Adv Comput Math (2020) 46:46
https://doi.org/10.1007/s10444-020-09749-3

Least-Squares Padé approximation of parametric and stochastic Helmholtz maps

Francesca Bonizzoni[1] · Fabio Nobile[2] · Ilaria Perugia[1] · Davide Pradovera[2]

Received: 20 July 2018 / Accepted: 29 November 2019 /
Published online: 8 May 2020

Abstract

The present work deals with rational model order reduction methods based on the single-point Least-Square (LS) Padé approximation techniques introduced in Bonizzoni et al. (ESAIM Math. Model. Numer. Anal., **52**(4), 1261–1284 2018, Math. Comput. **89**, 1229–1257 2020). Algorithmical aspects concerning the construction of rational LS-Padé approximants are described. In particular, we show that the computation of the Padé denominator can be carried out efficiently by solving an eigenvalue-eigenvector problem involving a Gramian matrix. The LS-Padé techniques are employed to approximate the frequency response map associated with two parametric time-harmonic acoustic wave problems, namely a transmission-reflection problem and a scattering problem. In both cases, we establish the meromorphy of the frequency response map. The Helmholtz equation with stochastic wavenumber is also considered. In particular, for Lipschitz functionals of the solution and their corresponding probability measures, we establish weak convergence of the measure derived from the LS-Padé approximant to the true one. 2D numerical tests are performed, which confirm the effectiveness of the approximation methods.

Keywords Hilbert space-valued meromorphic maps · Padé approximants · Convergence of Padé approximants · Parametric Helmholtz equation · PDE with random coefficients

Mathematics Subject Classifications (2010) 30D30 · 41A21 · 41A25 · 35A17 · 35J05 · 35R60 · 65D15

Communicated by: Anthony Nouy

This article belongs to the Topical Collection: *Model Reduction of Parametrized Systems*
Guest Editors: Anthony Nouy, Peter Benner, Mario Ohlberger, Gianluigi Rozza, Karsten Urban and Karen Willcox

✉ Francesca Bonizzoni
francesca.bonizzoni@univie.ac.at

Extended author information available on the last page of the article.

1 Introduction

Many applications require the fast and accurate numerical evaluation of Helmholtz frequency response functions, i.e., functions that map the wavenumber to the solution (or some quantity of interest related to the solution) of the corresponding time-harmonic wave-problem, for a large number of frequencies. In mid- and high-frequency regimes, very fine meshes or high polynomial degrees should be considered, in order to obtain accurate Finite Element (FE) solutions of the time-harmonic wave-problem. Moreover, low order FE schemes are affected by the pollution effect [2], namely, an increasing discrepancy between the best approximation error and the FE error, as the wavenumber increases. In the "many-queries" context, i.e., when many solutions of the underlying Partial Differential Equation (PDE) are needed, the "brute force" approach entails the solution of a large number of high-dimensional linear systems, and it is then out of reach.

Model order reduction methods aim at significantly reducing the computational cost by approximating the quantity of interest starting from evaluations at only few wavenumbers. They rely on a two-step strategy: the *offline* stage consists in the computation of a finite dimensional basis—e.g., the basis of snapshots (see, e.g., [6, 11, 17, 18, 23, 27, 31, 32, 34, 35]), or evaluations of the frequency response map and its derivatives at fixed centers (Padé method, see, e.g., [3, 10, 12, 13, 16]); the output of this phase, whose computational cost may be very high, is stored, to be used during the *online* phase, in which the approximation of the frequency response map corresponding to a given new value of the parameter is constructed. This stage does not involve the numerical solution of any large-scale PDE, and is expected to provide the output in real time.

In this work, we focus on the Padé-based model order reduction technique introduced in [3], defined for any given univariate meromorphic map $\mathcal{S} : \mathbb{C} \to V$, V being a Hilbert space (viz. $V = H^1(D)$, $D \subset \mathbb{R}^d$), and relying on a *single-point Least-Square (LS) Padé approximant*. In particular, the single-point LS-Padé approximant of $\mathcal{S}$ centered at $z_0 \in \mathbb{C}$, denoted by $\mathcal{S}_{[M/N]}$, is given by the rational V-valued map $\mathcal{S}_{[M/N]}(z) = \frac{\mathcal{P}_{[M/N]}(z)}{\mathcal{Q}_{[M/N]}(z)}$, where $\mathcal{P}_{[M/N]}(z) = \sum_{\alpha=0}^{M} p_\alpha (z - z_0)^\alpha$, with coefficients $p_\alpha \in V$ (we write $\mathcal{P}_{[M/N]} \in \mathbb{P}_M(\mathbb{C}; V)$), and $\mathcal{Q}_{[M/N]} \in \mathbb{P}_N^\star(\mathbb{C})$, where $\mathbb{P}_N^\star(\mathbb{C})$ is the set of all polynomials with complex coefficients $\{q_\alpha\}_{\alpha=0}^N$ such that $\sum_{\alpha=0}^N |q_\alpha|^2 = 1$.

In [3], we have analyzed the convergence of $\mathcal{S}_{[M/N]}$ to $\mathcal{S}$ as $M \to \infty$ for a fixed denominator degree N. In particular, the LS-Padé approximant $\mathcal{S}_{[M/N]}$ identifies the N poles of $\mathcal{S}$ closest to the center z_0, as limit of the roots of the denominator $\mathcal{Q}_{[M/N]}(z)$ for M going to ∞.

We also consider a modified version of the single-point LS Padé approximant, named fast LS-Padé that has been introduced and analyzed in [4].

In this paper, we describe in detail the *algorithmical aspects* of the construction of both single-point LS-Padé approximants. In particular, the identification of the LS-Padé denominator is proved to be equivalent to the identification of the normalized eigenvector corresponding to the smallest non-negative eigenvalue of the Gramian matrix of the set $\left\{\mathcal{S}(z_0), \left(\mathcal{S}\right)_{1,z_0}, \ldots, \left(\mathcal{S}\right)_{E,z_0}\right\}$, where $\left(\mathcal{S}\right)_{\alpha,z_0}$ denotes the Taylor coefficient of $\mathcal{S}$ of order α at z_0.

Moreover, we explore the effectiveness of the single-point LS-Padé techniques when applied to *parametric* frequency response problems which go beyond the setting considered in [3], namely, a *transmission-reflection* problem, and a *scattering* problem. In both cases, we first prove that the frequency response map associated with the considered problem is meromorphic. 2D numerical results are provided, which demonstrate the convergence of the LS-Padé approximation.

In the simulations we carry out, we compare both methods with a “natural” state-of-the-art competitor, the Krylov subspace method [7, 9]. Our numerical experiments show that the Krylov subspace and the fast LS-Padé methods yield comparable results, sharper than the standard LS-Padé approximation.

The *stochastic* Helmholtz boundary value problem is also considered. We refer to [28] for Uncertainty Quantification for frequency responses in vibroacoustics, to [8, 20, 21] for model order reduction for random frequency responses in structural dynamics, and to [15, 30] for the stochastic Helmholtz equation with uncertainty arising either in the forcing term or in the boundary data or in the shape of the scatterer.

Within the present framework, we propose a novel approach to the *stochastic* Helmholtz boundary value problem based on the LS-Padé technique, where the squared wavenumber k^2 is modeled as a K-valued random variable, with $K = [k^2_{min}, k^2_{max}]$. We approximate the random variable $X := \mathcal{L}(\mathcal{S}(k^2))$ with $X_P := \mathcal{L}(\mathcal{S}_{[M/N]}(k^2))$. Here, $\mathcal{L} : V \to \mathbb{R}$ is a Lipschitz functional representing a quantity of interest, $\mathcal{S}$ is the meromorphic frequency response map associated with the (stochastic) Helmholtz equation endowed with either homogeneous Dirichlet or homogeneous Neumann boundary conditions, and $\mathcal{S}_{[M/N]}$ is the LS-Padé approximation of $\mathcal{S}$. An upper bound on the approximation error for the characteristic function is derived.

All the considered boundary value problems fall into the following general setting. Let D be an open connected bounded Lipschitz domain in $\mathbb{R}^d$ ($d = 1, 2, 3$), and consider the following Helmholtz boundary value problem

$$\begin{cases} -\Delta u - k^2 \varepsilon_r \, u = f & \text{in } D, \\ u = g_D & \text{on } \Gamma_D, \\ \nabla u \cdot \mathbf{n} = g_N & \text{on } \Gamma_N, \\ \nabla u \cdot \mathbf{n} - iku = g_R & \text{on } \Gamma_R, \end{cases} \tag{1}$$

where the squared wavenumber k^2 is either a parameter or a random variable, which takes values into an interval of interest $K := [k^2_{min}, k^2_{max}] \subset \mathbb{R}^+$, $\varepsilon_r = \varepsilon_r(\mathbf{x}) \in L^\infty(D)$, $f \in L^2(D)$, $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in H^{-1/2}(\Gamma_N)$, $g_R \in H^{-1/2}(\Gamma_R)$, and $\{\Gamma_D, \Gamma_N, \Gamma_R\}$ is a partition of ∂D, i.e., $\overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R = \partial D$ and $\Gamma_D \cap \Gamma_N = \Gamma_D \cap \Gamma_R = \Gamma_N \cap \Gamma_R = \emptyset$. Throughout the paper, we denote by V the Hilbert space $H^1_{\Gamma_D}(D)$. Moreover, we assume the functions in V to be complex-valued.

The outline of the paper is the following. In Section 2, we recall the definition of the single-point LS-Padé approximant and the main convergence result of [3]. In Section 3, we describe the algorithm to compute LS-Padé approximant, and we describe briefly its modified version, leading to the fast LS-Padé approximants. Section 4 deals with a parametric transmission-reflection problem, whereas Section 5 deals with a parametric scattering problem. In Section 6, the LS-Padé methodology is applied to the stochastic setting. Finally, conclusions are drawn in Section 7.

2 Least-Squares Padé approximation of the parametric model problem

This section deals with the Least-Squares (LS) Padé approximation of the following parametric Helmholtz problem:

Problem 1 (**Parametric Model Problem**) *The Helmholtz equation* (1) *has parametric squared wavenumber* $k^2 \in K := [k^2_{min}, k^2_{max}] \subset \mathbb{R}^+$, $\varepsilon_r = 1$, *and is endowed with either Dirichlet or Neumann homogeneous boundary conditions on* ∂D, *i.e.,* $\Gamma_R = \emptyset$ *and either* $\Gamma_D = \partial D$ *and* $g_D = 0$, *or* $\Gamma_N = \partial D$ *and* $g_N = 0$.

Let $\mathcal{S}$ be the frequency response map which associates to each $z \in \mathbb{C}$, the solution $u_z \in V$ of the weak formulation of Problem 1:

$$\int_D \nabla u_z(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x})\, d\mathbf{x} - z \int_D u_z(\mathbf{x})\overline{v}(\mathbf{x})\, d\mathbf{x} = \int_D f(\mathbf{x})\overline{v}(\mathbf{x})\, d\mathbf{x} \quad \forall v \in V. \tag{2}$$

The following results were proved in [3].

Theorem 2.1 *Let* $z \in \mathbb{C} \setminus \Lambda$, Λ *being the set of (real, non negative) eigenvalues of the Laplace operator with the considered boundary conditions. Then, Problem* (2) *admits a unique solution.*

Theorem 2.2 *The frequency response map* $\mathcal{S} : \mathbb{C} \to V$ *defined above is meromorphic in* $\mathbb{C}$, *with a pole of order one at each* $\lambda \in \Lambda$, *i.e.* $z \mapsto (z-\lambda)\mathcal{S}(z)$ *is holomorphic over a neighborhood of* λ.

Remark 2.3 For the sake of simplicity, in Problem 1 we endow the Helmholtz equation with either homogeneous Dirichlet or homogeneous Neumann boundary conditions. Small modifications to the proofs of Theorem 3.1, Proposition 4.1, and Proposition 4.2 in [3] allow to handle both homogeneous mixed Dirichlet-Neumann and non-homogeneous Neumann boundary conditions, and to derive an analogous result as Theorems 2.1 and 2.2. In Section 4, we will show how to handle non-homogeneous Dirichlet boundary conditions.

We recall now the definition and the convergence theorem of the LS-Padé approximant of the frequency response map $\mathcal{S}$.

