




High order approximations of the operator Lyapunov equation have low rank

Luka Grubišić¹ · Harri Hakula² 

Received: 24 February 2021 / Accepted: 16 March 2022 / Published online: 7 April 2022
© The Author(s) 2022

Abstract

We present a low-rank greedily adapted hp -finite element algorithm for computing an approximation to the solution of the Lyapunov operator equation. We show that there is a hidden regularity in eigenfunctions of the solution of the Lyapunov equation which can be utilized to justify the use of high order finite element spaces. Our numerical experiments indicate that we achieve eight figures of accuracy for computing the trace of the solution of the Lyapunov equation posed in a dumbbell-domain using a finite element space of dimension of only 10^4 degrees of freedom. Even more surprising is the observation that hp -refinement has an effect of reducing the rank of the approximation of the solution.

Keywords Low-rank approximation · Exponential decay · Lyapunov equation · hp -finite element methods

Mathematics Subject Classification 65N30 · 65N50 · 65N25 · 93C05

Communicated by Daniel Kressner.

This research was funded by the Hrvatska Zaklada za Znanost (Croatian Science Foundation) under the Grant IP-2019-04-6268 - Randomized low-rank algorithms and applications to parameter dependent problems.

✉ Harri Hakula
Harri.Hakula@aalto.fi
Luka Grubišić
luka@math.hr

¹ Department of Mathematics, Faculty of Science, University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia

² Department of Mathematics and Systems Analysis, Aalto University, P.O. Box 11100, 00076 Aalto, Finland

1 Introduction

When dealing with non-local operators, low-rank approximation methods are turning out to be a method of choice both for theoretical analysis as well as a foundation for constructing high performance numerical algorithms. We will concentrate on non-local operators defined as solutions to the continuous time Lyapunov operator equation with a rank one operator as the right-hand side coefficient. This equation can be formally written as

$$AX + XA = bb' . \quad (1.1)$$

Here we assume that A is an unbounded (differential) positive definite self-adjoint operator in a Hilbert space and b' is a functional on $\mathcal{X} = \text{dom}(A^{1/2})$, the domain of the square root of A . A typical example of the operator A would be a reaction-diffusion operator in $\mathcal{H} = L^2(\Omega)$. The formal expression (1.1) can be justified in the framework of Gelfand triplets $\mathcal{X} \subset \mathcal{H} \subset \mathcal{W}$ where $\mathcal{W} = \mathcal{X}'$ is the topological dual of \mathcal{X} and $\|\cdot\|_{\bullet}, \bullet = \mathcal{X}, \mathcal{H}, \mathcal{W}$ denote the corresponding norms [20]. We use the notation b^* to denote the functional $b^* : \mathcal{H} \rightarrow \mathbb{C}$ and b' to denote the functional $b' : \mathcal{X} \rightarrow \mathbb{C}$. Note that \mathcal{H} and \mathcal{X} do not have the same scalar product and so the reason for the difference in notation. The notation $b' \in \mathcal{X}'$ implies that we have not identified \mathcal{X} and \mathcal{X}' , whereas the notation b^* assumes that \mathcal{H} was identified with its dual.

Under the assumption of the positive definiteness of A , the equation (1.1) has a unique positive solution $X \in \mathcal{L}(\mathcal{H})$. Furthermore, the fact that the operator bb' is of rank one implies that X is contained in every Schatten ideal \mathcal{S}_p , $p \in \mathbb{N}$, see [26]. In particular, this implies that the trace $\text{tr}(X)$ is finite. This will be the quantity of interest which we will monitor in the presented numerical experiments. The physical relevance of $\text{tr}(X)$ stems from the fact that it represents the total output energy of the formal Cauchy system

$$\begin{aligned} \dot{x} &= -Ax, \\ x(0) &= b. \end{aligned}$$

To see this, notice that the solution of this initial value problem is then given by the formula $x(t) = \exp(-tA)b$ and then

$$\begin{aligned} E^2 &= \int_0^\infty \|x(t)\|^2 dt = \int_0^\infty (x(t), x(t)) dt \\ &= \int_0^\infty \text{tr}(x(t)x(t)^*) dt \\ &= \text{tr} \left(\int_0^\infty (\exp(-tA)b)(\exp(-tA)b)^* dt \right) = \text{tr}(X) \end{aligned}$$

follows. A more common application of the Lyapunov equation is in the study of the model order reduction for the linear control systems $\dot{x}(t) = -Ax(t) + bu(t)$, by

means of balanced truncation [1, 22]. We use precisely this context to justify the term *loading* for the function b .

By $\|X\| = \sqrt{\text{spr}(X^*X)}$, where spr is the spectral radius of an operator, we denote the operator norm of X . The main goal of this paper is to construct a low-rank approximation $Z_r = \sum_{j=1}^r z_j z_j^*$ of X using a computationally efficient approximation method. For a given tolerance $\tau > 0$ such an algorithm constructs Z_r such that $\|Z_r - X\| \leq \tau$ and r is a small integer. In this paper we will be looking for r such that for the given $\tau > 0$ we have $\|Z_r - X\| \leq \tau$ and $|\text{tr}(Z_r) - \text{tr}(X)| \leq \tau$.

Let $\mu_1(H) \geq \mu_2(H) \geq \dots$ denote the eigenvalues, counting by multiplicity, of some compact self-adjoint operator H . For the solution X of (1.1) assume $\mu_r(X) > \mu_{r+1}(X)$ for some $r \in \mathbb{N}$. The Weyl's theorem [6, 36] for compact operators now implies

$$|\mu_i(X) - \mu_i(Z_r)| \leq \|X - Z_r\|, \quad i = 1, \dots, r.$$

Let now $P_{\mathcal{E}_r}$ be the orthogonal projection onto the eigensubspace belonging to the r largest eigenvalues of X and let $P_{\mathcal{Z}}$ be the orthogonal projection onto the subspace $\mathcal{Z} = \text{span}\{z_1, \dots, z_r\}$. The Davis-Kahan $\sin \Theta$ theorem [8] yields the estimate

$$\|P_{\mathcal{E}_r} - P_{\mathcal{Z}}\| \leq \frac{1}{\mu_{r+1}(X) - \rho_r(\mathcal{Z}, X)} \|X - Z_r\|,$$

where $\rho_r(\mathcal{Z}, X)$ is the Rayleigh Quotient, formally defined below. In the view of this the space $\mathcal{Z} = \text{span}\{z_1, \dots, z_r\}$ can be seen as an approximation to the eigenspace belonging to the r largest eigenvalues of X . In this paper we will interpret the low-rank approximation problem for the Lyapunov equation as the subspace approximation problem and use this context to assess the quality of the constructed vectors $z_i, i = 1, \dots, r$. The algorithm which we will present will be formulated for the divergence form operators posed in a polygonal planar domain. The vectors z_j will be constructed as elements of a low-dimensional space of continuous piecewise polynomial functions [23]. We will present our results, as far as possible, in their abstract form and specialize to particular divergence form operators when presenting numerical examples. The main theme of the paper is an interplay of the low-rank approximation methods based on spectral calculus [11] and the utilization of the regularity of the eigenvectors of X in the construction of higher order adaptive approximation methods [2, 23].

1.1 The novelty of the paper

We construct an a posteriori error estimator for an approximation to the solution of the Lyapunov equation for the divergence form operator posed in a polygonal domain in \mathbb{R}^2 . The estimator is based on the auxiliary subspace technique from [3]. We first show that the estimator can be approximated by refining the auxiliary subspace. This yields our first practical algorithm. Then, we present an analysis of the estimator based on the perturbation theory for the Rayleigh Quotients from [37] combined with the use of an eigenvalue saturation assumption from [25]. This yields an alternative computable error indicator. We justify the use of the eigenvalue saturation assumption

based on a perturbation argument, showing that the solution X is a small perturbation of an operator with A -analytic eigenvectors. We test the estimator on a sequence of graded meshes and observe that it correctly indicates the portions of the domain where the mesh should be refined but also does not indicate those which are not relevant for a particular solution. We also show that we observe exponential square root convergence—even in the case of non convex polygonal domains—as we would when approximating piecewise analytic functions in a domain of \mathbb{R}^2 .

