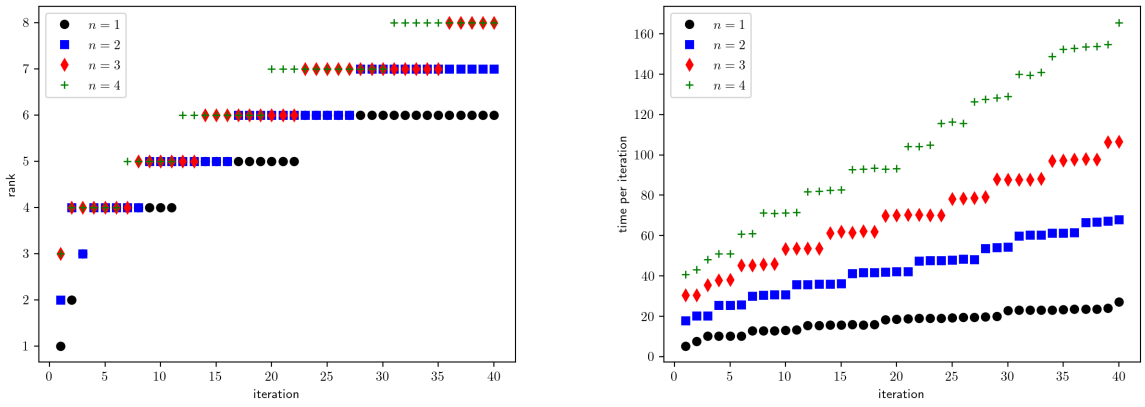
(a) (b)

Figure 2: Solution ranks (a) and computational time per iteration (b) as function of the GMRES iteration for Preconditioner (II).



(a) (b)

Figure 3: Solution residual norm as function of the GMRES iteration (a) and residual norm as function of computational time, in logarithmic scale (b) for Preconditioner (II)



(a) (b)

Figure 4: Solution ranks (a) and computational time per iteration (b) as function of the GMRES iteration for preconditioner (III)

$\mathbf{v}_x = \mathbf{1}_{N_x}$, and $\mathbf{v}_\vartheta = \mathbf{1}_{\mathbf{N}_\vartheta}$, the vectors whose components (N_x and N_ϑ respectively) are all equal to 1. In this case we have:

$$\beta_l = \frac{{\mathbf{1}_{\mathbf{N}_\vartheta}}^T D_l \mathbf{1}_{\mathbf{N}_\vartheta}}{N_\vartheta}, \quad 1 \leq l \leq 2,$$

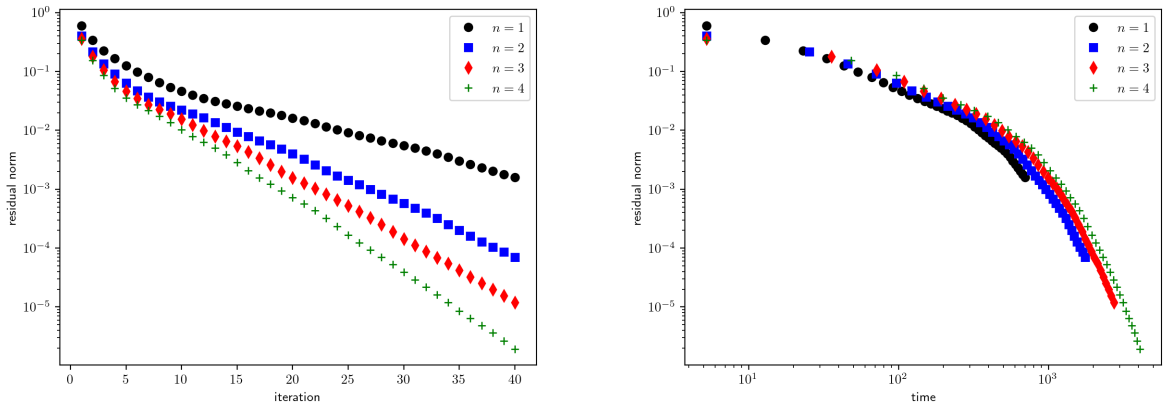
$$\gamma_l = \frac{{\mathbf{1}_{\mathbf{N_x}}}^T A_l \mathbf{1}_{\mathbf{N_x}}}{N_x}.$$