Let $K = [k^2_{min}, k^2_{max}] \subset \mathbb{R}^+$ be the interval of interest, and $z_0 \in \mathbb{C} \setminus \Lambda$ with $\mathrm{Re}\,(z_0) > 0$. To fix ideas, we take $z_0 = \frac{k^2_{min}+k^2_{max}}{2} + \delta i$, with $\delta \in \mathbb{R} \setminus \{0\}$ arbitrary. The (standard) LS-Padé approximant of $\mathcal{S}$, centered at z_0, is given by the ratio of two polynomials of degree M and N respectively:

$$\mathcal{S}_{[M/N]}(z) := \frac{\mathcal{P}_{[M/N]}(z)}{\mathcal{Q}_{[M/N]}(z)}. \tag{3}$$

The denominator $\mathcal{Q}_{[M/N]}(z)$ is a function of z only, and belongs to the space $\mathbb{P}^{\star}_{N}(\mathbb{C})$ of all polynomials of degree at most N, $q = \sum_{i=0}^{N} q_i (z - z_0)^i \in \mathbb{P}_N(\mathbb{C})$, such that $\sum_{i=0}^{N} |q_i|^2 = 1$. The numerator $\mathcal{P}_{[M/N]} : \mathbb{C} \to V$ is a function of both the complex variable z and the space variable $\mathbf{x} \in D$. More precisely, $\mathcal{P}_{[M/N]}(z) = \sum_{i=0}^{M} p_i (z - z_0)^i$, with coefficients $p_i \in V$. In the following, we denote by $\mathbb{P}_M(\mathbb{C}; V)$ the space of polynomials of degree at most M in $z \in \mathbb{C}$ with coefficients in V.

The construction of the LS-Padé approximant proposed in [3] relies on the minimization of the functional $j_{E,\rho} : \mathbb{P}_M(\mathbb{C}; V) \times \mathbb{P}^{\star}_{N}(\mathbb{C}) \to \mathbb{R}$, parametric in $E \in \mathbb{N}$ and $\rho \in \mathbb{R}^+$, defined as

$$j_{E,\rho}(P, Q) = \left(\sum_{\alpha=0}^{E} \left(\rho^{2\alpha} \left\| \big(Q(z)\mathcal{S}(z) - P(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \right) \right)^{1/2}, \tag{4}$$

where the brackets $\big(\cdot\big)_{\alpha,z_0}$ denote the α-th Taylor coefficient of the Taylor series centered at z_0 (i.e., for a map $\mathcal{T} : \mathbb{C}\backslash\Lambda \to V$, $\big(\mathcal{T}(z)\big)_{\alpha,z_0} = \frac{1}{\alpha!}\frac{d^\alpha \mathcal{T}}{dz^\alpha}(z_0)$), and $\|\cdot\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ denotes the weighted $H^1(D)$-norm (equivalent to the standard one) induced by the inner product

$$\langle u, v\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} := \int_D \big(\nabla u(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) + \mathrm{Re}\,(z_0)\, u(\mathbf{x})\overline{v}(\mathbf{x})\big)\, d\mathbf{x}. \tag{5}$$

We recall the formal definition of the LS-Padé approximant of the solution map $\mathcal{S}$, and we refer to Section 3 for the proof of the existence of a (not in general unique) LS-Padé approximant.

Definition 2.4 Let $M, N \in \mathbb{N}$, $E \geq M + N$, and $\rho \in \mathbb{R}^+$. A LS-Padé approximant $\mathcal{S}_{[M/N]}$, centered at z_0, of the solution map $\mathcal{S}$ is a quotient $\frac{P}{Q}$ with $P \in \mathbb{P}_M(\mathbb{C}; V)$, $Q \in \mathbb{P}^{\star}_{N}(\mathbb{C})$, such that

$$j_{E,\rho}(P, Q) \leq j_{E,\rho}(R, S) \quad \forall R \in \mathbb{P}_M(\mathbb{C}; V)\,,\ \forall S \in \mathbb{P}^{\star}_{N}(\mathbb{C})\,. \tag{6}$$

Note that the LS-Padé approximant depends on E, $E \geq M + N$, as it is the minimizer of the functional $j_{E,\rho}(\cdot, \cdot)$. For simplicity, we have omitted this dependence in the notation $\mathcal{S}_{[M/N]}$.

In [3], the following convergence result has been proved (in a weaker form, see Remark 2.6).

Theorem 2.5 *Let $N \in \mathbb{N}$ be fixed, and let $R \in \mathbb{R}^+$ be such that the disk $\overline{\mathcal{B}(z_0, R)}$ contains exactly N poles of $\mathcal{S}$. Then, for any $z \in \mathcal{B}(z_0, R) \setminus \Lambda$ and for any $|z - z_0| < \rho < R$, it holds*

$$\lim_{M\to\infty} \left\| \mathcal{S}(z) - \mathcal{S}_{[M/N]}(z) \right\|_{V,\sqrt{\mathrm{Re}(z_0)}} = 0.$$

Moreover, given $\alpha > 0$ small enough, introduce the open subset

$$K_\alpha := \bigcup_{\lambda\in\Lambda\cap K} (\lambda - \alpha, \lambda + \alpha) \subset K.$$

Then for any $0 < \rho < R$ *such that* $\mathcal{B}(z_0, \rho) \supset K$*, there exists* $M^\star \in \mathbb{N}$ *such that, for any* $M \geq M^\star$ *and for any* $z \in K \setminus K_\alpha$*, it holds*

$$\left\| \mathcal{S}(z) - \mathcal{S}(z)_{[M/N]} \right\|_{V,\sqrt{\mathrm{Re}(z_0)}} \leq C \frac{1}{\alpha^3} \left(\frac{\rho}{R} \right)^{M+1}, \tag{7}$$

where the constant $C > 0$ *depends on* ρ*,* R*,* N*,* z_0*,* $\lambda_{min} = \min\{\lambda \in \Lambda\}$*,* $\|f\|_{L^2(D)}$*, and* $g(z) = \prod_{\lambda \in \Lambda \cap \mathcal{B}(z_0,R)} (z - \lambda)$*.*

Remark 2.6 The actual bound proved in [3] is

$$\left\| \mathcal{S}(z) - \mathcal{S}(z)_{[M/N]} \right\|_{V,\sqrt{\mathrm{Re}(z_0)}} \leq c(\alpha) \frac{1}{\alpha} \left(\frac{\rho}{R} \right)^{M+1},$$

whose right-hand-side shows a linear dependence on $c(\alpha)/\alpha$. To obtain (7), it suffices to show that $c(\alpha) \leq c'/\alpha^2$, with c' independent of α. To this aim, we observe that $c(\alpha) = c''/g^2_{K\setminus K_\alpha}$, with

$$g_{K\setminus K_\alpha} := \min_{z \in K \setminus K_\alpha} |g(z)|,$$

$g(z) = \prod_{\lambda \in \Lambda \cap \mathcal{B}(z_0,R)} (z - \lambda)$, and c'' independent of α. Since the frequency response map $\mathcal{S}$ presents only simple poles (given by the Dirichlet-Neumann Laplace eigenvalues), and the interval of interest K contains a finite number of poles of $\mathcal{S}$, it follows that there exists $c_g > 0$ independent of α such that

$$|g(z)| \geq c_g \min_{\lambda \in \Lambda \cap \mathcal{B}(z_0,R)} |z - \lambda| \quad \forall z \in K.$$

Hence, $g_{K\setminus K_\alpha} \geq c_g \alpha$, and (7) follows with $C = c''/c_g^2$.

We can draw the following conclusions:

a. The roots of the LS-Padé denominator $\mathcal{Q}_{[M/N]}$ approximate the N poles of $\mathcal{S}$, closest to z_0.
b. The region of convergence of $\mathcal{S}_{[M/N]}$ is an open disk centered at z_0 (excluding the poles of $\mathcal{S}$), whose radius is equal to the distance between z_0 and the $(N+1)$-th pole of $\mathcal{S}$ closest to z_0.

Remark 2.7 Numerical experiments show that the estimate (7) is not sharp with respect to z, see Fig. 4 below. Indeed, the approximation error at z decreases faster (with respect to M) the closer z and the center of approximation z_0 are: in practice, we observe that a bound of the form

$$\left\| \mathcal{S}(z) - \mathcal{S}(z)_{[M/N]} \right\|_{V,\sqrt{\mathrm{Re}(z_0)}} \leq C \frac{1}{\alpha^3} \left(\frac{|z - z_0|}{R} \right)^{M+1}, \tag{8}$$

holds true with the same constants and for the same values of z as in Theorem 2.5.

3 Algorithmical aspects

In this section, we describe an algorithm for the computation of a LS-Padé approximant (defined according to Definition 2.4) of the Helmholtz frequency response map $\mathcal{S}$ introduced in the previous section. We underline that the presented algorithm can be likewise applied to any V-valued meromorphic map. For instance, since the exact solution of the Helmholtz problem (2) is generally not available in closed form, one could apply LS-Padé approximants to the frequency response map of a FE discretization of (2).

As a first instructive step, we recall the proof of the existence of such an approximant, which was developed in [3, Proposition 4.1].

Proposition 3.1 *For any $M, N \in \mathbb{N}$, $E \geq M+N$, and $\rho \in \mathbb{R}^+$, there exists a LS-Padé approximant centered at z_0.*

Proof We want to show that the minimization problem (6) admits at least one solution. Since P has degree M, then $\big(P(z)\big)_{\alpha,z_0} = 0$ for all $\alpha > M$. Hence, we can rewrite $j_{E,\rho}$ as

$$\begin{aligned} j_{E,\rho}(P,Q)^2 \quad &= \textstyle\sum_{\alpha=0}^{M} \left\| \big(Q(z)\mathcal{S}(z) - P(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\ &+ \textstyle\sum_{\alpha=M+1}^{E} \left\| \big(Q(z)\mathcal{S}(z) - P(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\ &= \textstyle\sum_{\alpha=0}^{M} \left\| \big(Q(z)\mathcal{S}(z) - P(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\ &+ \textstyle\sum_{\alpha=M+1}^{E} \left\| \big(Q(z)\mathcal{S}(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha}. \end{aligned}$$

Now, let Q be fixed. Taking $P = \bar{P}(Q)$, where $\bar{P}(Q)$ satisfies

$$\big(\bar{P}(z)\big)_{\alpha,z_0} = \big(Q(z)\mathcal{S}(z)\big)_{\alpha,z_0} \quad \text{for } 0 \leq \alpha \leq M,$$

problem (6) can be formulated as a minimization problem in Q only: find $Q \in \mathbb{P}^\star_N(\mathbb{C})$ such that

$$\bar{j}_{E,\rho}(Q) \leq \bar{j}_{E,\rho}(S) \quad \forall S \in \mathbb{P}^\star_N(\mathbb{C}), \tag{9}$$

where

$$\bar{j}_{E,\rho}(Q) := j_{E,\rho}(\bar{P}(Q), Q) = \left(\sum_{\alpha=M+1}^{E} \left\| \big(Q(z)\mathcal{S}(z)\big)_{\alpha,z_0} \right\|^2_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \right)^{1/2}. \tag{10}$$

By considering the monomial basis in $\mathbb{P}_N(\mathbb{C})$, it is trivial to show that $\mathbb{P}^\star_N(\mathbb{C})$ is homeomorphic to the unit sphere in $\mathbb{C}^{N+1}$, hence it is compact, see [25]. Thus, since the functional $\bar{j}_{E,\rho}$ is continuous, it has a global minimum on $\mathbb{P}^\star_N(\mathbb{C})$, and the minimization problem (9) admits at least one solution. □

In the following proposition, we express an equivalent formulation of the constrained minimization problem (9).

Proposition 3.2 *The constrained minimization problem (9) is equivalent to the identification of a normalized eigenvector corresponding to the smallest non-negative eigenvalue of the Hermitian positive-semidefinite matrix $G_{E,\rho} \in \mathbb{C}^{(N+1)\times(N+1)}$ with entries*

$$\left(G_{E,\rho}\right)_{i,j} = \sum_{\alpha=M+1}^{E} \left\langle \left(\mathcal{S}\right)_{\alpha-j,z_0}, \left(\mathcal{S}\right)_{\alpha-i,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha}, \qquad i,j = 0,\ldots,N. \tag{11}$$

(We set $\left(\mathcal{S}\right)_{\beta,z_0} = 0$ whenever $\beta < 0$.)