1.2 The outline of the paper

The rest of the paper is structured as follows: We will first in Sect. 2 introduce the basic notation and present the motivating example. In Sect. 3 we review the background results on the approximation methods for the Lyapunov equation and present the result on approximate A -analyticity of the eigenvectors of the solution operator. We then introduce the basic conventions from the hp -finite element approximation theory in Sect. 4. In Sect. 5 we present an a posteriori error indicator and a framework for an analysis of its reliability. We then present a greedy hp -refinement strategy based on the error balancing approach. We continue with further Numerical experiments in Sect. 6. A sketch of the theoretical result related to the case in which the forcing $b \in \mathcal{X}'$ is such that $\|A^{-\alpha}b\| < \infty$ for some $0 \leq \alpha < 1/2$ is outlined in the “Appendix A”.

2 Basic definitions and a motivating example

In this section we will present a motivating numerical example and introduce the basic notation and results which are needed to interpret the result. This will in particular include the introduction of the regularity classes associated with an operator A and the review of the perturbation analysis of the Rayleigh Quotient.

Let us recall the following definitions of the regularity classes defined by the positive definite operator A , see [24, 30]. According to [30, Section 7.4] a vector is called the infinite vector of the operator A if it is the element of the set $\mathcal{A}_\infty = \bigcap_{n \in \mathbb{N}} \text{dom}(A^n)$. A vector $x \in \mathcal{A}_\infty$ is called an A -bounded vector if there exists a constant $B_x > 0$ such that

$$\|A^n x\| \leq B_x^n, \quad n \in \mathbb{N}_0,$$

and we write $x \in \mathcal{A}_b$. A vector $x \in \mathcal{A}_\infty$ is called an A -analytic vector if there exists a constant $C_x > 0$ such that

$$\|A^n x\| \leq C_x^n n!, \quad n \in \mathbb{N}_0,$$

and we write $x \in \mathcal{A}_a$. It holds that $\mathcal{A}_b \subset \mathcal{A}_a \subset \mathcal{A}_\infty$ and so eigenvectors of an operator A are A -bounded vectors. The vector $v = \exp(-tA)b$, for $b \in \mathcal{H}$ and $t > 0$, is an example of an A -analytic vector, see [30, Example 7.5].

Let now $\Omega \subset \mathbb{R}^2$ be a bounded domain and let A be a divergence type positive definite (elliptic and self-adjoint) operator in $\mathcal{H} = L^2(\Omega)$. Then we have a more

detailed description of the regularity spaces. In the case in which A is a divergence form operator with analytic coefficients and b is also analytic and Ω has at least a C^2 regular boundary, an A -analytic function is an (real) analytic function in the classical sense, and in particular the function $v = A^{-1}b$ is analytic in the interior of Ω , [24]. In the case in which A is a divergence type operator with piecewise analytic coefficients, b is piecewise analytic, and the boundary of Ω is also piecewise analytic the solution $v = A^{-1}b$ is still infinitely differentiable. Further, the function v can be represented as a sum of an analytic function and a function whose singularities are concentrated at the corners. So even though we cannot control the growth of the derivatives in the classical sense, the solution is contained in all weighted Sobolev spaces, where the weighting function is the distance of a point to the corners of the domain [23].

The two prototype domains with the associated low-rank approximations to the solution of the associated Lyapunov equation are presented in the following examples. For the coefficients of the Lyapunov equation $AX + XA = bb^*$ we choose $b(x_1, x_2)$ to be the bell curve concentrated at $(1/2, 1/2)$:

$$b(x_1, x_2) = \exp\left(-\delta\left(x_1 - \frac{1}{2}\right)^2 - \delta\left(x_2 - \frac{1}{2}\right)^2\right), \quad \delta > 0$$

and we take A to be the Laplace operator with the Dirichlet boundary conditions in the corresponding domain Ω .

Example 2.1 (Ellipse) Let us choose the computational domain Ω as an ellipse. The operator A in the Lyapunov equation (1.1) is taken to be the Laplace operator with the zero Dirichlet boundary conditions and we set $\delta = 5$ for the load vector b . This is the regularity setting as in the classical paper by Nelson [24] and it ensures that A -analytic vectors are real analytic functions. The solution X is remarkably of (numerical) rank = 2. See Fig. 1 for illustrations of the loading, the first column of X , and two eigenmodes of X .

Example 2.2 (Dumbbell A) Let us consider a classical Laplace dumbbell problem with computational domain $\Omega = ([0, 2.4] \times [0, 1]) \setminus (([1, 1.4] \times [0, 0.3]) \cup ([1, 1.4] \times [0.7, 1]))$. We choose the operator A in the Lyapunov equation (1.1) to be the Laplace operator with the zero Dirichlet boundary conditions and set $\delta = 50$ for the load vector b . The solution X is again of (numerical) rank = 2 and we see that the vectors z_1 and z_2 are highly regular away from the corners of the domain. See Fig. 2 for illustrations of the loading, the first column of X , and two eigenmodes of X .

We will analyze the eigenvalues of a positive compact operator H using variational techniques. Let $x \in \mathcal{H} \setminus \{0\}$ be given, then

$$\rho(x, H) = \frac{(x, Hx)}{(x, x)}$$

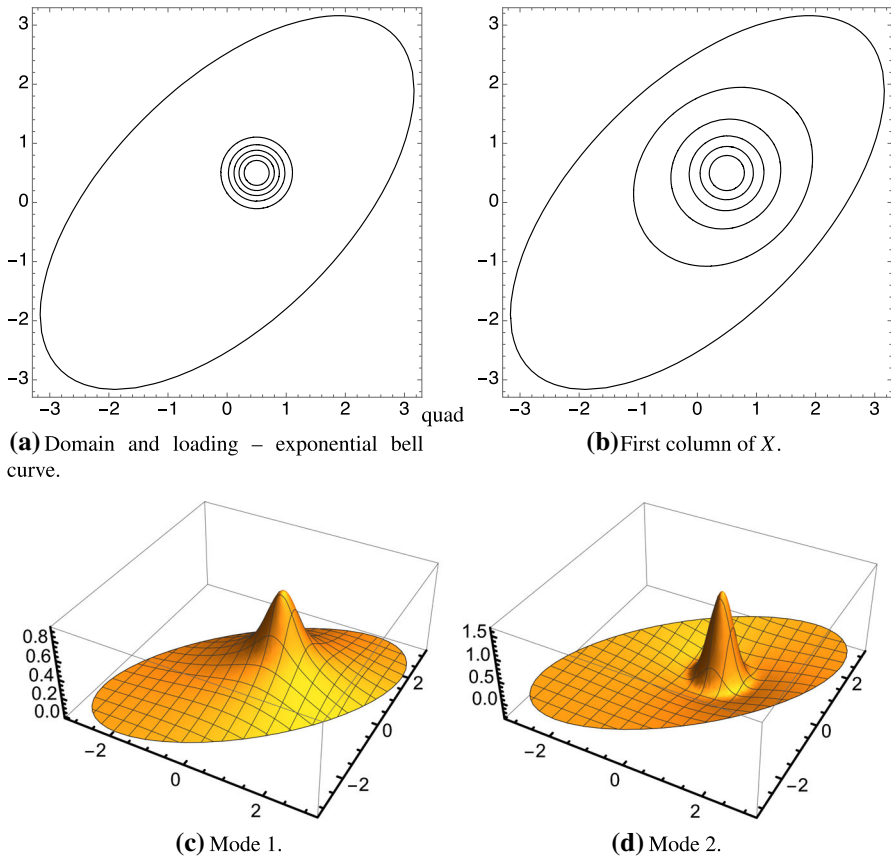


Fig. 1 A representative example: An elliptic domain with an exponential bell curve concentrated at $(1/2, 1/2)$. The columns of X can be plotted and the first one is illustrated. In this case the operator has numerical rank = 2, and the corresponding modes are shown. The dimension of the finite element space is 10^4

is the Rayleigh Quotient of the vector x for the operator H . In the case in which we are given a r -dimensional subspace $\mathcal{L} \subset \mathcal{H}$ we define the Rayleigh Quotients

$$\rho_i(\mathcal{L}, H) = \max_{\substack{\mathcal{F} \subset \mathcal{L} \\ \dim \mathcal{F} = i}} \min_{x \in \mathcal{F} \setminus \{0\}} \frac{(x, Hx)}{(x, x)}.$$