With this choice, given the fact that $g_1(\vartheta) = \vartheta_1 > 0$ and $g_2(\vartheta) = \vartheta_2 > 0$, since the mass and the stiffness matrices are positive definite, it follows that $P^{(x)}$ is positive definite, hence invertible. Similarly, we can prove that the denominator of the functions y_1, y_2 does not vanish.

We consider $Q = 1, 2, 3, 4$. As done for the other type of preconditioner, we report hereafter the observations on the rank, residual norm and computational time.

In Figure 4.2.(a) we show the solution rank as function of the GMRES iteration. The behaviour is similar to the one observed for the second type of preconditioner. A preconditioner with larger Q would tend to decrease the residual faster with respect to the iteration, inducing a higher rank. In Figure 4.2.(b) we show the computational time per iteration as function of the iteration. In this type of preconditioner as well, the cost is determined by Q and by the rank.

In Figure 4.2.(a) we show the residual Frobenius norm as function of the GMRES iteration. As we can see, also in this case, there are roughly two orders of magnitude after 40 GMRES iteration between the resolution with $Q = 1$ and the one with $Q = 4$. In Figure 4.2.(b) we show the residual Frobenius norm as function of the computational



(a) (b)

Figure 5: Solution residual norm as function of the GMRES iteration (a) and residual norm as function of computational time, in logarithmic scale (b) for preconditioner (III).

time in logarithm scale. We see here a pattern which is similar to the one observed for the second type of preconditioner.

Overall, for this numerical test, we can see that the third type of preconditioner has a better trade off between accuracy and computational cost.

5 Conclusion.

In this work we addressed the problem of setting up preconditioners for parametric PDEs, and, in particular, when the solution is approximated in a low-rank format in which the space-time variables are separated from the parameters. Three different preconditioning strategies have been investigated and compared. In the three of them, the goal is to approximate the action of the inverse of the (parametric) matrix obtained when semi-discretising in space(-time) the PDEs. The two first type of preconditioners are based on the idea that the matrix $A(\overline{\vartheta})$ evaluated for one instance of the parameter is a good approximation of $A(\vartheta)$ when ϑ belongs to a neighbourhood of $\overline{\vartheta}$. The third type of preconditioner is based on the approximation of the action of the inverse by repeated application of a rank-one tensor operator, which is the product of a function depending upon the parameters and a constant matrix. Some theoretical results have been presented concerning the three preconditioning strategies. The numerical tests have been performed by considering the accuracy and the computational burden of the preconditioners used in a TT-GMRES method. For the three strategies, the gain with respect to a non-preconditioned strategy were significant. On the numerical tests

performed, the third preconditioning strategy proved to be the most effective one in terms of trade-off between accuracy and computational cost after a fix number of TT-GMRES iteration.

Numerous perspectives arise concerning all these strategies. For the preconditioner type (II) further investigations will try to understand whether we can optimise in a relatively inexpensive way, the distribution of the centres of the Voronoi cells. For the preconditioner type (III) we will investigate how an iteration-dependent preconditioner could be set up by limiting the computational overheads.

References

[1] Jan Ackmann, Peter D Düben, Tim N Palmer, and Piotr K Smolarkiewicz. Machine-learned preconditioners for linear solvers in geophysical fluid flows. *arXiv preprint arXiv:2010.02866*, 2020.

[2] Yael Azulay and Eran Treister. Multigrid-augmented deep learning preconditioners for the helmholtz equation. *SIAM Journal on Scientific Computing*, (0):S127–S151, 2022.

[3] Markus Bachmayr. Low-rank tensor methods for partial differential equations. *Acta Numerica*, 32:1–121, 2023.

[4] Markus Bachmayr and Reinhold Schneider. Iterative methods based on soft thresholding of hierarchical tensors. *Foundations of Computational Mathematics*, 17:1037–1083, 2017.