Proof Set $q_\alpha := \left(Q\right)_{\alpha,z_0}$ for $\alpha = 0,\ldots,N$. Since

$$\left(Q\mathcal{S}\right)_{\alpha,z_0} = \sum_{n=0}^{\alpha} q_n \left(\mathcal{S}\right)_{\alpha-n,z_0} = \sum_{n=0}^{N} q_n \left(\mathcal{S}\right)_{\alpha-n,z_0}$$

according to our convention that $\left(\mathcal{S}\right)_{\beta,z_0} = 0$ for $\beta < 0$, we have

$$\begin{aligned}
\bar{j}_{E,\rho}(Q)^2 &= \sum_{\alpha=M+1}^{E} \left\langle \left(Q\mathcal{S}\right)_{\alpha,z_0}, \left(Q\mathcal{S}\right)_{\alpha,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\
&= \sum_{\alpha=M+1}^{E} \left\langle \sum_{j=0}^{N} q_j \left(\mathcal{S}\right)_{\alpha-j,z_0}, \sum_{i=0}^{N} q_i \left(\mathcal{S}\right)_{\alpha-i,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\
&= \sum_{\alpha=M+1}^{E} \sum_{i,j=0}^{N} q_i^* q_j \left\langle \left(\mathcal{S}\right)_{\alpha-j,z_0}, \left(\mathcal{S}\right)_{\alpha-i,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\
&= \sum_{i,j=0}^{N} q_i^* q_j \sum_{\alpha=M+1}^{E} \left\langle \left(\mathcal{S}\right)_{\alpha-j,z_0}, \left(\mathcal{S}\right)_{\alpha-i,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}} \rho^{2\alpha} \\
&= \mathbf{q}^\star G_{E,\rho} \mathbf{q},
\end{aligned}$$

where $G_{E,\rho} \in \mathbb{C}^{(N+1)\times(N+1)}$ is defined in (11), and $\mathbf{q} = (q_0,\ldots,q_N)^T$. By definition, $G_{E,\rho}$ is Hermitian. Moreover, (10) implies that $G_{E,\rho}$ is positive-semidefinite, so that all its eigenvalues are real non-negative. Finally, observe that the constraint $\sum_{\alpha=0}^{N} \left|\left(Q\right)_{\alpha,z_0}\right|^2 = 1$ is equivalent to the condition $\|\mathbf{q}\|_2 = 1$. The claim follows. □

The Hermitian matrix $G_{E,\rho}$ defined in (11) is obtained as weighted sum of submatrices of the Gram matrix $G \in \mathbb{C}^{(E+1)\times(E+1)}$ associated with the solution map $\mathcal{S}$, namely the matrix with entries $G_{i,j} = \left\langle \left(\mathcal{S}\right)_{i,z_0}, \left(\mathcal{S}\right)_{j,z_0} \right\rangle_{V,\sqrt{\mathrm{Re}(z_0)}}$, for $i,j = 0,\ldots,E$. See Fig. 1 for a graphical representation.

By following the steps performed in the proof of Proposition 3.1, and applying Proposition 3.2, we devise Algorithm 1 for the computation of the LS-Padé approximant.

$$G = \left[\begin{array}{cccccc} \langle\mathcal{S},\mathcal{S}\rangle_V & \langle\mathcal{S},\mathcal{S}_1\rangle_V & \langle\mathcal{S},\mathcal{S}_2\rangle_V & \cdots & & \\ \langle\mathcal{S}_1,\mathcal{S}\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_1\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_3\rangle_V & \cdots & \\ \langle\mathcal{S}_2,\mathcal{S}\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_1\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_3\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_4\rangle_V & \cdots \\ \vdots & \langle\mathcal{S}_3,\mathcal{S}_1\rangle_V & \langle\mathcal{S}_3,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_3,\mathcal{S}_3\rangle_V & \langle\mathcal{S}_3,\mathcal{S}_4\rangle_V & \cdots \\ & \vdots & \langle\mathcal{S}_4,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_4,\mathcal{S}_3\rangle_V & \langle\mathcal{S}_4,\mathcal{S}_4\rangle_V & \cdots \\ & & \vdots & \vdots & \vdots & \end{array}\right]$$

$$G_{E,\rho} = \ldots + \rho^6 \begin{bmatrix} \langle\mathcal{S}_3,\mathcal{S}_3\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_3\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_3\rangle_V \\ \langle\mathcal{S}_3,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_2\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_2\rangle_V \\ \langle\mathcal{S}_3,\mathcal{S}_1\rangle_V & \langle\mathcal{S}_2,\mathcal{S}_1\rangle_V & \langle\mathcal{S}_1,\mathcal{S}_1\rangle_V \end{bmatrix} + \ldots$$

Fig. 1 Gram matrix (top) associated with the frequency response map $\mathcal{S}$. To lighten the notation, we omit both the argument (z_0) of the Taylor coefficients $\mathcal{S}_\alpha$, and the weight $\sqrt{\mathrm{Re}\,(z_0)}$ of the scalar product $\langle\cdot,\cdot\rangle_V$. In blue the sub-matrix corresponding to $N = 2$ and $\alpha = 3$, which provides a contribution to $G_{E,\rho}$ (bottom) with weight ρ^6. Observe that a transposition with respect to the secondary diagonal is carried out before computing the sum

Algorithm 1 Construction of the LS-Padé approximant.

```
Fix $z_0 \in \mathbb{C} \setminus \Lambda$ with $\mathrm{Re}\,(z_0) > 0$, $\rho \in \mathbb{R}^+$, $M,\ N,\ E \in \mathbb{N}$, with $E \geq M + N$
Evaluate $\mathcal{S}$ in the center $z_0$
for $\beta = 1, \ldots, E$ do
    Compute the Taylor coefficient of $\mathcal{S}$ at $z_0$ of order $\beta$, $\left(\mathcal{S}\right)_{\beta,z_0}$
end for
Define the matrix $G_{E,\rho} \in \mathbb{R}^{(N+1)\times(N+1)}$ according to (11)
Compute the (normalized) eigenvector $\xi = (\xi_0, \ldots, \xi_N)$ corresponding to the smallest non-negative eigenvalue of the matrix $G_{E,\rho}$
Define the denominator as $\mathcal{Q}_{[M/N]}(z) = \sum_{\alpha=0}^{N} \xi_\alpha (z - z_0)^\alpha$
for $\alpha = 0, \ldots, M$ do
    Compute the Taylor coefficient of $\mathcal{S}\mathcal{Q}_{[M/N]}$ at $z_0$ of order $\alpha$ using the formula $\left(\mathcal{S}\mathcal{Q}_{[M/N]}\right)_{\alpha,z_0} = \sum_{n=0}^{\alpha} \xi_n \left(\mathcal{S}\right)_{\alpha-n,z_0}$
end for
Define the numerator as $\mathcal{P}_{[M/N]}(z) = \sum_{\alpha=0}^{M} \left(\mathcal{S}\mathcal{Q}_{[M/N]}\right)_{\alpha,z_0} (z - z_0)^\alpha$
Define the single-point LS-Padé approximant as $\mathcal{S}_{[M/N]} = \frac{\mathcal{P}_{[M/N]}}{\mathcal{Q}_{[M/N]}}$
```

Remark 3.3 The choice of ρ impacts the algorithm only by determining the weights in the computation of $G_{E,\rho}$. Specifically, small (respectively large) values of ρ emphasize the contributions from the sub-matrices located in the top-left (respectively bottom-right) portion of G.

It is important to observe that Algorithm 1 computes $E + 1$ derivatives (including the 0th order one) of the solution map $\mathcal{S}$, but defines the LS-Padé numerator $\mathcal{P}_{[M/N]}$ as the truncated Taylor polynomial of $\mathcal{S}\mathcal{Q}_{[M/N]}$ only up to order M. The condition $E \geq M + N$ is needed for the derivation above, concerning LS-Padé approximants, due to the specific structure of the functional $j_{E,\rho}$. However, this condition is unnecessary from an implementation viewpoint. This consideration, together with Remark 3.3, motivated the definition of a simplified version of our approximants, named *fast LS-Padé approximation* [4], whose definition is independent of the parameter ρ.

Definition 3.4 Consider $M, N \in \mathbb{N}$, and $E \geq \max\{M, N\}$. A fast LS-Padé approximant $\widetilde{\mathcal{S}}_{[M/N]}$, centered at z_0, of the solution map $\mathcal{S}$ is a quotient $\frac{P}{Q}$ with $P \in \mathbb{P}_M(\mathbb{C}; V)$, $Q \in \mathbb{P}_N^\star(\mathbb{C})$, such that

$$j_E(P, Q) \leq j_E(R, S) \quad \forall R \in \mathbb{P}_M(\mathbb{C}; V), \ \forall S \in \mathbb{P}_N^\star(\mathbb{C}), \tag{12}$$

where the functional $j_E : \mathbb{P}_M(\mathbb{C}; V) \times \mathbb{P}_N^\star(\mathbb{C}) \to \mathbb{R}^+$ is defined as

$$j_E(P, Q) = \left(\sum_{\alpha=0}^{M} \left\| \big(Q(z)\mathcal{S}(z) - P(z)\big)_{\alpha, z_0} \right\|^2_{V, \sqrt{\operatorname{Re}(z_0)}} + \left\| \big(Q(z)\mathcal{S}(z)\big)_{E, z_0} \right\|^2_{V, \sqrt{\operatorname{Re}(z_0)}} \right)^{1/2}. \tag{13}$$

The main differences between fast and standard LS-Padé approximants are:

i. the definition of fast LS-Padé approximants is independent of the parameter ρ;
ii. the number of derivatives of the solution map to be computed is required to be larger than $\max\{M, N\}$, rather than $M + N$ as for the standard definition;
iii. the functional to be minimized, j_E, involves the Taylor coefficients $\big(Q\mathcal{S} - P\big)_{\alpha, z_0}$ for $\alpha \in \{0, \ldots, M-1, M, E\}$, whereas the standard version requires $\alpha \in \{0, \ldots, E-1, E\}$.

Proposition 3.2 can be applied to the simplified functional j_E if one replaces the matrix $G_{E,\rho}$ in (11) by

$$(G_E)_{i,j} = \left\langle \big(\mathcal{S}\big)_{E-j, z_0}, \big(\mathcal{S}\big)_{E-i, z_0} \right\rangle_{V, \sqrt{\operatorname{Re}(z_0)}}, \qquad i, j = 0, \ldots, N. \tag{14}$$

As such, Algorithm 1 can be easily modified to compute fast LS-Padé approximants.

Since the number of computed derivatives $E + 1$ can be as large as the number of Taylor terms in the definition of the fast LS-Padé numerator $\mathcal{P}_{[M/N]}$, we can interpret fast LS-Padé approximation as a more "data-intensive" version of the technique. However, the price to pay for this improved efficiency is the difficulty in developing a theory as general as the one for standard LS-Padé approximants, which has been discussed in Section 2. Nevertheless, in [4] a partial theory for fast LS-Padé approximants has been developed, proving, among others, the one discussed in Remark 2.7, under the following assumption on the target solution map.

Assumption 1 *There exists $z_0 \in \mathbb{C}$, a countable set $\Lambda = \{\lambda_j\} \subset \mathbb{C}$ and a countable family $\{v_j\} \subset V$, with*

(i) Λ *having no finite limit points,*
(ii) $\langle v_i, v_j \rangle_{V, \sqrt{\operatorname{Re}(z_0)}} = 0$ *for all $i \neq j$,*

such that

$$\mathcal{S}(z) = \sum_j \frac{v_j}{\lambda_j - z} \quad \textit{in the } V\textit{-topology.}$$

Assumption 1 is more restrictive than just demanding $\mathcal{S}$ to be meromorphic. In fact, it requires that all poles in Λ are simple, and the residues $\{v_j\}$ are V-orthogonal (these conditions are not fulfilled, for instance, in the problems discussed in Sections 4 and 5. Nonetheless, Assumption 1 can be shown to hold for problems of practical interest, for instance Problem 1. We refer to [4] for a more thorough discussion.

We have not yet commented on how to compute the first $E+1$ Taylor coefficients of $\mathcal{S}$ at z_0, since different maps $\mathcal{S}$ may require different strategies. For our purposes, let us assume (as is the case for the map introduced in Section 1) that $\mathcal{S} : \mathbb{C} \to V$ is implicitly defined by the equation

$$A(z)\mathcal{S}(z) = F(z), \quad \text{for all } z \text{ for which } A(z) \text{ has bounded inverse,} \tag{15}$$

where $A : \mathbb{C} \to \mathcal{L}(V; V)$ and $F : \mathbb{C} \to V$ are holomorphic over a neighborhood of z_0, and $A(z_0)$ has bounded inverse. Then

$$A(z_0)\big(\mathcal{S}\big)_{\beta,z_0} = \big(F\big)_{\beta,z_0} - \sum_{j=0}^{\beta-1} \big(A\big)_{\beta-j,z_0}\big(\mathcal{S}\big)_{j,z_0} \quad \text{for } \beta = 0, 1, \ldots. \tag{16}$$

Computing the first $E+1$ derivatives of $\mathcal{S}$ through (16) requires the solution of $E+1$ problems with the same operator on the left-hand-side. As we discuss in Section 4.2 below, this can be exploited to achieve a better computational efficiency. However, there are also some drawbacks: for instance, since the computation of the β-th derivative involves the ones with order smaller than β, numerical errors accumulate at an exponential rate in β. Thus, small to moderate values of E (depending on the distance between z_0 and the closest pole of $\mathcal{S}$) are necessary to guarantee a stable method.