Obviously, for the r dimensional subspace \mathcal{L} we have $\mu_i(H) \geq \rho_i(\mathcal{L}, H)$, $i = 1, \dots, r$. We will now review some basic results on the Rayleigh Quotient analysis from [15, 17, 37]. Let $H : \mathcal{H} \rightarrow \mathcal{H}$ be a positive compact operator and let u , $\|u\| = 1$ be an eigenvector so that $Hu = \mu u$, $\mu \in \text{Spec}(H)$. For a given non-zero vector $\psi \in \mathcal{H}$ we have the estimate

$$|\mu - \rho(\psi, H)| \leq \mu_1(H) \sin^2(\angle\{\psi, u\}). \tag{2.1}$$

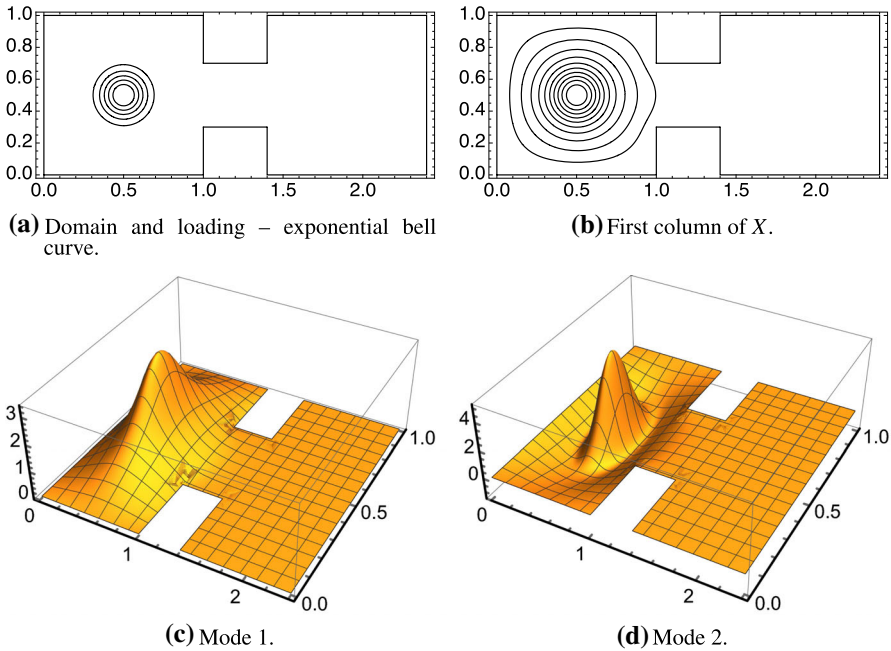


Fig. 2 A representative example: A dumbbell domain with an exponential bell curve concentrated at $(1/2, 1/2)$. The columns of X can be plotted and the first one is illustrated. In this case the operator has numerical rank = 2, and the corresponding modes are shown. The dimension of the finite element space is 10^4

This estimate is obviously very accurate for Rayleigh Quotients $\rho(\psi, H)$ which are close to $\mu_1(H)$. This is precisely the setting in which we expect to find ourselves. This bound has a subspace extension which can be used to treat eigenvalues with higher multiplicities or clusters of eigenvalues, [15, 18]. We measure the distance between finite dimensional subspaces \mathcal{X} and \mathcal{Y} using the concept of the principal angle. Let $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ be the orthogonal projections onto \mathcal{X} and \mathcal{Y} respectively. The vector of the sines squared $\sin_p^2 \Theta(\mathcal{X}, \mathcal{Y})$ of the principal angles between \mathcal{X} and \mathcal{Y} is defined using eigenvalues of the positive self-adjoint operator $S = I - P_{\mathcal{X}} P_{\mathcal{Y}} P_{\mathcal{X}}$ as $\sin_p^2 \Theta_i(\mathcal{X}, \mathcal{Y}) = \mu_i(S)$. Since the operator S is a self-adjoint operator, we define other trigonometric functions of the principal angles using spectral calculus. We define the sine of the maximal principal angle as $\sin^2 \Theta(\mathcal{X}, \mathcal{Y}) = \|I - P_{\mathcal{X}} P_{\mathcal{Y}} P_{\mathcal{X}}\|$.

Under the assumption that $\mu_r(H) > \mu_{r+1}(H)$, the crudest estimate from [18, Theorem 2.2] reads

$$\sum_{i=1}^r \mu_i(H) - \sum_{i=1}^r \rho_i(\mathcal{L}, H) \leq \mu_1(H) \sum_{i=1}^r \sin_p^2 \Theta_i(\mathcal{E}_r, \mathcal{L}) \leq \mu_1(H) r \sin^2 \Theta(\mathcal{E}_r, \mathcal{L}).$$

Here \mathcal{E}_r denotes the eigensubspace associated to the eigenvalues $\mu_i(H), i = 1, \dots, r$. This estimate implies, using [16, Lemma 5.5] and [35],

$$\sum_{i=1}^r \mu_i(H) - \sum_{i=1}^r \rho_i(\mathcal{L}, H) \leq \mu_1(H) \sum_{i=1}^r \|v_i - z_i\|^2, \tag{2.2}$$

where $Xv_i = \mu_i(X)v_i$, and v_i are an orthonormal set of vectors.

3 The Lyapunov equation

In this section we will review the low-rank approximation estimates for the solution X of the Lyapunov equation (1.1). Also, we will review basic results on the projection based approximation methods for the Lyapunov equation.

Given a Gelfand triple $\mathcal{X} \subset \mathcal{H} \subset \mathcal{W}$ of Hilbert spaces, where $\mathcal{W} = \mathcal{X}'$ is the dual space to \mathcal{X} , we consider an unbounded operator A such that its range is in \mathcal{W} and its domain of definition is given by $\mathcal{X} = \text{dom}_{\mathcal{W}}(A) = \{x \in \mathcal{H} : \|Ax\|_{\mathcal{W}} < \infty\}$. We let $A' : \mathcal{W} \rightarrow \mathcal{X} \subset \mathcal{W}$ denote the dual operator to A in the duality pairing $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{W} \times \mathcal{X}}$. Moreover, we consider a (not necessarily bounded) linear operator $B : \mathcal{U} \rightarrow \mathcal{W}$ for a Hilbert space \mathcal{U} with inner product $(\cdot, \cdot)_{\mathcal{U}}$.

The operators A, B give rise to the *Lyapunov operator equation* in a linear operator X :

$$AX + XA' = -BB', \tag{3.1}$$

which formally stands for the variational formulation

$$\langle AXz_1, z_2 \rangle_{\mathcal{W} \times \mathcal{X}} + \langle XA'z_1, z_2 \rangle_{\mathcal{W} \times \mathcal{X}} = \mathfrak{b}(z_1, z_2), \quad z_1, z_2 \in \mathcal{X}, \tag{3.2}$$

with the sesquilinear form $\mathfrak{b}(z_1, z_2) = -(B'z_1, B'z_2)_{\mathcal{U}}$. We refer to, e.g., [7, 14, 29] for a more detailed discussion of this equation.

We now consider the situation when A is positive definite self-adjoint on \mathcal{H} , and in addition has a compact resolvent. We choose $\mathcal{W} = \mathcal{H}_{-1/2}$, which is equipped with the scalar product $(\cdot, A^{-1}\cdot) = (A^{-1/2}\cdot, A^{-1/2}\cdot)$, and $\mathcal{X} = \mathcal{H}_{1/2} = \text{dom}_{\mathcal{H}}(A^{1/2})$.