[5] Markus Bachmayr, Reinhold Schneider, and André Uschmajew. Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. *Foundations of Computational Mathematics*, 16:1423–1472, 2016.

[6] Jonas Ballani and Lars Grasedyck. A projection method to solve linear systems in tensor format. *Numerical linear algebra with applications*, 20(1):27–43, 2013.

[7] Peter Benner, Thomas Richter, and Roman Weinhandl. A low-rank approach for nonlinear parameter-dependent fluid-structure interaction problems. In *Numerical Mathematics and Advanced Applications ENUMATH 2019: European Conference, Egmond aan Zee, The Netherlands, September 30-October 4*, pages 1157–1165. Springer, 2020.

[8] Martijn Boussé, Nico Vervliet, Ignat Domanov, Otto Debals, and Lieven De Lathauwer. Linear systems with a canonical polyadic decomposition constrained solution: Algorithms and applications. *Numerical Linear Algebra with Applications*, 25(6):e2190, 2018.

[9] Michael Brazell, Na Li, Carmeliza Navasca, and Christino Tamon. Solving multilinear systems via tensor inversion. *SIAM Journal on Matrix Analysis and Applications*, 34(2):542–570, 2013.

[10] Alexandra Bünger, Valeria Simoncini, and Martin Stoll. A low-rank matrix equation method for solving pde-constrained optimization problems. *SIAM Journal on Scientific Computing*, 43(5):S637–S654, 2021.

[11] Sergey V Dolgov. Tt-gmres: solution to a linear system in the structured tensor format. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 28(2):149–172, 2013.

[12] Sergey V Dolgov and Dmitry V Savostyanov. Alternating minimal energy methods for linear systems in higher dimensions. *SIAM Journal on Scientific Computing*, 36(5):A2248–A2271, 2014.

[13] Mareike Dressler, André Uschmajew, and Venkat Chandrasekaran. Kronecker product approximation of operators in spectral norm via alternating sdp. *SIAM Journal on Matrix Analysis and Applications*, 44(4):1693–1708, 2023.

[14] Markus Götz and Hartwig Anzt. Machine learning-aided numerical linear algebra: Convolutional neural networks for the efficient preconditioner generation. In *2018 IEEE/ACM 9th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (scalA)*, pages 49–56. IEEE, 2018.

[15] Lars Grasedyck, Daniel Kressner, and Christine Tobler. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

[16] Alena Kopaničáková and George Em Karniadakis. Deeponet based preconditioning strategies for solving parametric linear systems of equations. *arXiv preprint arXiv:2401.02016*, 2024.

[17] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Preconditioned low-rank riemannian optimization for linear systems with tensor product structure. *SIAM Journal on Scientific Computing*, 38(4):A2018–A2044, 2016.

[18] Daniel Kressner and Christine Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM journal on matrix analysis and applications*, 31(4):1688–1714, 2010.

[19] Dongdong Liu, Wen Li, and Seak-Weng Vong. A new preconditioned sor method for solving multi-linear systems with an m-tensor. *Calcolo*, 57(2):15, 2020.

[20] Anthony Nouy. A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 199(23-24):1603–1626, 2010.

[21] Davide Palitta, Marcel Schweitzer, and Valeria Simoncini. Sketched and truncated polynomial krylov subspace methods: Matrix equations. *arXiv preprint arXiv:2311.16019*, 2023.

[22] V Simoncini. Numerical solution of a class of third order tensor linear equations. *Bollettino dell'Unione Matematica Italiana*, 13(3):429–439, 2020.

[23] André Uschmajew and Bart Vandereycken. *Geometric methods on low-rank matrix and tensor manifolds*. Springer, 2020.

[24] Nicolas Venkovic. *Stratégies de préconditionnement pour équations stochastiques elliptiques aux dérivées partielles*. PhD thesis, Bordeaux, 2023.

[25] Henry Wolkowicz and George PH Styan. Bounds for eigenvalues using traces. *Linear algebra and its applications*, 29:471–506, 1980.