4 Application to a transmission-reflection problem

In this section, we test LS-Padé approximations on a model transmission-reflection problem involving a fluid-fluid interface, which was originally treated in [22]. Let a plane wave $e^{i\mathbf{k}\cdot\mathbf{x}}$ with wavevector $\mathbf{k} = (\kappa\cos(\theta), \kappa\sin(\theta))$ impinge on the interface between two fluids of different refractive indices $n_1 < n_2$. More precisely, we assume the two fluids to occupy each of the two regions into which $D = (-1, 1)^2$ is partitioned by the horizontal axis. The Helmholtz problem is the following

$$-\Delta u - \kappa^2\varepsilon_r^2 u = 0, \quad \text{with } \varepsilon_r(x_1, x_2) = \begin{cases} n_1 & \text{if } x_2 < 0, \\ n_2 & \text{if } x_2 > 0. \end{cases} \tag{17}$$

For any angle $0 \le \theta < \pi/2$, the following function is a solution of (17):

$$u_{ex}(x_1, x_2) = \begin{cases} T\exp\{i\mathbf{K}\cdot\mathbf{x}\} & \text{if } x_2 > 0, \\ \exp\{in_1\mathbf{k}\cdot\mathbf{x}\} + R\exp\{in_1\mathbf{k}\cdot(x_1, -x_2)\} & \text{if } x_2 < 0, \end{cases} \tag{18}$$

where $\mathbf{K} = (n_1 k_1, \sqrt{(\kappa n_2)^2 - (n_1 k_1)^2})$, $R = \frac{n_1 k_2 - K_2}{n_1 k_2 + K_2}$ and $T = 1 + R$. We couple the Helmholtz equation (17) with Dirichlet boundary conditions derived from the exact solution (18), i.e., $u|_{\partial D} = u_{ex}|_{\partial D}$.

Depending on the value of θ (angle of the incident wave), the solution may exhibit two types of behavior:

- if $\theta < \theta_{crit} := \arccos(n_2/n_1)$, then $\mathrm{Im}(K_2) \neq 0$, and u_{ex} decays exponentially for $x_2 > 0$. Physically, this phenomenon is called *total internal reflection*;
- if $\theta > \theta_{crit}$, then $\mathbf{k}$ is close to normal incidence, and the wave is *refracted* at the interface.

The two behaviors are depicted in Fig. 2.

4.1 Frequency response map

We are interested in the following boundary value problem:

Problem 2 (Transmission-Reflection Problem) *The squared wavenumber κ^2 lies within the interval of interest $K = [\kappa^2_{min}, \kappa^2_{max}]$, and the Helmholtz equation is endowed with Dirichlet boundary conditions on $\Gamma_D = \partial D$:*

$$\begin{cases} -\Delta u - \kappa^2 \varepsilon_r^2 u = 0 & \text{in } D, \\ u = g_D & \text{on } \partial D, \end{cases} \tag{19}$$

where $g_D := u_{ex}|_{\partial D}$, and u_{ex} is given by formula (18) with $\kappa = 11$ and either $\theta = 29°$ or $\theta = 69°$.

A weak formulation of problem (19) with $z \in \mathbb{C}$ replacing κ^2 reads: find $\mathring{u}_z \in V = H_0^1(D)$ such that

$$\int_D \nabla \mathring{u}_z(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) d\mathbf{x} - z \int_D \varepsilon_r^2(\mathbf{x}) \mathring{u}_z(\mathbf{x}) \overline{v}(\mathbf{x}) d\mathbf{x}$$
$$= z \int_D \varepsilon_r^2(\mathbf{x}) w_g(\mathbf{x}) \overline{v}(\mathbf{x}) d\mathbf{x} \quad \forall v \in V, \tag{20}$$

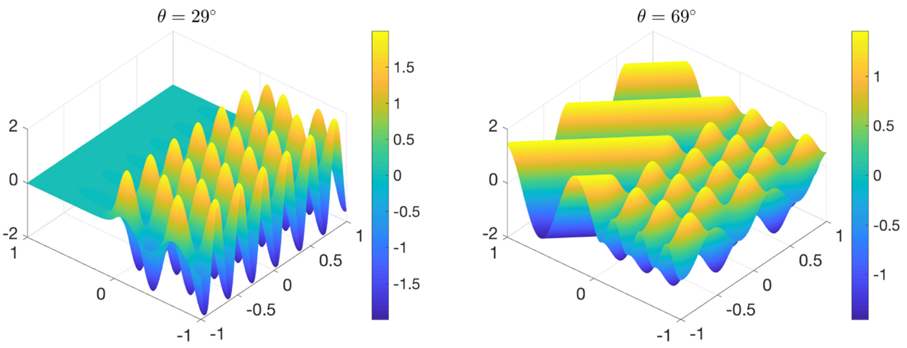

Fig. 2 Exact solution of the transmission-reflection problem with $n_1 = 2$, $n_2 = 1$, $\kappa = 11$ and $\theta = 29°$ (left), $\theta = 69°$ (right)

where $w_g \in H^1(D)$ is the unique harmonic extension of g_D, i.e., $\Delta w_g = 0$ in D and $w_g|_{\partial D} = g_D$, and $\mathring{u} := u - w_g$.

By generalizing [3, Theorem 2.1], it can be proved that problem (20) admits a unique solution for all $z \in \mathbb{C} \setminus \Lambda$, Λ being the set of eigenvalues of the Laplacian (w.r.t. the weighted $L^2(D)$ inner product $\langle u, v\rangle_{\varepsilon_r} = \int_D \varepsilon_r^2(\mathbf{x}) u(\mathbf{x}) \overline{v}(\mathbf{x}) d\mathbf{x}$) with homogeneous Dirichlet boundary conditions. Moreover, given

$$0 < \alpha < \min_{j:\lambda_j \in \Lambda} \left|\lambda_j - z\right|, \tag{21}$$

the unique solution satisfies the a priori bound

$$\|\mathring{u}_z\|_{V,\sqrt{\mathrm{Re}(z_0)}} \leq \max\{1, n_1, n_2\} \frac{\sqrt{|\lambda_{min} - z| + |\mathrm{Re}\,(z)| + \mathrm{Re}\,(z_0)}}{\alpha} |z| \left\|w_g\right\|_{L^2(D)}, \tag{22}$$

where $\lambda_{min} := \min\{\lambda \in \Lambda\}$. By triangular inequality, an analogous upper bound on $\|u_z\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ follows.

Let us denote by $\mathcal{S} : \mathbb{C} \to V := H^1(D)$ the frequency response map that associates to each complex wavenumber squared z, the function $\mathcal{S}(z) = \mathring{u}_z + w_g$, with $\mathring{u}_z$ the weak solution of (20).

Proposition 4.1 *The frequency response map $\mathcal{S}$ is meromorphic in $\mathbb{C}$, having a pole of order one in each $\lambda \in \Lambda$, where Λ is the set of eigenvalues of the Laplacian (w.r.t. the $\langle\cdot,\cdot\rangle_{\varepsilon_r}$-induced topology) with homogeneous Dirichlet boundary conditions.*

Proof Let $\{\varphi_j\}$ be the set of eigenfunctions of the Laplacian w.r.t. the $\langle\cdot,\cdot\rangle_{\varepsilon_r}$-induced topology, and let $\{\lambda_j\}$ be the corresponding eigenvalues, i.e., $-\Delta\varphi_j = \lambda_j \varepsilon_r^2 \varphi_j$ in D and $\varphi_j|_{\partial D} = 0$ (see, e.g., [26, Theorem 2.36]). Plugging $v = \varphi_j$ into (20) and integrating by parts yields

$$z\langle w_g, \varphi_j\rangle_{\varepsilon_r} = \int_D \left(\nabla\mathring{u}_z(\mathbf{x}) \cdot \overline{\nabla\varphi_j}(\mathbf{x}) - z\varepsilon_r^2(\mathbf{x})\mathring{u}_z(\mathbf{x})\overline{\varphi}_j(\mathbf{x})\right) d\mathbf{x} = \left(\lambda_j - z\right) \langle\mathring{u}_z, \varphi_j(\mathbf{x})\rangle_{\varepsilon_r}.$$

Thus, denoting with $w_j = \langle w_g, \varphi_j\rangle_{\varepsilon_r}$, the eigenfunction expansion of $\mathring{u}_z$ reads

$$\mathring{u}_z = \sum_j \frac{z\, w_j}{\lambda_j - z}\varphi_j \tag{23}$$

with convergence in $L^2(D)$ and, due to the regularity of the eigenfunction basis $\{\varphi_j\}$, also in $H^1(D)$. Accordingly,

$$\mathcal{S}(z) = w_g + \sum_j \frac{z\, w_j}{\lambda_j - z}\varphi_j \tag{24}$$

in $H^1(D)$. Hence, $\mathcal{S}$ is meromorphic over $\mathbb{C}$, and each λ_j is a pole of order one for $\mathcal{S}$. □

4.2 LS-Padé approximants of the frequency response map

Since the frequency response map is meromorphic, it is appropriate to use the LS-Padé technology to catch the singularities of $\mathcal{S}$, and provide sharp approximations of

$\mathcal{S}(z)$, when z is close to the center z_0. We apply Algorithm 1 as well as its fast variant, and compute the coefficients of the denominator as the entries of the eigenvector corresponding to the minimal eigenvalue of matrix (11) (or (14) for the fast version). The Taylor coefficient of order $\beta \geq 1$, $\big(\mathcal{S}\big)_{\beta,z_0} = \frac{1}{\beta!}\frac{d^\beta \mathcal{S}}{dz^\beta}|_{z=z_0} \in H_0^1(D)$, satisfies [3]

$$\int_D \nabla\big(\mathcal{S}\big)_{\beta,z_0}(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x})d\mathbf{x} - z_0 \int_D \varepsilon_r^2(\mathbf{x})\big(\mathcal{S}\big)_{\beta,z_0}(\mathbf{x})\overline{v}(\mathbf{x})d\mathbf{x}$$

$$= \int_D \varepsilon_r^2(\mathbf{x})\big(\mathcal{S}\big)_{\beta-1,z_0}(\mathbf{x})\overline{v}(\mathbf{x})d\mathbf{x} \quad \forall v \in H_0^1(D). \tag{25}$$

Problem (25) admits a unique solution for all $z \in \mathbb{C}\setminus\Lambda$, since the PDE operator is the same as in (20) and the right-hand side is a bounded linear form. Since an analytic expression of the exact solution of (20) and (25) is unavailable, in our experiment we replace the solution map $\mathcal{S}$ by its $\mathbb{P}^3$ FE approximation $\mathcal{S}_h$. This discrete version of the frequency response can easily be proven to be meromorphic using the same theoretical tools employed in the previous section.

Let $K = [z_{min}, z_{max}] = [3, 12]$ be the interval of interest and $\theta = 29°$. In order to get convergence of the LS-Padé approximant on the whole interval of interest K, see Theorem 2.5, in all numerical tests we set $\rho = \max\{|z_0 - z_{min}|, |z_0 - z_{max}|\}$ and $E = M + N$. In Fig. 3, the $H^1(D)$-weighted norm of $\mathcal{S}_h$ is compared with the norm of its standard LS-Padé approximant $\mathcal{S}_{h,P}$ centered at $z_0 = 7.5 + 0.5i$, for various degrees (M, N). As expected, we verify that the approximant of type $[M/N]$ can identify with good accuracy N poles of $\mathcal{S}_h$.

In order to obtain each plot, the norms of exact and approximate solutions $\mathcal{S}_h$ and $\mathcal{S}_{h,P}$ were evaluated on a grid of $n = 200$ uniformly spaced (squared) wavenumbers. We remark that the direct evaluation of $\mathcal{S}_h$ in those $n = 200$ points entails assembling and solving n FE systems with different matrices and different right-hand-sides. Overall, due to the simple parametric dependence in (25), only two matrices (mass and stiffness) and one vector need to be assembled to compute the n samples. Thus, the computational cost of this procedure is dominated by the n solves of the FE linear systems.