Additionally, we assume that the product $A^{-1/2}B$ is bounded. This is equivalent to the assumption that

$$b(\psi, \phi) = -\mathfrak{b}(A^{-1/2}\psi, A^{-1/2}\phi)$$

is everywhere defined and bounded on \mathcal{H} . The substitutions $\psi = A^{1/2}z_1$ and $\phi = A^{1/2}z_2$ then allow us to turn (3.2) into the equivalent equation

$$(A^{1/2}\psi, XA^{-1/2}\phi) + (XA^{-1/2}\psi, A^{1/2}\phi) = b(\psi, \phi), \quad \psi, \phi \in \mathcal{X}. \tag{3.3}$$

3.1 Approximate A -analyticity of the eigenvectors

A result of [11] showed that in the case in which $b \in \mathcal{H}$ and A has the compact resolvent the following estimate holds

$$\|X - \sum_{p=-k}^k \omega_p \exp(-t_p A)b(\exp(-t_p A)b, \cdot)\| \leq \frac{C_{St} \|b\|^2}{2\lambda_1} \exp(-\pi \sqrt{k}). \tag{3.4}$$

The constant C_{St} can be bounded to be less than 3, and given $k \in \mathbb{N}$ the weights ω_p and the nodes t_p are given by an explicit formula (see [11])

$$\begin{aligned} t_p &= \log(\exp(p\pi/\sqrt{k}) + \sqrt{1 + \exp(2p\pi/\sqrt{k})})/(2\lambda_1), \\ \omega_p &= \pi/(2\sqrt{k(1 + \exp(p\pi/\sqrt{k})\lambda_1)}), \end{aligned} \tag{3.5}$$

for $p = -k, \dots, k$ and λ_1 denoting the smallest eigenvalue of A . This construction can still be performed under the weaker assumption that only $\|A^{-\alpha}b\| < \infty$ for some $\alpha, 0 \leq \alpha < 1/2$. Under this assumption one needs to modify ω_p and t_p to reflect the changed asymptotic (as $t \rightarrow 0$) behavior of $\exp(-tA)b$ and the fact that we are using a sinc quadrature formula for the sector rather than for the strip as was used in [11], see [27]. For further details of this construction and for ramifications in the context of this paper, see ‘‘Appendix A’’. The assumption $\|A^{-\alpha}b\| < \infty$ also covers the standard boundary control setting for the Lyapunov equation, see [20, 27].

The ramifications of this result are twofold. First it indicates that X can be represented by the sum of rank one operators and that the error decays exponentially in the square root of the number of terms in the sum. But equally important is the second consequence, the range of the operator

$$X_k = \sum_{p=-k}^k \omega_p (\exp(-t_p A)b)(\exp(-t_p A)b, \cdot) \tag{3.6}$$

is finite dimensional and it is spanned by the A -analytic vectors $v_p = \exp(-t_p A)b$.

Proposition 3.1 *Assume A is self-adjoint and positive definite and let $\tau > 0$ be given. By X denote the unique self-adjoint and positive solution of (3.2). Then there exist numbers r and k such that*

$$|\text{tr}(X) - \sum_{i=1}^r \mu_i(X_k)| \leq \tau \text{tr}(X)$$

and eigenvectors $\hat{u}_i \neq 0, X_k \hat{u}_i = \mu_i(X_k) \hat{u}_i$ are A -analytic.

Proof Since X is trace class, there exists a number r such that $|\operatorname{tr}(X) - \sum_{i=1}^r \mu_i(X)| \leq (\tau \operatorname{tr}(X))/2$. Further, let $k \in \mathbb{N}$ be the smallest k such that

$$\|X - X_k\| \leq \frac{C_{Sr} \|b\|^2}{2\lambda_1} \exp(-\pi\sqrt{k}) \leq \frac{\tau}{2r} \operatorname{tr}(X)$$

then

$$\begin{aligned} |\operatorname{tr}(X) - \sum_{i=1}^r \mu_i(X_k)| &= |\operatorname{tr}(X) - \sum_{i=1}^r \mu_i(X) + \sum_{i=1}^r \mu_i(X) - \sum_{i=1}^r \mu_i(X_k)| \\ &\leq (\tau \operatorname{tr}(X))/2 + \sum_{i=1}^r |\mu_i(X) - \mu_i(X_k)| \\ &\leq (\tau \operatorname{tr}(X))/2 + (\tau \operatorname{tr}(X))/2. \end{aligned}$$

The last inequality follows by the use of the Weyl’s theorem which implies

$$|\mu_i(X) - \mu_i(X_k)| \leq \|X - X_k\| \leq \tau/(2r) \operatorname{tr}(X),$$

from which the final estimate follows by summing over i . □

We can interpret (3.4) as a substitute regularity result. If we consider the actual operator X as a perturbation of the operator X_k then we can say that with exponentially (in \sqrt{k}) decaying tolerance the eigenvectors of X are close to \mathcal{A}_a eigenvectors of the operator X_k . This in turn implies that a high order method, such as an hp -adaptive finite element method for operators posed in $\mathcal{H} = L^2(\Omega)$ might be able to exploit this regularity to construct a high performance solver which will mix low-rank numerical linear algebra with full finite element piecewise polynomial adaptivity. Also, this indicates that we might view the low-rank approximation task as a task of computing the approximation to the r dominant eigenvalues of the operator X .

3.2 Galerkin approximation of the Lyapunov equation

We will first present abstract approximation results. Let A be positive definite and self-adjoint operator with a compact resolvent. Let us further assume that we have constructed a sequence of finite dimensional subspaces $\mathcal{V}_s \subset \mathcal{X} = \operatorname{dom}(A^{1/2})$, $s \in \mathbb{N}$ such that $\mathcal{V}_{s_1} \subset \mathcal{V}_{s_2}$ for $s_1 < s_2$ and the orthogonal projections P_s onto \mathcal{V}_s converge strongly to the identity operator as $s \rightarrow \infty$.

According to [20, Section 4.1.2 and Section 5], the Galerkin projection $A_s : \mathcal{V}_s \rightarrow \mathcal{V}_s$ is given by the formula

$$A_s = (A^{1/2} P_s)^*(A^{1/2} P_s).$$

It holds that

$$\|A_s^{-1} P_s v - A^{-1} v\|_{\operatorname{dom}(A^{1/2})} \leq 2\|(I - P_s)A^{-1} v\|_{\operatorname{dom}(A^{1/2})}, \tag{3.7}$$

and furthermore $\|A_s^{-1}P_s - A^{-1}\| \rightarrow 0$. For more on the convergence of discrete operator approximations see [20, Section 4] and classical references [6, 28]. With this notation we define the operator $X_s : \mathcal{V}_s \rightarrow \mathcal{V}_s$ —in a generic situation when instead we have some finite element subspace \mathcal{V} we tacitly write $X_{\mathcal{V}}$ —as the solution of the finite dimensional operator equation

$$A_s X_s + X_s A_s = -(P_s b)(P_s b)^* . \tag{3.8}$$

Under additional assumptions on the uniformity of the sequence P_s , a general result from [20, Theorem 4.1.4.1] states that $\|X_s P_s - X\| \rightarrow 0$. The precise formulation of the convergence result is quite technical. Intuitively it could be condensed to checking that the Galerkin projections A_s are uniformly coercive in s and that the orthogonal projections onto \mathcal{V}_s converge strongly to identity in a monotone way (in the Loewner order) and at a guaranteed rate. An example of spaces for which these assumption hold are spaces associated with a hierarchical finite element scheme defined on a quasi-uniform grid. For further details see [20, Section 5.2] where the convergence rates are established for the specific case when \mathcal{V}_s are spaces of piecewise linear functions and A is the Laplace operator posed in a polygonal domain. Since both X_s and X are bounded operators, norm convergence implies that the eigenvalues converge with multiplicity and that the associated spectral projections converge in norm.