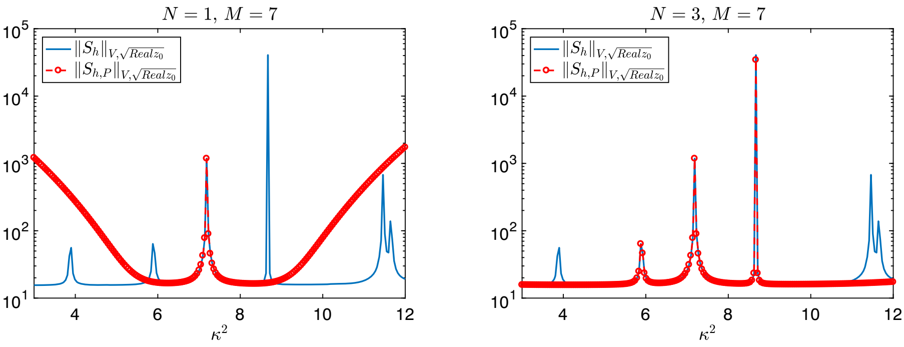

Fig. 3 Comparison between the $H^1(D)$-weighted norm of $\mathcal{S}_h$ (with $\theta = 29°$) and of its LS-Padé approximant $\mathcal{S}_{h,P}$ centered at $z_0 = 7.5 + 0.5i$

In contrast, the evaluation of the approximate map $\mathcal{S}_{h,P}$ in the same $n = 200$ points has a computational cost that is dominated by the (offline) computation of the LS-Padé approximant. Such operation requires the assembly of a single FE matrix and of a single right-hand-side vector, and the solution of $E + 1$ linear systems. The matrix of the system is the same for all problems, so one can, as a preliminary step, factorize it (or compute a preconditioner in case of an iterative solver) to speed up each subsequent linear system solve. Then, the online computation only requires n inexpensive evaluations of 2 polynomials (one V-valued and one $\mathbb{C}$-valued), and no expensive FE computation is needed anymore.

In Fig. 4 (left), we perform a quantitative convergence analysis of the approximation error $\left\|\mathcal{S}_h(z) - \mathcal{S}_{h,P}(z)\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ at $z = 8$, with respect to the number E of computed derivatives. The solid lines show the error achieved by the standard LS-Padé approximants for $N = 1, 2, 3$ and $M = E - N$. In all three cases, the error bound in Theorem 2.5 turns out not to be sharp. In fact, the observed convergence rates are as in (8) instead.

On the same plot, we show by dashed lines the approximation error obtained by applying fast LS-Padé approximants with $N = 1, 2, 3$ and $M = E \geq N$. Even though the theoretical Assumption 1 is not satisfied for this problem, the numerical results show that the fast LS-Padé approximation technique leads to the same exponential convergence rate with respect to E as the standard approximants; we refer to [4] for a full discussion and theoretical justification on the observed rate (8) in the case of fast LS-Padé approximants. On the other hand, the magnitude of the error is much smaller in the fast case, particularly for larger N. This is to be expected since, for fixed E and N, the fast method uses a higher polynomial degree $M = E$ of the numerator than the standard method, which uses $M = E - N$, instead.

Having numerically verified the exponential convergence of the error for fixed N, we consider now the diagonal case, i.e. approximants with $M = N$. While it is definitely trickier to derive sharp convergence bounds in this case, such approach can be convenient in practical applications, e.g. when the number of poles inside the region of interest is unknown. In Fig. 4 (right), we show the error achieved by diagonal LS-Padé approximants (both in their standard and fast version). The convergence of fast approximants is faster, due to the fact that only one new sample is required to increase

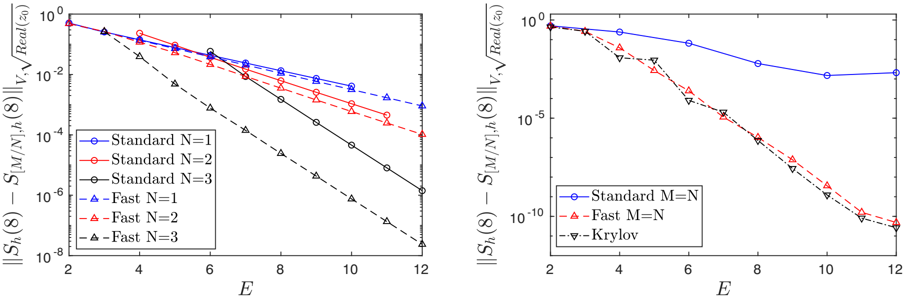

Fig. 4 Approximation error at $z = 8$. On the left, LS-Padé approximants (solid) are compared with their fast version (dashed) for fixed $N = 1, 2, 3$. On the right, diagonal ($M = N$) LS-Padé approximants (blue and red, standard and fast respectively) are compared with a projection-based method relying on Krylov subspaces (black)

both M and N by one, while standard approximants require two new samples to achieve the same.

In the same plot, we also compare LS-Padé approximation with its natural state-of-the-art benchmark, namely the Krylov subspace method (see, e.g., [7, 9]). This method obtains a surrogate for $\mathcal{S}$ by projecting the algebraic version of the original problem (20) onto a Krylov subspace of dimension $E + 1$, and then by solving the resulting problem (of size $E + 1$. Referring to (15) and its FE approximation

$$A_h(z)\mathcal{S}_h(z) = F_h(z),$$

for our comparison we have used the Krylov subspace of dimension $E + 1$

$$K_{E+1} = span\{A_h^{-1}(z_0)F_h(z_0), \ldots, A_h^{-(E+1)}(z_0)F_h(z_0)\}$$

which coincides with the span of the first $E + 1$ derivatives of the solution map $\mathcal{S}$. In this way, the three considered numerical methods employ the same offline information, i.e. they build approximants starting from the same finite dimensional subspace.

The numerical results show that fast LS-Padé approximation and the Krylov subspace method achieve comparable results. However, the former technique could be considered superior to the latter for several reasons: firstly, due to its explicit nature, it has a lower online cost; also, it allows approximation of general problems, possibly with non-linearities, and with non-affine parametric dependence. The main price to pay for the increased applicability of the method is a lower stability: this can be observed in Fig. 5, where the error achieved at (among other points) the two extrema

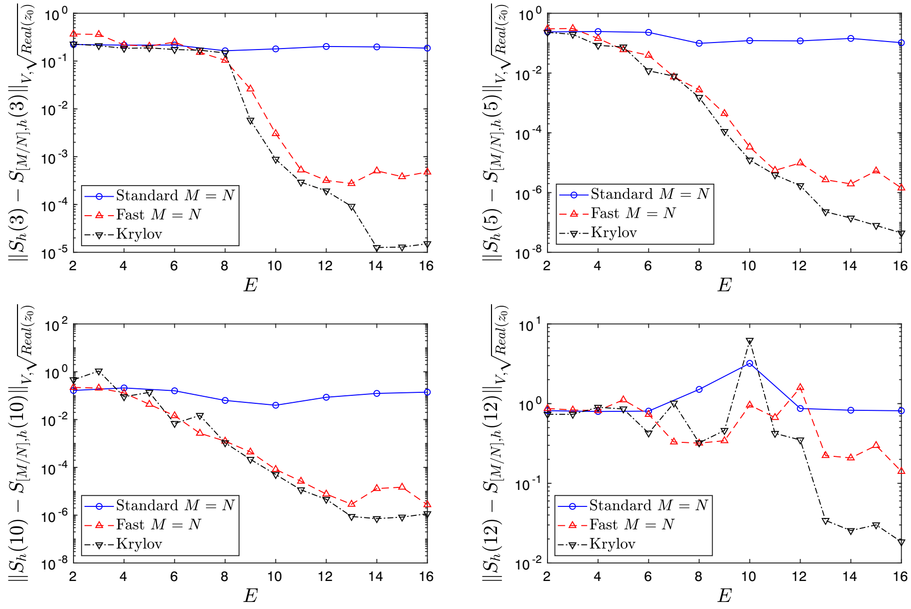

Fig. 5 Approximation error at $z = z_{min} = 3$ (top left), $z = 5$ (top right), $z = 10$ (bottom left), and $z = z_{max} = 12$ (bottom right). Diagonal ($M = N$) LS-Padé approximants (blue and red, standard and fast respectively) are compared with a projection-based method relying on Krylov subspaces (black)

of the parameter range is shown. We can observe that standard LS-Padé approximants do not achieve a reasonable accuracy in either case, whereas their fast version seems to yield better results, comparable with the ones obtained by the Krylov subspace method.

For fixed E, all methods rely on the same information, namely the same number of derivatives of $\mathcal{S}_h$ at z_0. As such, the difference in the results can be mainly explained in terms of conditioning issues, which affect the three methods at different degrees. In the Krylov subspace method, the ill-conditioning is quite mild (mostly due to the implicit enforcement of the moment-matching conditions), influencing only the projection step and, possibly, the resolution of the reduced system. Instead, in LS-Padé approximation, conditioning problems arise in the optimization of (4)–(13), once in the computation of the denominator via SVD, and once in the construction of the numerator via explicit moment-matching. The overall impact of ill-conditioning on these two techniques appears heavier, particularly for the standard version of the method.

As a countermeasure to these issues, we believe necessary a change of paradigm in the choice of approximation strategy. Namely, what we envision is a *multi-point* extension of LS-Padé approximation, relying on evaluations of the solution map $\mathcal{S}$ and, possibly, of its derivatives at multiple parameter values. Among the advantages of such strategy, we expect a reduced ill-conditioning, the ability to estimate a larger number of poles, and, more generally, an improved effectiveness in approximating the solution map on wider parameter ranges. A preliminary investigation of this technique can be found in [29], and some numerical results comparing it with the multi-moment-matching method [9] are available in [5].

5 Application to a scattering problem

In this section, we consider the two-dimensional scattering of an acoustic wave on an object occupying the domain $B = \mathcal{B}(\mathbf{0}, 0.5) \subset \mathbb{R}^2$ (ball of radius 0.5 centered at the origin). The incident wave u^i is the time-harmonic plane wave with wavevector $\mathbf{k} = (k\cos(\theta), k\sin(\theta))$ and unit amplitude, i.e., $u^i(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}$. The total field u, given by the sum of the incident wave u^i with the scattered wave u^s, satisfies the following boundary value problem in the infinite domain $\mathbb{R}^2 \setminus B$

$$\begin{cases} -\Delta u - k^2 u = 0 & \text{in } \mathbb{R}^2 \setminus \overline{B}, \\ u = 0 & \text{on } \Gamma_D := \partial B, \\ \lim_{|\mathbf{x}|\to\infty} |\mathbf{x}|^{1/2} \left(\frac{\partial u^s(\mathbf{x})}{\partial |\mathbf{x}|} - iku^s(\mathbf{x}) \right) = 0. \end{cases} \tag{26}$$

The FE approximation of problem (26) entails the truncation of the unbounded domain $\mathbb{R}^2 \setminus B$ into the bounded domain

$$D := ([-2, 2] \times [-2, 2]) \setminus B,$$

whose outer boundary will be denoted by Γ_R. Approximating the Sommerfeld radiation condition at infinity in problem (26) by a first order absorbing boundary condition, we write the following parametric problem:

Problem 3 (**Scattering Problem**) *The wavenumber k ranges in the interval of interest $K := [k_{min}, k_{max}] \subset \mathbb{R}^+$, $\mathbf{n}$ is the outgoing normal vector field to Γ_R, and $g_k := \frac{\partial u^i}{\partial \mathbf{n}} - iku^i$ is the impedance trace of the incoming wave u^i. We consider the Helmholtz boundary value problem*

$$\begin{cases} -\Delta u - k^2 u = 0 & \text{in } D, \\ u = 0 & \text{on } \Gamma_D, \\ \frac{\partial u}{\partial \mathbf{n}} - iku = g_k & \text{on } \Gamma_R. \end{cases} \tag{27}$$

5.1 Regularity of the frequency response map

We extend problem (27) to complex wavenumbers. Given a complex wavenumber $z \in \mathbb{C}$, we introduce the incident plane wave $u^i = e^{i\mathbf{k}\cdot\mathbf{x}}$ and its impedance trace $g_z := \frac{\partial u^i}{\partial \mathbf{n}} - izu^i$, and we define the frequency response map $\mathcal{S} : z \mapsto \mathcal{S}(z) := u_z \in V := H^1_{\Gamma_D}(D)$, where u_z satisfies

$$\int_D \nabla u_z(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) d\mathbf{x} - z^2 \int_D u_z(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} - iz \int_{\Gamma_R} u_z(\mathbf{x})\overline{v}(\mathbf{x}) ds$$
$$= \int_{\Gamma_R} g_z(\mathbf{x})\overline{v}(\mathbf{x}) ds \quad \forall\, v \in V. \tag{28}$$

If $z \in \mathbb{R}$, problem (28) admits a unique solution (see, e.g., [14]), which implies that the frequency response map is well-defined on $\mathbb{R}$. The following Theorem extends this result to the complex half plane $\{z \in \mathbb{C} : \operatorname{Im}(z) \geq 0\}$. Since the wavenumber in (28) coincides with the parameter z, we will endow the Hilbert space V with the weighted $H^1(D)$-norm, with weight $w = \operatorname{Re}(z_0)$ (and not $w = \sqrt{\operatorname{Re}(z_0)}$, as was done before).

Theorem 5.1 *Problem (28) admits a unique solution in all compact subsets of*

$$\mathbb{C}^+ := \{z \in \mathbb{C} : \operatorname{Im}(z) \geq 0\}. \tag{29}$$

Proof Given $z \in \mathbb{C}$, we introduce the bilinear and linear forms which define problem (28):

$$B_z(u, v) := \int_D \nabla u_z(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) d\mathbf{x} - z^2 \int_D u_z(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} - iz \int_{\Gamma_R} u_z(\mathbf{x})\overline{v}(\mathbf{x}) ds, \tag{30}$$

$$L_z(v) := \int_{\Gamma_R} g_z(\mathbf{x})\overline{v}(\mathbf{x}) ds. \tag{31}$$

We first show that either the coercivity or the Gårding inequality (see [26]) holds, provided that $\operatorname{Im}(z)$ is non-negative. For the bilinear form in (30), we have

$$\begin{aligned} \operatorname{Re}(B_z(u, u)) &= \|\nabla u\|^2_{L^2(D)} - (\operatorname{Re}(z)^2 - \operatorname{Im}(z)^2)\, \|u\|^2_{L^2(D)} + \operatorname{Im}(z)\, \|u\|^2_{L^2(\Gamma_R)} \\ &\geq \|\nabla u\|^2_{L^2(D)} - (\operatorname{Re}(z)^2 - \operatorname{Im}(z)^2)\, \|u\|^2_{L^2(D)} \\ &= \|u\|^2_{V,\operatorname{Re}(z_0)} - (\operatorname{Re}(z)^2 - \operatorname{Im}(z)^2 + \operatorname{Re}(z_0)^2)\, \|u\|^2_{L^2(D)}\,. \end{aligned}$$

If $C := \mathrm{Re}\,(z)^2 - \mathrm{Im}\,(z)^2 + \mathrm{Re}\,(z_0)^2 \leq 0$, then $B(\cdot,\cdot)$ is coercive, whereas if $C > 0$, then $B_z(\cdot,\cdot)$ satisfies the Gårding inequality.

The bilinear form (30) is bounded, with constant $C = \max\left\{1, \frac{|z|^2}{\mathrm{Re}(z_0)}, \frac{|z|C_{tr}^2}{\mathrm{Re}(z_0)}\right\}$. Indeed, using the trace inequality

$$\|u\|_{L^2(\Gamma_R)} \leq C_{tr}\,\|u\|_{H^1(D)}\,,$$

we get

$$\begin{aligned}
|B_z(u,v)| &\leq \|\nabla u\|_{L^2(D)}\,\|\nabla v\|_{L^2(D)} + |z|^2\,\|u\|_{L^2(D)}\,\|v\|_{L^2(D)} + |z|\,\|u\|_{L^2(\Gamma_R)}\,\|v\|_{L^2(\Gamma_R)}\\
&\leq \|\nabla u\|_{L^2(D)}\,\|\nabla v\|_{L^2(D)} + |z|^2\,\|u\|_{L^2(D)}\,\|v\|_{L^2(D)} + |z|\,C_{tr}^2\,\|u\|_{H^1(D)}\,\|v\|_{H^1(D)}\\
&\leq \|\nabla u\|_{L^2(D)}\,\|\nabla v\|_{L^2(D)} + \frac{|z|^2\,\mathrm{Re}\,(z_0)^2}{\mathrm{Re}\,(z_0)^2}\,\|u\|_{L^2(D)}\,\|v\|_{L^2(D)}\\
&\quad + |z|\,C_{tr}^2\,\max\left\{1, \frac{1}{\mathrm{Re}\,(z_0)^2}\right\}\|u\|_{V,\mathrm{Re}(z_0)}\,\|v\|_{\mathrm{Re}(z_0)}\\
&\leq \max\left\{1, \frac{|z|^2}{\mathrm{Re}\,(z_0)^2}, \frac{|z|\,C_{tr}^2}{\mathrm{Re}\,(z_0)^2}\right\}\|u\|_{V,\mathrm{Re}(z_0)}\,\|v\|_{\mathrm{Re}(z_0)}\,.
\end{aligned}$$

Moreover, the linear functional (31) is bounded, with constant

$$C = C_{tr}^2\,\max\left\{1, \frac{1}{\mathrm{Re}\,(z_0)^2}\right\}\|g_z\|_{V,\mathrm{Re}(z_0)}\,.$$

Problem (28) admits a unique solution (continuously dependent on the data) if and only if its homogeneous adjoint problem admits only trivial solutions: see [26, Theorem 4.11]. We consider the case $\mathrm{Im}\,(z) > 0$, and we refer to [14] for $\mathrm{Im}\,(z) = 0$. The bilinear form associated with the adjoint problem with $g_z = 0$ reads:

$$B_z^*(\varphi,v) := \overline{B_z(v,\varphi)} = \int_D \nabla\varphi(\mathbf{x})\cdot\overline{\nabla v}(\mathbf{x})d\mathbf{x} - \overline{z}^2\int_D \varphi(\mathbf{x})\overline{v}(\mathbf{x})d\mathbf{x} - \overline{iz}\int_{\Gamma_R}\varphi(\mathbf{x})\overline{v}(\mathbf{x})ds,$$

and the condition $B_z^*(u,u) = 0$ is equivalent to

$$\begin{cases}
\mathrm{Re}\left(B_z^*(u,u)\right) = \|\nabla u\|_{L^2(D)}^2 - (\mathrm{Re}\,(z)^2 - \mathrm{Im}\,(z)^2)\,\|u\|_{L^2(D)}^2 + \mathrm{Im}\,(z)\,\|u\|_{L^2(\Gamma_R)}^2 = 0\\
\mathrm{Im}\left(B_z^*(u,u)\right) = \mathrm{Re}\,(z)\left(2\mathrm{Im}\,(z)\,\|u\|_{L^2(D)}^2 + \|u\|_{L^2(\Gamma_R)}^2\right) = 0
\end{cases}$$

If $\mathrm{Re}\,(z) \neq 0$ and $\mathrm{Im}\,(z) > 0$, then $\mathrm{Im}\left(B_z^*(u,u)\right) = 0$ is equivalent to $\|u\|_{L^2(D)} = \|u\|_{L^2(\Gamma_R)} = 0$, that is, $u = 0$ in D, whereas, if $\mathrm{Re}\,(z) = 0$ and $\mathrm{Im}\,(z) > 0$, then $\mathrm{Re}\left(B_z^*(u,u)\right) = 0$ implies $\|\nabla u\|_{L^2(D)} = \|u\|_{L^2(D)} = \|u\|_{L^2(\Gamma_R)} = 0$, hence $u = 0$. □

We recall here the following theorem, see [33, Theorem 1], which will be used in the proof of Proposition 5.3.

Theorem 5.2 *Let B be an open and connected subset of the complex plane. Given $\{T(z)\}_{z\in B}$ an analytic family of compact operators defined on a given Banach space, either $(I - T(z))$ is nowhere invertible over B or $(I - T(z))^{-1}$ is meromorphic over B.*

Proposition 5.3 *The frequency response map $\mathcal{S}$ associated with problem (28) is meromorphic in all open bounded and connected subsets of $\mathbb{C}$, and all its poles have negative imaginary part.*

Proof We proceed as in [24, Proposition 2]. We add and subtract the term $\int_D u_z \overline{v} dx$ to the left-hand side of (28), and we get

$$\int_D \nabla u_z(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) d\mathbf{x} + \int_D u_z(\mathbf{x})\overline{v}(\mathbf{x}) dx - (1+z^2) \int_D u_z(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x}$$
$$-iz \int_{\Gamma_R} u_z(\mathbf{x})\overline{v}(\mathbf{x}) ds = \int_{\Gamma_R} g_z(\mathbf{x})\overline{v}(\mathbf{x}) ds \quad \forall\, v \in V,$$

which can be written equivalently as

$$(I - T(z))u_z = G_z \quad \text{in } V, \tag{32}$$

where $T(z), G_z : V \to V$ are defined, respectively, as

$$\begin{aligned} \langle T(z)u, v\rangle_V &= (1+z^2) \int_D u(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} + iz \int_{\Gamma_R} u(\mathbf{x})\overline{v}(\mathbf{x}) ds \quad \forall\, v \in V, \\ \langle G_z, v\rangle_V &= \int_{\Gamma_R} g_z \overline{v}(\mathbf{x}) ds \quad \forall\, v \in V. \end{aligned}$$

Therefore, $\mathcal{S}(z) = (I - T(z))^{-1} G_z$. We prove that $T(z)$ is compact in all open bounded connected subsets of the complex plane $\mathbb{C}$. We write $T(z)$ as $T(z) = \widetilde{T}(z) \circ J$, where J is the compact embedding $J : V \to H^{1/2+\varepsilon}(D)$, and $\widetilde{T}(z) : H^{1/2+\varepsilon}(D) \to V$. Hence, in order to prove the compactness of $T(z)$, it is enough to show that $\widetilde{T}(z)$ is continuous. For all $u \in H^{1/2+\varepsilon}(D)$, we have

$$\begin{aligned} \left\|\widetilde{T}(z)u\right\|_V &= \sup_{v\in V, \|v\|_V=1} \left|\left\langle \widetilde{T}(z)u, v\right\rangle_V\right| \\ &= \sup_{v\in V, \|v\|_V=1} \left|(1+z^2) \int_D u(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} + iz \int_{\Gamma_R} u(\mathbf{x})\overline{v}(\mathbf{x}) ds\right| \\ &\le \sup_{\|v\|_V=1} \left(\left|1+z^2\right| \|u\|_{L^2(D)} \|v\|_{L^2(D)} + |z| \|u\|_{L^2(\Gamma_R)} \|v\|_{L^2(\Gamma_R)}\right) \\ &\le \sup_{\|v\|_V=1} \left(\left|1+z^2\right| \|u\|_{L^2(D)} \|v\|_{L^2(D)} + |z| \|u\|_{L^2(\partial D)} \|v\|_{L^2(\partial D)}\right) \\ &\le \sup_{\|v\|_V=1} \left(\left|1+z^2\right| \|u\|_{L^2(D)} \|v\|_{L^2(D)} + C_{tr}^2 |z| \|u\|_{H^{1/2+\varepsilon}(D)} \|v\|_{H^{1/2+\varepsilon}(D)}\right) \\ &\le \max\{\left|1+z^2\right|, C_{tr}^2 |z|\} \|u\|_{H^{1/2+\varepsilon}(D)}, \end{aligned}$$

where C_{tr} is the continuity constant of the trace operator $\gamma : H^{1/2+\varepsilon}(D) \to L^2(\partial D)$ (see, e.g., [1, Theorem 5.36]). Applying Theorem 5.2, we conclude that $(I - T(z))^{-1}$ is meromorphic in all open bounded and connected subsets of $\mathbb{C}$ and, since G_z is linear in z (hence holomorphic in $\mathbb{C}$), the same conclusion applies to the frequency response function $\mathcal{S}(z) = (I - T(z))^{-1} G_z$. Moreover, since Theorem 5.1 states that $\mathcal{S}$ is well defined in $\mathbb{C}^+$, we deduce that all poles of $\mathcal{S}$ must have negative imaginary part. □

5.2 LS-Padé approximant of the frequency response map

We construct the LS-Padé approximants of the frequency response map $\mathcal{S}$ following Algorithm 1 and its fast variant. Having fixed $z_0 \in \mathbb{C}^+$, N, M, and $E \geq M+N$ ($E \geq \max\{M, N\}$ for the fast version), the coefficients of the denominator are computed by identifying the eigenvector corresponding to the smallest eigenvalue of the matrix (11) or (14), where the β-th Taylor coefficient of $\mathcal{S}$, $\left(\mathcal{S}\right)_{\beta,z_0}$, solves the following recursive problem:

$$\int_D \nabla\left(\mathcal{S}\right)_{\beta,z_0}(\mathbf{x}) \cdot \overline{\nabla v}(\mathbf{x}) d\mathbf{x} - z_0^2 \int_D \left(\mathcal{S}\right)_{\beta,z_0}(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} - i z_0 \int_{\Gamma_R} \left(\mathcal{S}\right)_{\beta,z_0}(\mathbf{x})\overline{v}(\mathbf{x}) ds$$
$$= 2z_0 \int_D \left(\mathcal{S}(\mathbf{x})\right)_{\beta-1,z_0}\overline{v}(\mathbf{x}) d\mathbf{x} + i \int_{\Gamma_R} \left(\mathcal{S}(\mathbf{x})\right)_{\beta-1,z_0}\overline{v}(\mathbf{x}) ds$$
$$+ \int_D \left(\mathcal{S}(\mathbf{x})\right)_{\beta-2,z_0}\overline{v}(\mathbf{x}) d\mathbf{x} + \frac{1}{\beta!}\int_{\Gamma_R} \frac{d^\beta}{dz} g_z(\mathbf{x})|_{z=z_0} \cdot \overline{v}(\mathbf{x}) ds \quad \forall\, v \in V. \qquad (33)$$

Since the PDE operator in (33) is the same as in (28), and the linear form on the right-hand-side is bounded, by applying Theorem 5.1, we conclude that problem (33) is well-posed for any $z \in \mathbb{C}^+$.