For practical computations we typically do not have an access to an orthonormal basis of the space \mathcal{V}_s . One either has to solve a linear system in order to compute the action of the orthogonal projection P_s or seek a computationally more efficient and stable alternative. Simply, the problem (3.8) can be rephrased as a generalized Lyapunov equation, which is obtained by dropping the requirement for an access to an orthonormal basis of \mathcal{V}_s . We pay for this flexibility by the introduction of the Gram matrix M_s of the chosen (non-orthonormal) basis of \mathcal{V}_s . The matrix M_s is called the mass matrix and the equation (3.8) takes the form of the generalized Lyapunov equation

$$K_s Y_s M_s + M_s Y_s K_s = -(M_s \underline{b})(M_s \underline{b})^* . \tag{3.9}$$

Here K_s and M_s are the finite element stiffness and the mass matrix and \underline{b} is the matrix representation of the load vector b in the chosen finite element basis of \mathcal{V}_s , see [21, 33]. Let us note that the finite dimensional operator X_s is represented as the pencil $(M_s Y_s M_s, M_s)$ and so the generalized eigenvalue of the pencil coincide with the eigenvalues of X_s .

The generalized Lyapunov equation (3.9) can be efficiently solved by a projection iterative method. Our method of choice is the projection onto the extended Krylov subspace generated by A and b . This is implemented in MATLAB as the `krylov` algorithm of Simoncini [32]. See also [19] for a possibility to solve such operator equations even when the systems are so large that computing the action of A^{-1} by a sparse direct solver is not feasible.

3.3 Measuring the residual of a low-rank approximation of a Lyapunov equation

Given $z_i \in \mathcal{X}$ let $\hat{Y}_r = \sum_{k=1}^r z_i z_i^*$ be as before. Its Lyapunov equation residual is the sesquilinear form

$$\begin{aligned} \tau(\psi, \phi) = \sum_{i=1}^r & \left[(\psi, A^{1/2} z_i)(A^{-1/2} z_i, \phi) + (A^{-1/2} z_i, \psi)(A^{1/2} z_i, \phi) \right] \\ & - (\psi, A^{-1/2} b)(A^{-1/2} b, \phi), \end{aligned}$$

for $\psi, \phi \in \mathcal{H}$. This form is bounded on \mathcal{H} , it is of finite rank of at most $2r + 1$ and there exists a unique operator $R(\hat{Y}_r)$ such that

$$\tau(\psi, \phi) = (\psi, R(\hat{Y}_r)\phi).$$

Subsequently, the solution $Y = X - X_r$ of the equation

$$(A^{1/2}\psi, Y A^{-1/2}\phi) + (Y A^{-1/2}\psi, A^{1/2}\phi) = \tau(\psi, \phi)$$

can be estimated by

$$\|X - X_r\| \leq \frac{1}{2} \sup_{\psi \neq 0, \phi \neq 0} \frac{|\tau(\psi, \phi)|}{\|\psi\| \|\phi\|} = \frac{1}{2} \|R(X_r)\|.$$

Furthermore, $Y = X - X_r$ is contained in every Schatten ideal, since r is of finite rank.

Let us now formulate an approximation result which might serve as means to construct an approximation to the size of the residual.

Proposition 3.2 *Let $z \in \mathcal{X}$ be given and let the sequence of orthonormal projections P_s be such that $P_s z = z$ for all $s > 0$ and let $P_s \rightarrow I$ strongly. Define*

$$\tau_s(\psi, \phi) = (\psi, A_s^{1/2} z)(A_s^{-1/2} z, \phi) + (A_s^{-1/2} z, \psi)(A_s^{1/2} z, \phi) - (\psi, A_s^{-1/2} b_s)(A_s^{-1/2} b_s, \phi)$$

then $\|R_s(z z^) - R(z z^*)\| \rightarrow 0$ as $s \rightarrow \infty$. The operator $R_s(z z^*)$ is a bounded operator representing the form τ_s and $b_s = P_s b$.*

Proof Given $x, y \in \mathcal{H}$ it follows that

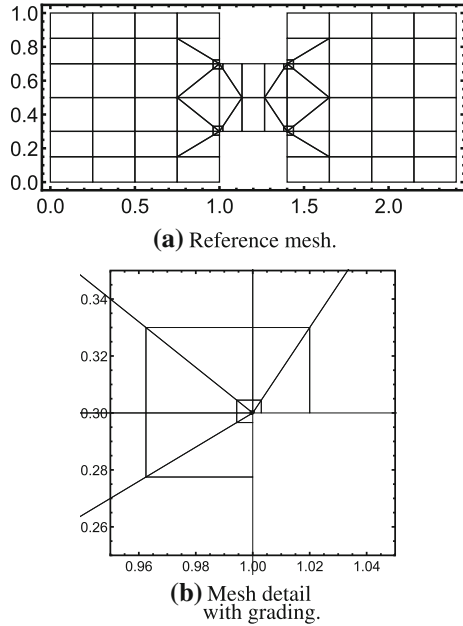
$$\|x x^* - y y^*\| \leq 2 \max\{\|x\|, \|y\|\} \|x - y\| \quad (3.10)$$

and so the strong convergence of $x \rightarrow y$ implies the norm convergence of $x x^*$ to $y y^*$. Now,

$$\tau_s(\psi, \phi) = (\psi, [(A_s^{1/2} z)(A_s^{-1/2} z)^* + (A_s^{-1/2} z)(A_s^{1/2} z)^* - (A_s^{-1/2} b_s)(A_s^{-1/2} b_s)^*] \phi)$$

and the conclusion follows from the norm resolvent convergence of A_s to A and the estimate (3.10) for the norm convergence of the rank one operators. \square

Fig. 3 Geometrically graded mesh. At every reentrant corner the mesh has been graded by applying the replacement rule for ℓ times. Here $\ell = 8$



4 p and hp finite element discretization

As mentioned above, it is crucial to use high order FEM to exploit the regularity of the solution. In this section we give a general overview of the *hp* solver at our disposal. We emphasize that the techniques described here can be applied in general curvilinear setting without any modification. However, we state the families of finite element spaces only in the case of domains partitioned into triangles and quadrilaterals.

Let $\Omega \subset \mathbb{R}^2$ be a open, bounded domain, with Lipschitz boundary $\partial\Omega$, $\mathcal{H} \in H^1(\Omega)$, and let $\mathcal{T} = \{T\}$ be a conforming partition of Ω into convex (curvilinear) triangles and quadrilaterals, which we call a *mesh* or *triangulation*, see Fig. 3. We do not impose any restriction on the number of curved edges. Any curved elements are handled using standard blending function techniques (cf. [34]).

For a given element T and non-negative integer m , we define the local polynomial space $\mathbb{Q}_m(T)$ as follows. If T is a triangle, then $\mathbb{Q}_m(T)$ consists of the polynomials of total degree $\leq m$, so $\dim \mathbb{Q}_m(T) = (m + 2)(m + 1)/2$. If T is a quadrilateral, then $\mathbb{Q}_m(T)$ consists of polynomials of degree $\leq m$ in each variable, so $\dim \mathbb{Q}_m(T) = (m + 1)^2$.

For a given triangulation, \mathcal{T} , let $\mathbf{p} : \mathcal{T} \rightarrow \mathbb{N}$ be a function that assigns a positive integer to each element $T \in \mathcal{T}$. This map is called a *p*-vector. We define the corresponding finite element space

$$\mathcal{V} = \mathcal{V}(\mathcal{T}, \mathbf{p}) = \{v \in \mathcal{X} : v|_T \in \mathbb{Q}_{\mathbf{p}(T)}(T) \text{ for all } T \in \mathcal{T}\}. \tag{4.1}$$

We note that $\mathcal{V} \subset C(\overline{\Omega})$.

Let $\mathcal{F} = \{\mathcal{T}_\ell\}$ be a family of nested meshes obtained from successive refinements of an initial coarse mesh, where the index $\ell \geq 0$ refers to a refinement level. To account for possible *singular points* at points on the boundary where there are non-convex corners, we apply a geometric grading of element sizes toward singular points that takes into account this a priori knowledge [31, Section 4.5]. Beginning with a coarse mesh \mathcal{T}_0 in which the vertex graph distance between singular points (i.e., the minimal number of edges in a path connecting these points) is at least two, the mesh grading approach is implemented using element-level replacement rules employing exact geometry description as described in [13]. The element layers are created by nested application of replacement rules on every element touching a singular point. At each step, only the elements touching the singular point created at the previous one are refined making the bookkeeping of the layers simple. This is illustrated in Fig. 3. Notice that these replacement rules need not be the same at different levels, since a rule for a quadrilateral element may result in a triangle touching the singular point.