Let u^i be an incident wave with wavevector $\mathbf{k} = (2, 0)$, and fix $z_0 = 3 + 0.5i$ and $K = [2, 4]$. In all numerical tests we set $\rho = |z_0 - 2|$ - to get convergence on the whole interval of interest K - and $E = M + N$. In Fig. 6 we plot the relative LS-Padé approximation error, for both the standard (solid lines) and fast (dashed lines) methods, at $z = 2$ versus the degree of the LS-Padé numerator (left) and the number of computed derivatives (right), for different values of denominator degree. Also, the diagonal LS-Padé approximant is considered. Even though Assumption 1 is not satisfied for this problem, the fast LS-Padé approximant can be computed.

As in Section 4.2, a subspace projection method is also considered (dash-dotted lines). Once more, the approximate solution is obtained by projecting the original problem onto the subspace spanned by the Taylor coefficients of (the FE approximation of) $\mathcal{S}$:

$$V_{E+1} = span\{\left(\mathcal{S}_h\right)_{0,z_0}, \left(\mathcal{S}_h\right)_{1,z_0}, \ldots, \left(\mathcal{S}_h\right)_{E,z_0}\}.$$

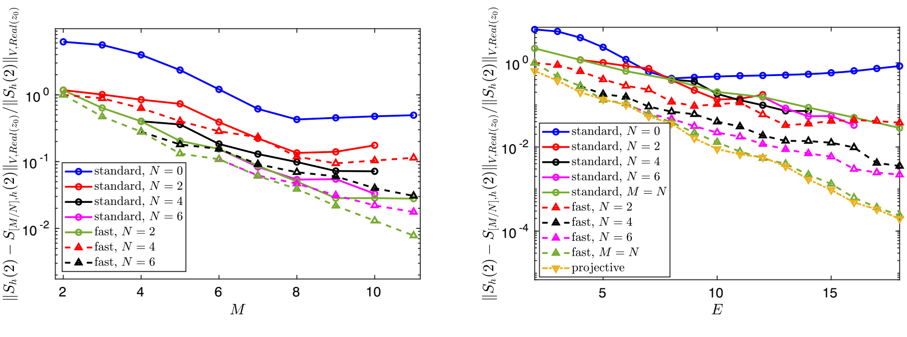

Fig. 6 Relative approximation error at $z = 2$ plotted versus M (left) and E (right) for the standard (solid lines) and fast (dashed) LS-Padé approximants

Note that, due to the nonlinear dependence on z, the subspace V_{E+1} does not satisfy the Krylov property. As such, the procedure we actually employ is somewhat closer to a POD-reduced basis approach [34].

We remark that the approximant for $N = 0$, i.e., the truncated Taylor series of $\mathcal{S}$ with center z_0, does not converge. This indicates the lack of holomorphy of $\mathcal{S}$ within the disk centered at z_0 with radius $|2 - z_0|$. Moreover, both the standard and the fast LS-Padé approximants with $N = 2$ yield approximation errors that stagnate for large M. This suggests that the disk centered at z_0 with radius $|2 - z_0|$ contains more than two poles, and a higher denominator degree should be employed. Indeed, the LS-Padé approximation error decreases exponentially in M with the choice $N = 4, 6$.

The two versions of LS-Padé approximants yield similar results for the same values of M and N (see Fig. 6 left). In Fig. 6 right, the same quantities are plotted versus E: fast approximants yield smaller approximation errors than standard approximants for the same number of computed derivatives.

In the same figure, the Padé approximation techniques are also compared with the above-mentioned projective method. We notice again that the performances of the fast diagonal LS-Padé and the projective method are quite similar.

6 LS-Padé approximant of the stochastic model problem

This section deals with the stochastic counterpart of Problem 1:

Problem 4 (**Stochastic Model Problem**) *The squared wavenumber k^2 of the Helmholtz equation is modeled as a random variable with bounded probability density function $\mathscr{F}_{k^2}$. In this section, either Dirichlet or Neumann or mixed Dirichlet-Neumann homogeneous boundary conditions on ∂D are considered.*

We introduce a Lipschitz functional $\mathcal{L} : V \to \mathbb{R}$ representing a quantity of interest of the frequency response map $\mathcal{S}$, and we define the following two random variables:

$$X := \mathcal{L}(\mathcal{S}(k^2)) \tag{34}$$

and

$$X_P := \mathcal{L}(\mathcal{S}_{[M/N]}(k^2)) \tag{35}$$

where $\mathcal{S}_{[M/N]} = \frac{\mathcal{P}_{[M/N]}}{\mathcal{Q}_{[M/N]}}$ is the LS-Padé approximant of $\mathcal{S}$ centered at z_0, with $\mathrm{Re}\,(z_0) = \frac{k^2_{min}+k^2_{max}}{2}$ and $\mathrm{Im}\,(z_0) \neq 0$; this guarantees that $z_0 \notin \Lambda$, Λ being the set of eigenvalues of the Laplacian, with the considered boundary conditions.

Note that, due to the pole singularities of the solution map $\mathcal{S}$ as well as of its Padé approximant $\mathcal{S}_{[M/N]}$, the random variables X and X_P have no finite statistical moments in general. On the other hand, their characteristic functions $\phi_X(t) := \mathbb{E}\left[e^{itX}\right]$ and $\phi_{X_P}(t) := \mathbb{E}\left[e^{itX_P}\right]$ are well-defined on $\mathbb{R}$. In Theorem 6.1 we state the pointwise convergence of $\phi_{X_P}(t)$ to $\phi_X(t)$ for any $t \in \mathbb{R}$, as $M \to +\infty$. The upper bound (36) in Theorem 6.1 implies, in particular, the uniform exponential convergence of $\phi_{X_P}(t)$ to $\phi_X(t)$ on any compact subset of $\mathbb{R}$ (see Corollary 6.2), from which the convergence in distribution of X_P to X follows, as $M \to +\infty$, follows (see Corollary 6.2).

Theorem 6.1 *Let $\mathcal{L} : V \to \mathbb{R}$ be a Lipschitz functional with Lipschitz constant L, and let X, X_P be the random variables defined in (34) and (35), respectively. Given $\alpha > 0$, then it holds*

$$\left|\phi_X(t) - \phi_{X_P}(t)\right| \leq \left(2\left|K_\alpha\right| + |t|\; L\, C\frac{1}{\alpha^3}\left(\frac{\rho}{R}\right)^{M+1}|K|\right)\sup_{x\in K}\mathscr{F}_{k^2}(x), \quad \forall\, t \in \mathbb{R}, \tag{36}$$

with the same definitions of R, ρ, and K_α, and the same characterization of $C > 0$ as in Theorem 2.5, and $|\cdot|$ denoting the Lebesgue measure.

Proof Using the definition of the characteristic function and the linearity of the expected value we find

$$\begin{aligned}\left|\phi_X(t) - \phi_{X_P}(t)\right| &= \left|\mathbb{E}\left[e^{itX}\right] - \mathbb{E}\left[e^{itX_P}\right]\right| = \left|\mathbb{E}\left[e^{itX} - e^{itX_P}\right]\right| \\ &= \left|\int_K \left(e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right| \\ &\leq \left|\int_{K_\alpha} \left(e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right| \\ &\quad + \left|\int_{K\setminus K_\alpha} \left(e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right|.\end{aligned}$$

We bound the two integrals separately. For the integral over K_α, we have

$$\begin{aligned}&\left|\int_{K_\alpha} \left(e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right| \\ &\quad\leq \int_{K_\alpha}\left|e^{it\mathcal{L}(\mathcal{S}(x))}\right|\mathscr{F}_{k^2}(x)dx + \int_{K_\alpha}\left|e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right|\mathscr{F}_{k^2}(x)dx \leq 2\left|K_\alpha\right|\sup_{x\in K_\alpha}\mathscr{F}_{k^2}(x).\end{aligned} \tag{37}$$

Consider now the integral over $K \setminus K_\alpha$. Since e^{itx} is Lipschitz as a function of x with constant $|t|$, and $\mathcal{L}$ is Lipschitz with constant L, we find

$$\begin{aligned}&\left|\int_{K\setminus K_\alpha} \left(e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right| \\ &\quad\leq \int_{K\setminus K_\alpha}\left|e^{it\mathcal{L}(\mathcal{S}(x))} - e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right|\mathscr{F}_{k^2}(x)dx \\ &\quad\leq |t|\int_{K\setminus K_\alpha}\left|\mathcal{L}(\mathcal{S}(x)) - \mathcal{L}(\mathcal{S}(x))_{[M/N]}\right|\mathscr{F}_{k^2}(x)dx \\ &\quad\leq |t|\; L\int_{K\setminus K_\alpha}\left\|\mathcal{S}(x) - \mathcal{S}(x)_{[M/N]}\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}\mathscr{F}_{k^2}(x)dx.\end{aligned}$$

From the bound (7) of Theorem 2.5, we obtain

$$\begin{aligned}\left|\int_{K\setminus K_\alpha}\left(e^{it\mathcal{L}(\mathcal{S}(x))}-e^{it\mathcal{L}(\mathcal{S}_{[M/N]}(x))}\right)\mathscr{F}_{k^2}(x)\,dx\right| \\ \le |t|\;L\,C\frac{1}{\alpha^3}\left(\frac{\rho}{R}\right)^{M+1}|K|\sup_{x\in K\setminus K_\alpha}\mathscr{F}_{k^2}(x).\end{aligned} \tag{38}$$

The conclusion follows from inequalities (37) and (38). □

The following corollary establishes uniform exponential convergence of ϕ_{X_P} to ϕ_X on any compact subset of $\mathbb{R}$.

Corollary 6.2 *Under the same assumptions as in Theorem 6.1, it holds*

$$\lim_{M\to\infty}\left|\phi_X(t)-\phi_{X_P}(t)\right|=0\quad\forall t\in\mathbb{R}. \tag{39}$$

In particular, there exists $C>0$ such that for any $t\in\mathbb{R}$

$$\left|\phi_X(t)-\phi_{X_P}(t)\right|\le C\,|t|^{1/4}\left(\frac{\rho}{R}\right)^{\frac{M+1}{4}}.$$

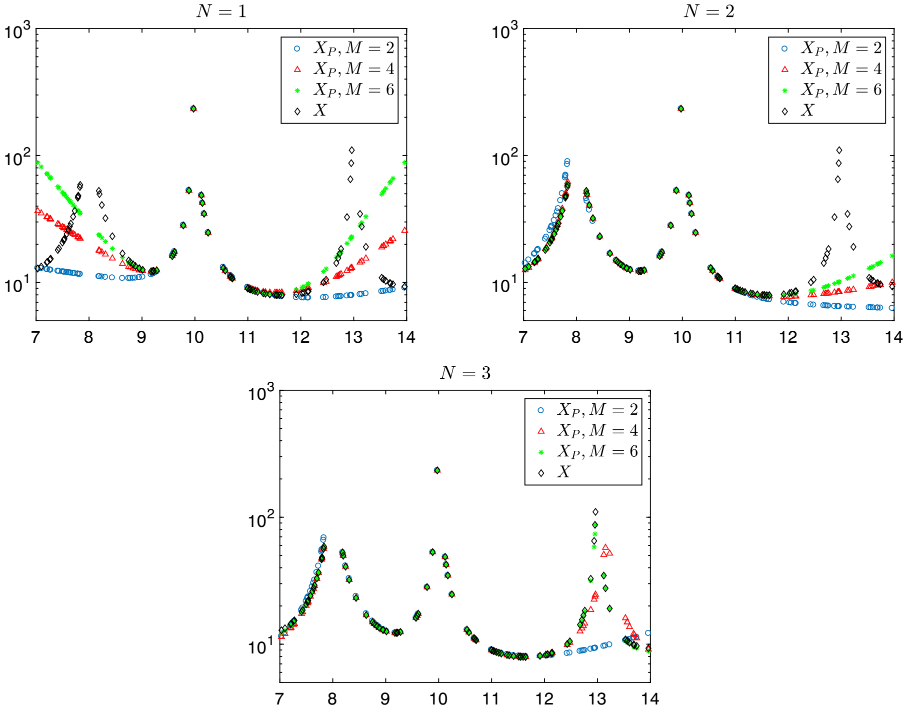

Fig. 7 Comparison between $X=\left\|\mathcal{S}_h(k^2)\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ and $X_P=\left\|\mathcal{S}_{h,P}(k^2)\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ evaluated at 100 sample points uniformly distributed in $K=[7,14]$

Proof We have $|K_\alpha| \leq \alpha n$, with $n \leq N$ the number of poles of $\mathcal{S}$ in K. From Theorem 6.1 it holds

$$\left|\phi_X(t) - \phi_{X_P}(t)\right| \leq \inf_{\alpha>0}\left(C_1\alpha + C_2(t)\frac{1}{\alpha^3}\left(\frac{\rho}{R}\right)^{M+1}\right),$$

with $C_1 = 2n \sup_{x\in K} \mathscr{F}_{k^2}(x)$ and $C_2(t) = |t|\, LC\, |K|\, \mathscr{F}_{k^2}(x)$. By optimizing the expression in α we obtain

$$\left|\phi_X(t) - \phi_{X_P}(t)\right| \leq C_t\left(\frac{\rho}{R}\right)^{\frac{M+1}{4}}$$

with $C_t = C_1^{3/4} C_2(t)^{1/4}(3^{1/4} + 3^{-3/4})$. □

A straightforward application of Lévy's Continuity Theorem [19, Chapter 19] leads now to the convergence in distribution.