Given such a family of meshes, we distinguish two families of finite element spaces defined on them. We refer to the first as the p -method family because it uses a fixed polynomial degree for every element in the mesh. For this family, the polynomial degree p is chosen and applied to each element in the p th mesh in the family, $\mathcal{T}_p \in \mathcal{F}$, i.e. $\mathbf{p}(T) = p$ for all $T \in \mathcal{T}_p$. We denote the finite element spaces in this family by $\mathcal{V}_{1,p}$, and use $1 \leq p \leq 8$ for our experiments. We note that the spaces are nested, $\mathcal{V}_{1,p} \subset \mathcal{V}_{1,p+1}$. We refer to the second family as the hp -family because it uses variable polynomial degrees in the mesh. For the second family, given a polynomial degree p , the mesh \mathcal{T}_p is chosen as in the first family, but polynomial degrees are no longer assigned uniformly throughout the mesh. All elements touching a singular point are assigned polynomial degree 1, the next layer of elements are assigned polynomial degree 2, and so on, until polynomials of degree p are achieved at the p th layer. Any elements that are greater than p layers away from all singular points are also assigned polynomial degree p . The initial mesh and refinement scheme ensures that there is no ambiguity in how polynomial degrees are assigned to each element. We denote the finite element spaces in this by $\mathcal{V}_{2,p}$, using $1 \leq p \leq 12$ for our experiments. As before, the spaces are nested, $\mathcal{V}_{2,p} \subset \mathcal{V}_{2,p+1}$, and we also note that $\mathcal{V}_{2,p} \subset \mathcal{V}_{1,p}$.

For the hp -family it is necessary to construct the local basis functions in such a way that varying the local polynomial order still results in a continuous formulation. We distinguish between three types of polynomial functions on an element: *vertex functions*, which vanish on all vertices except one; *edge functions*, which vanish on all edges except one; and *element functions* (interior bubble functions), which vanish on all edges. On the global (mesh) level, vertex functions are supported in the patch of elements sharing that vertex, edge functions are supported in the (one or two) elements sharing an edge, and element functions are supported in a single element. It is this distinction between the types of polynomial functions that enables one to build elements in which the degrees of the element functions may differ from those of the edge functions, and the degree used on one edge may differ from that used on another. In fact, this is precisely what is done in the hp -family, $\mathcal{V}_{2,p}$, to allow for variable $p(T)$. In particular, when T and T' are adjacent elements whose assigned polynomial degrees differ by one, say $p(T) = m$ and $p(T') = m + 1$, the polynomial degree of the edge functions associated with their shared edge is taken to be $m + 1$.

5 A greedy auxiliary subspace mesh refinement

In what follows we will assume that we have a sequence of spaces $\mathcal{V}_s = \mathcal{V}_{s, p_s}$ such that $\mathcal{V}_{s_1} \subset \mathcal{V}_{s_2}$ for $s_1 < s_2$ and $P_s \rightarrow I$ strongly as $s \rightarrow \infty$. Further, we will assume that A is a divergence type operator with analytic coefficients and posed in a polygonal domain Ω . We will use the notation X for the solution of the Lyapunov equation (3.2) and X_s will be the solution of the Lyapunov equation projected onto the subspace \mathcal{V}_s . The X_k will denote the operator (3.6), which is the function of the operator A .

In this section we aim to present error indicator which will be shown to be converging to the error estimate in the sense of Proposition 3.2. We will also show what role approximate A analyticity of the eigenvectors of X might play. Also note that Proposition 3.2 gives a computable approximation of the error estimator. Namely the operator $R_s(Z_r)$ is representable as a short sum of rank one operators and so its maximal singular value can be efficiently computed by a standard iterative procedure. However, we will now take an alternative avenue, but before we do so, let us present an extension to Proposition 3.1.

Theorem 5.1 *Let $r \in \mathbb{N}$ be given such that $\mu_r(X) > \mu_{r+1}(X)$ and let $b \in \mathcal{H}$. Then for $\mathcal{Z} = \text{span}\{z_1, \dots, z_r\} \subset \mathcal{V}$ and $z_i, i = 1, \dots, r$ pairwise orthogonal and of norm one, we have, for $i = 1, \dots, r$, the estimate*

$$|\mu_i(X) - \rho(z_i, X_k)| \leq \frac{C_{St} \|b\|^2}{2\lambda_1} \exp(-\pi\sqrt{k}) + \frac{\|b\|^2(1 + C_{St} \exp(-\pi\sqrt{k}))}{2\lambda_1} \|\tilde{u}_i - z_i\|^2.$$

Here $\tilde{u}_i, \|\tilde{u}_i\| = 1$ are A -analytic eigenvectors such that $\mu_i(X_k) = (\tilde{u}_i, X_k \tilde{u}_i)$. Further, we have

$$\begin{aligned} \|\mu_i(X)u_iu_i^* - \rho(z_i, X_k)z_iz_i^*\| &\leq \frac{C_{St} \|b\|^2}{2\lambda_1} \exp(-\pi\sqrt{k}) \\ &+ \frac{\|b\|^2(1 + C_{St} \exp(-\pi\sqrt{k}))}{2\lambda_1} \left[\|\tilde{u}_i - z_i\|^2 + \|\tilde{u}_i - z_i\| \right]. \end{aligned}$$

Proof The proof follows by combining Proposition 3.1 and estimate (2.2). Recall that X_k is a positive self-adjoint operator, and then using (2.2) compute

$$\begin{aligned} |\mu_i(X) - \rho(z_i, X_k)| &\leq |\mu_i(X) - \mu_i(X_k)| \\ &+ |\mu_i(X_k) - \rho(z_i, X_k)| \\ &\leq \frac{C_{St} \|b\|^2}{2\lambda_1} \exp(-\pi\sqrt{k}) + \|X_k\| \|\tilde{u}_i - z_i\|^2. \end{aligned}$$

To finish the proof we need to estimate $\|X_k\|$ by an a priori bound. To this end, let ψ_i be an orthonormal system of eigenvectors of the self-adjoint and positive operator A . Then, see also [11, Equation (3.2)],

$$\|X_k\| \leq \|X\| + \|X_k - X\|$$

and

$$|(\psi_i, X\psi_j)| = \frac{|(\psi_i, b)| |(b, \psi_j)|}{\lambda_i + \lambda_j} \leq \frac{|(\psi_i, b)| |(b, \psi_j)|}{2\lambda_1}$$

follows. We now compute

$$\|X\| \leq \sqrt{\text{tr}(X^*X)} \leq \frac{1}{2\lambda_1} \|b\|^2.$$

and then conclude

$$\|X_k\| \leq \frac{1}{2\lambda_1} \|b\|^2 (1 + C_{St} \exp(-\pi\sqrt{k})).$$

The last inequality of the theorem follows by combining the estimate for $|\mu_i(X) - \rho(z_i, X_k)|$ and the estimate (3.10). □

This theorem indicates that in the case of A being the divergence type operator with analytic coefficients and posed in the polygonal domain the functions \tilde{u}_i are going to be infinitely differentiable functions which are also A -analytic. For such functions there exist constants $C_i > 0$ and $\gamma_i > 0$ such that

$$\min\{\|\tilde{u}_i - z\| : z \in \mathcal{Y}_{l_s, p_s}\} \leq C_i \exp(-\gamma_i \sqrt{\dim(\mathcal{Y}_{l_s, p_s})}). \tag{5.1}$$

Subsequently one concludes—using the second inequality of Theorem 5.1 r times—that there exists a rank r of approximation X , whose eigenvectors will belong to the space \mathcal{Y}_{l_s, p_s} and the eigenvalues will approximate the eigenvalues of the operator X —in the sense of Theorem 5.1—with the a priori estimate of the form (5.1).

We now present an analysis of our approximation of the a posteriori error estimator. It is based on the norm convergence of the solutions of the projected Lyapunov equations. Given that the operator A is self-adjoint and that the sequence of subspaces is monotonic we conclude, based on [20, Section 4.1.2 and Section 5], that $\|X - X_s\| \rightarrow 0$. Then a simple calculation shows that for a given s_1 and for each $\varepsilon > 0$ exists $s_2 > s_1$ such that

$$\|X - X_{s_1}\| \leq (1 + \varepsilon) \|X_{s_2} - X_{s_1}\|.$$

In the case in which $X_{s_i} = \sum_{j=1}^r z_j^{(s_i)} (z_j^{(s_i)})^*$, $i = 1, 2$ we have, using (3.10), the estimate

$$\begin{aligned} \|X - X_{s_1}\| &\leq (1 + \varepsilon) \|X_{s_2} - X_{s_1}\| \\ &\leq (1 + \varepsilon) \max\{\|z_j^{(s_i)}\| : i = 1, 2, j = 1, \dots, r\} \sum_{j=1}^r \|z_j^{(s_2)} - z_j^{(s_1)}\|. \end{aligned} \tag{5.2}$$

In the case of A being a divergence type operator and $\mathcal{V}_s = \mathcal{V}_{l_s, p_s}$, we might take $\mathcal{V}_{s_1} = \mathcal{V}_{l_{s_1}, p_{s_1}}$ and $\mathcal{V}_{s_2} = \mathcal{V}_{l_{s_1}, p_{s_2}}$. That is to keep the refinement level l_s constant and studying the pure p refinement for improving the auxiliary subspace.

We now argue that instead of the auxiliary subspace adaptively we opt for a fixed refinement by uniformly increasing the polynomial degree by two. This argument will be partly heuristic.

A more refined information on the error can be obtained if one interprets the low-rank approximation task in the context of spectral approximations, as was done in Proposition 3.1. Recall the estimate (2.1), which according to [15, 18]—in the case in which ψ and u are of norm one—implies

$$|\mu - \rho(\psi, H)| \leq \mu_1 \|\psi - u\|^2 .$$

Since we typically do not have access to the vector u , we will use the auxiliary subspace technique from [3, 12] in a combination with the saturation assumption to approximate $\|u - \psi\|$. Let us demonstrate the auxiliary subspace error estimation for the interpolation error. Let $\mathcal{V} \subset \mathcal{V}^{aux} \subset \mathcal{H}$ be two finite element spaces and let \mathcal{H} be an appropriate Sobolev space. Note that in this section u^{aux} will not be denoting the duality paring for u , but rather that u^{aux} is an element of \mathcal{V}^{aux} for some function u . Then for the finite element approximation $\psi \in \mathcal{V}$ of the solution $u \in \mathcal{H}$ we can define the approximate error function $\varepsilon \approx u - \psi$ by projecting the error $u - \psi$ onto the space $\mathcal{V}_{\ominus}^{aux} = \mathcal{V}^{aux} \ominus \mathcal{V}$. With this we immediately get the efficiency bound $\|\varepsilon\| \leq \|u - \psi\|$ and the reliability bound is obtained as a combination of the strong Cauchy inequality for the subspace $\mathcal{V}_{\ominus}^{aux}$ and the saturation assumption for the subspace \mathcal{V}^{aux} . In the context of the eigenvalue problem, a saturation assumption for analyzing the eigenvalue approximation error has been used in [25]. We will adapt this assumption to the case of a compact operator. Assuming we are interested in the top r eigenvalues of some compact operator H , the saturation assumption holds if there is a constant $0 \leq \gamma < 1$ such that

$$\mu_i(H) - \rho_i(\mathcal{V}^{aux}, H) \leq \gamma(\rho_i(\mathcal{V}^{aux}, H) - \rho_i(\mathcal{V}, H)), \quad i = 1, \dots, r .$$

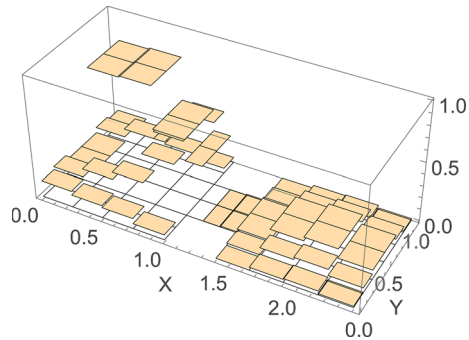
It was then shown in [25] that this is equivalent to stating

$$(\mu_i(H) - \rho_i(\mathcal{V}, H)) \leq \frac{1}{1 - \gamma}(\rho_i(\mathcal{V}^{aux}, H) - \rho_i(\mathcal{V}, H)). \tag{5.3}$$

This statement is adapted here from the case of an unbounded operator with a compact resolvent, which was considered in [25].

Remark 5.1 As the spaces \mathcal{V} and \mathcal{V}^{aux} , for practical hp -finite element computations, we choose the given hp -space and the hp -space obtained on the same mesh by increasing the polynomial degree by two.

Fig. 4 The approximate error function. $\varepsilon_1 = \hat{u}_1^{aux} - \hat{u}_1$, with \hat{u}_1^{aux} and \hat{u}_1 at $p = 6$ and $p = 4$, respectively. Elemental L^2 -error of ε_1 distributed over the mesh, relative scale over $[0, 1]$



Proposition 5.1 *Let the saturation assumption hold for the eigenvalues of some positive compact operator H . Then*

$$\mu_i(H) - \rho_i(\mathcal{V}, H) \leq \frac{\mu_1(H)}{1 - \gamma} \|\hat{u}_i^{aux} - \hat{u}_i\|^2$$

where \hat{u}_i^{aux} , $\|\hat{u}_i^{aux}\| = 1$ is a vector such that $\rho_i(\mathcal{V}^{aux}, H) = (\hat{u}_i^{aux}, H\hat{u}_i^{aux})$ and \hat{u}_i , $\|\hat{u}_i\| = 1$ is such that it verifies $\rho_i(\mathcal{V}, H) = (\hat{u}_i, H\hat{u}_i)$.

Proof The result of Knyazev [37] implies

$$\begin{aligned} |\rho_i(\mathcal{V}^{aux}, H) - \rho_i(\mathcal{V}, H)| &\leq \rho_1(\mathcal{V}^{aux}, H) \|\hat{u}_i^{aux} - \hat{u}_i\|^2 \\ &\leq \mu_1(H) \|\hat{u}_i^{aux} - \hat{u}_i\|^2 \end{aligned} \tag{5.4}$$

where \hat{u}_i , $\|\hat{u}_i\| = 1$ is such that $\rho_i(\mathcal{V}, H) = (\hat{u}_i, H\hat{u}_i)$. The statement of the theorem follows when one combines (5.4) with (5.3). □

Our greedy strategy for mesh refinement for such operators is to refine those triangles where the restrictions of the functions $\varepsilon_i = \hat{u}_i^{aux} - \hat{u}_i$, $i = 1, \dots, r$ are above a given threshold. One sees on Fig. 4 how ε_1 picks out the elements near the first two reentrant corners for refinement, but not the second two.

Corollary 5.1 *Let $r \in \mathbb{N}$ be given such that $\mu_r(X) > \mu_{r+1}(X)$, $b \in \mathcal{H}$. Given $k \in \mathbb{N}$, and let $0 \leq \gamma < 1$ be the saturation constant for X_k . Then for $\mathcal{L} = \text{span}\{z_1, \dots, z_r\} \subset \mathcal{V}$ and z_i , $i = 1, \dots, r$ pairwise orthogonal and of norm one, we have the estimate*

$$|\mu_i(X) - \rho(z_i, X_k)| \leq \frac{C_{St} \|b\|^2}{2\lambda_1} \exp(-\pi\sqrt{k}) + \frac{\|b\|^2(1 + C_{St} \exp(-\pi\sqrt{k}))}{2\lambda_1(1 - \gamma)} \|\tilde{u}'_i - z_i\|^2.$$

Here \tilde{u}'_i , $\|\tilde{u}'_i\| = 1$ are such that $\rho_i(\mathcal{V}^{aux}, X_k) = \rho(\tilde{u}'_i, X_k)$.