Corollary 6.3 *Under the same assumptions as in Theorem 6.1, X_P converges in distribution to X, as $M \to +\infty$.*

6.1 Numerical example

Let us consider a square domain $D = [0, \pi]^2$, with homogeneous Dirichlet boundary conditions ($\partial D = \Gamma_D$). Let $K = [7, 14]$ be the interval of interest (which contains three eigenvalues of the Dirichlet-Laplace operator: 8, 10, 13), and let the squared wavenumber be modeled as a random variable uniformly distributed on K, i.e., $k^2 \sim \mathcal{U}(K)$. Given the functional $\mathcal{L} = \|\cdot\|_{V,\sqrt{\mathrm{Re}(z_0)}}$, where $z_0 = 10 + 0.5i$, we consider the random variables $X = \left\|\mathcal{S}_h(k^2)\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}$ and $X_P = \left\|\mathcal{S}_{h,P}(k^2)\right\|_{V,\sqrt{\mathrm{Re}(z_0)}}$. We define as $\mathcal{S}_h$ the $\mathbb{P}^3$ FE approximation of $\mathcal{S}$; then $\mathcal{S}_{h,P}$ is the (standard) LS-Padé approximant of $\mathcal{S}_h$, centered at z_0 and with polynomial degrees (M, N). In Fig. 7, we display the random variables X and X_P evaluated at 100 sample points uniformly distributed in K. When the degree of the LS-Padé denominator is $N = 3$, all the poles are correctly identified by the LS-Padé approximant, provided that M is larger than 4. In Fig. 8 we plot the characteristic function of the random variable X_P, $\phi_{X_P}(t)$,

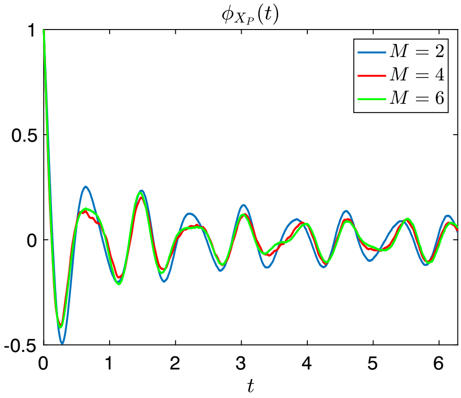

Fig. 8 Characteristic function $\phi_{X_P}(t)$, with $N = 3$ and $M = 2, 4, 6$

where the degrees of the Padé denominator and denominator are $N = 1, 2, 3$, and $M = 2, 4, 6$, respectively. The expected value has been computed by the Monte Carlo method, using 10^5 samples.

7 Conclusions

The present paper concerns two model order reduction methods based on the single-point LS-Padé approximation techniques introduced in [3] and [4], respectively. We have described an algorithm to compute LS-Padé approximants of Hilbert space-valued maps, and we have explored the applicability and potential of both methods via 2D numerical experiments in different frameworks involving a parametric Helmholtz equation. The time-harmonic wave equation with random wavenumber has been analyzed, as well.

We are currently investigating the extension of the proposed methodology and of its convergence analysis to the case of *multi-point* LS-Padé expansions, where evaluations of the frequency response map $\mathcal{S}$ and of its derivatives at *multiple* frequencies are used. We believe that this technique will outperform the single-point one, when a large number of singularities of $\mathcal{S}$ needs to be identified. Moreover, we expect such extension to overcome the conditioning issues of computing Taylor coefficients of the frequency response map.

Funding information Open access funding provided by Austrian Science Fund (FWF). F. Bonizzoni acknowledges partial support from the Austrian Science Fund (FWF) through the project F 65, and has been supported by the FWF Firnberg-Program, grant T998.

I. Perugia has been funded by the Vienna Science and Technology Fund (WWTF) through the project MA14-006, and by the Austrian Science Fund (FWF) through the projects P 29197-N32 and F 65.

D. Pradovera has been funded by the Swiss National Science Foundation (SNF) through project 182236.

References

1. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, vol. 140. Academic Press (2003)
2. Babuška, I.M., Sauter, S.A.: Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? SIAM J. Numer. Anal. **34**(6), 2392–2423 (1997)
3. Bonizzoni, F., Nobile, F., Perugia, I.: Convergence analysis of Padé, approximations for Helmholtz frequency response problems. ESAIM Math. Model. Numer Anal. **52**(4), 1261–1284 (2018)
4. Bonizzoni, F., Nobile, F., Perugia, I., Pradovera, D.: Fast Least-Squares Padé approximation of problems with normal operators and meromorphic structure. Math. Comput. **89**, 1229–1257 (2020)

5. Bonizzoni, F., Pradovera, D.: Distributed sampling for rational approximation of the acoustic scattering of an airfoil. Proc. Appl. Math. Mech. https://doi.org/10.1002/pamm.201900422 (2019)
6. Chen, Y., Hesthaven, J.S., Maday, Y., Rodríguez, J.: Certified reduced basis methods and output bounds for the harmonic Maxwell's equations. SIAM J. Sci Comput. **32**(2), 970–996 (2010)
7. Daniel, L., Siong, O.C., Chay, L.S., Lee, K.H., White, J.: A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models. IEEE Trans Comput.-Aided Design Integr. Circuits Syst. **23**(5), 678–693 (2004)
8. Ezvan, O., Batou, A., Soize, C., Gagliardini, L.: Multilevel model reduction for uncertainty quantification in computational structural dynamics. Comput. Mech. **59**(2), 219–246 (2017)
9. Feng, L., Benner, P.: A robust algorithm for parametric model order reduction based on implicit moment matching. Reduced order methods for modeling and computational reduction. In: Quarteroni, A., Rozza, G. (eds.) MS&A Series. Springer, Cham (2014)
10. Guillaume, P., Huard, A., Robin, V.: Generalized multivariate Padé approximants. J. Approx. Theory **95**(2), 203–214 (1998)
11. Hain, S., Ohlberger, M., Radic, M., Urban, K.: A hierarchical a-posteriori error estimatorfor the reduced basis method. Adv. Comput. Math. **45**, 2191–2214 (2019)
12. Hetmaniuk, U., Tezaur, R., Farhat, C.: Review and assessment of interpolatory model order reduction methods for frequency response structural dynamics and acoustics problems. Int. J. Numer. Methods Eng. **90**(13), 1636–1662 (2012)
13. Hetmaniuk, U., Tezaur, R., Farhat, C.: An adaptive scheme for a class of interpolatory model reduction methods for frequency response problems. Int. J. Numer. Methods Eng. **93**(10), 1109–1124 (2013)
14. Hiptmair, R., Moiola, A., Perugia, I.: Trefftz discontinuous Galerkin methods for acoustic scattering on locally refined meshes. Appl. Numer. Math. **79**, 79–91 (2014)
15. Hiptmair, R., Scarabosio, L., Schillings, C., Schwab, C.: Large deformation shape uncertainty quantification in acoustic scattering. Adv. Comput. Math. **44**, 1475–1518 (2018)
16. Huard, A., Robin, V.: Continuity of approximation by least-squares multivariate Padé approximants. J. Comput. Appl. Math. **115**(1–2), 255–268 (2000)
17. Huynh, D.B.P., Rozza, G., Sen, S., Patera, A.T.: A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. Comptes Rendus Mathematique **345**(8), 473–478 (2007)
18. Huynh, D.B.P., Knezevic, D.J., Patera, A.T.: A static condensation reduced basis element approximation: Application to three-dimensional acoustic muffler analysis. Int. J. Comput. Methods **11**(03), 1343010 (2014)
19. Jacod, J., Protter, P.: Probability Essentials. Springer, Berlin (2004)
20. Jacquelin, E., Dessombz, O., Sinou, J.-J., Adhikari, S., Friswell, M.I.: Polynomial chaos-based extended Padé expansion in structural dynamics. Int. J. Numer. Methods Eng. **111**(12), 1170–1191 (2017)
21. Jacquelin, E., Dessombz, O., Sinou, J.-J., Adhikari, S., Friswell, M.I.: Steady-state response of a random dynamical system described with Padápproximants and random eigenmodes. Procedia Eng. **199**, 1104–1109 (2017). X International Conference on Structural Dynamics, EURODYN 2017
22. Kapita, S., Monk, P., Warburton, T.: Residual-based adaptivity and PWDG methods for the Helmholtz equation. SIAM J. Sci. Comput. **37**(3), A1525–A1553 (2015)
23. Lassila, T., Manzoni, A., Rozza, G.: On the approximation of stability factors for general parametrized partial differential equations with a two-level affine decomposition. ESAIM: Math. Modell. Numer. Anal. **46**(6), 1555–1576 (2012)
24. Lenoir, M., Vullierme-Ledard, M., Hazard, C.: Variational formulations for the determination of resonant states in scattering problems. SIAM J. Math. Anal. **23**(3), 579–608 (1992)
25. Manetti, M.: Topology. Springer International Publishing (2015)
26. McLean, W.C.H.: Strongly Elliptic Systems and Boundary Integral Equations. Cambridge University Press (2000)
27. Modesto, D., Zlotnik, S., Huerta, A.: Proper generalized decomposition for parameterized Helmholtz problems in heterogeneous and unbounded domains: Application to harbor agitation. Comput. Methods Appl. Mech. Eng. **295**, 127–149 (2015)
28. Ohayon, R., Soize, C.: Computational vibroacoustics in low- and medium- frequency bands: damping, ROM, and UQ modeling. Appl. Sci.-Basel, **7**(6) (2017)
29. Pradovera, D.: Interpolatory rational model order reduction of parametric problems lacking uniform inf-sup stability, ArXiv e-prints (2019)

30. Schwab, C., Gittelson, C.J.: Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. Acta Numerica **20**, 291–467 (2011)
31. Sen, S.: Reduced basis approximation and a posteriori error estimation for non-coercive elliptic problems: Applications to acoustics. PhD thesis, Massachusetts Institute of Technology (2007)
32. Sen, S., Veroy, K., Huynh, D.B.P., Deparis, S., Nguyen, N.C., Patera, A.T.: "Natural norm" a posteriori error estimators for reduced basis approximations. J. Comput. Phys. **217**(1), 37–62 (2006)
33. Steinberg, S.: Meromorphic families of compact operators. Arch. Ration. Mech. Anal. **31**(5), 372–379 (1968)
34. Tonn, T., Urban, K., Volkwein, S.: Comparison of the reduced-basis and POD a posteriori error estimators for an elliptic linear-quadratic optimal control problem. Math. Comput. Model. Dyn. Syst. **17**(4), 355–369 (2011)
35. Veroy, K., Prud'Homme, C., Rovas, D.V., Patera, A.T.: A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations (2003)

Affiliations

Francesca Bonizzoni[1] · **Fabio Nobile**[2] · **Ilaria Perugia**[1] · **Davide Pradovera**[2]

Fabio Nobile
fabio.nobile@epfl.ch

Ilaria Perugia
ilaria.perugia@univie.ac.at

Davide Pradovera
davide.pradovera@epfl.ch

[1] Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090, Wien, Austria

[2] CSQI – MATH, Ecole Polytechnique Fédérale de Lausanne, Station 8, CH-1015,
Lausanne, Switzerland