Proof Let $0 \leq \gamma < 1$ be the saturation constant for X_k , then

$$\begin{aligned}
 |\mu_i(X) - \rho(z_i, X_k)| &\leq |\mu_i(X) - \mu_i(X_k)| + |\mu_i(X_k) - \rho(z_i, X_k)| \\
 &\leq |\mu_i(X) - \mu_i(X_k)| + \frac{\|X_k\|}{1 - \gamma} \|\tilde{u}_i^{k'} - z_i\|^2.
 \end{aligned}$$

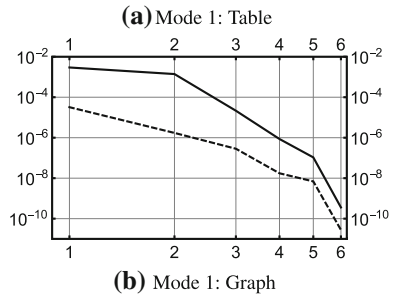
The proof follows using the estimate from Theorem 5.1. □

Remark 5.2 Let us first address the assumptions of Proposition 5.1 and Corollary 5.1. Recall that X_k is the function of the operator A . The eigenvalue saturation assumption with $0 \leq \gamma < 1$ for operator $H = X_k$ in the case in which A is a divergence type operator with analytic coefficients can be justified in the same way as it was done [25]. Namely, Neymeyer argues in [25, Section 4] that the saturation assumption with $\gamma < 1$ holds for the Laplace eigenvalue problem and \mathcal{V} being chosen as the space of piecewise linear functions and \mathcal{V}^{aux} as the space of piecewise quadratic functions. The argument rests on the fact that eigenfunctions of the Laplace operator are harmonic and a harmonic function whose restriction to the open set of positive measure is zero is a zero function. Equivalently, eigenfunctions of the operator X_k are analytic functions, this follows from Nelson [24] for this particular class of operators, and so the same argument holds. Note that this does not necessarily imply that $1/(1 - \gamma)$ is small, only that it is finite. Further more detailed analysis would be needed to assess the size of γ or to prove that it only depends on the shape regularity of the triangulation and possibly on the polynomial degree. For similar considerations in the context of the eigenfunction approximations of an unbounded operator see [10].

Note that we substitute \hat{u}_i^{aux} , the i -th eigenvector of X_{γ_1} —the solution of the Lyapunov equation projected onto \mathcal{V}^{aux} . We could quantify, in principle, the error $\|\hat{u}_i^{aux} - \tilde{u}_i'\|$ by a direct perturbation analysis. The operators X_k are given by an explicit formula and the action of X_k onto a vector is in principle computable using contour integration techniques from [4, 5, 9]. This analysis would however be quite technical, would require additional technical apparatus and the estimates will likely be unnecessarily pessimistic. For instance, estimates from [4, 5] are only valid for quasi-uniformly refined meshes and low order finite elements. They also depend on the approximability of the loading vector b , even though the eigenvectors of the operator X_k are A -analytic for any $b \in \mathcal{H}$. An extension to higher order finite elements is plausible, but would require extended technical work which is beyond the scope of this paper. Even then, the constant $1/(1 - \gamma)$ can potentially be very large and the overall estimate would be quite pessimistic. Instead, we opted to make a heuristic choice and monitor only $\|\hat{u}_i^{aux} - \hat{u}_i\|$. We will report on the numerical experiments where we will compare $\|\hat{u}_i^{aux} - \hat{u}_i\|^2$ with the error in the i -th eigenvalue. Note also that Proposition 3.2 gives further justification, or an alternative interpretation, for estimating the residual from an auxiliary subspace. We can compute the norm $\|R_{s_2}(\sum_{i=1}^r \hat{u}_i \hat{u}_i^*)\|$ by a data sparse SVD computation. Instead of going down this avenue, we can observe that $\sum_{i=1}^r \|\hat{u}_i^{aux} - \hat{u}_i\|$ is an estimator of this norm and this is the approach which we take (Fig. 5).

Fig. 5 Convergence of $\|\hat{u}_1^{aux} - \hat{u}_1\|^2$ (solid line) and $|\mu_1(X) - \mu_1(X_r)|$ (dashed line), as $p = 1, \dots, 6$, on a graded mesh with constant p at $\ell = 4$

p	$\ \hat{u}_1^{aux} - \hat{u}_1\ ^2$	$ \mu_1(X) - \mu_1(X_r) $
1	2.9516e-03	3.2195e-05
2	1.3962e-03	1.7227e-06
3	2.1491e-05	2.8466e-07
4	8.6205e-07	1.7452e-08
5	1.0649e-07	6.9308e-09
6	3.5918e-10	2.7812e-11



6 Numerical experiments

We will now present an example with a loading b which is not a separable function. Bearing in mind the a priori error estimate (5.1), we will be presenting the convergence plots in which the logarithm of the error will be on the y axis and \sqrt{N} , $N = \dim \mathcal{V}_{l,p}$ on the x axis. A straight line plot in this coordinate system implies the convergence of the order $O(\exp(-\gamma\sqrt{N}))$, for some $\gamma > 0$. We will see that we observe such convergence—if with a smaller γ —even for the domain with reentrant corners and a less regular b .

Example 6.1 (Dumbbell B) Let us consider a classical Laplace dumbbell problem with computational domain $\Omega = ([0, 2.4] \times [0, 1]) \setminus (([1, 1.4] \times [0, 0.3]) \cup ([1, 1.4] \times [0.7, 1]))$. We chose the operator A in the Lyapunov equation

$$AX + XA^* = -bb^*$$

to be the Laplace operator with the zero Dirichlet boundary conditions. The function $b(x_1, x_2)$ is taken to be the indicator function:

$$b(x_1, x_2) = \text{ind}\{|x_1 - 1/2| + |x_2 - 0.45| \leq 1/4\}.$$

The solution X is remarkably of numerical rank = 2. See Fig. 6 for illustrations of the function b , the first column of X , and the two dominant eigenmodes of X .

In our numerical experiments we quantify more precisely the performance of the p - and hp -discretizations of Examples 2.2 (Dumbbell A) and 6.1 (Dumbbell B). In both examples the configuration is exactly the same except for the loading which is smooth in Example 2.2 and discontinuous and not symmetric with respect to the domain in Example 6.1. The two quantities of interest are the sum of eigenvalues of the solution X and the rank of X .

The mesh grading strategy described in Sect. 4 is used in two different sets of experiments for both examples. First, we consider levels $\ell = 0, \dots, 8$, and for every ℓ we compute the solutions for all (constant) $p = 1, \dots, 8$. Here $\ell = 0$ means that the background mesh is used without any refinements, this is sometimes referred to as the “pure p -version”-approach. Second, we compute a proper hp -sequence, where as $\ell = 1, \dots, 11$, we compute the p -vector \mathbf{p} using maximal $p = \ell + 1$. The final solution of the hp -sequence is taken as the reference solution.

For every individual experiment the quantities of interest have been computed.

6.1 Convergence in eigenvalues

The observed convergence in the relative error in the sum of eigenvalues $\sum_i \lambda_i$ is illustrated in Figs. 7 and 8. In both cases an overall picture over the set of experiments is given with a detail plot indicating the region where the loss of convergence rate is observed. As expected, the hp -discretization is the most efficient one in both cases. Interestingly, the effect of the singularities is evident especially if one focuses on the levels $\ell = 0$ and $\ell = 2$, where it is evident that if the geometric grading is not taken to sufficiently high level, there is a loss of convergence rate. On the other hand, for the p -discretization, there appears to be an optimal level (here $\ell = 4$) beyond which the observed rate does not increase, yet the constant does. This is the reason why for the Dumbbell B the $p = 6, 8$ graphs have been omitted.

6.2 Asymptotic behavior of the numerical rank

More unusual error measure is the observed numerical rank of the solution. This is meant in the sense of the Proposition 3.1 with $\tau = 10^{-8}$. At first one could suspect that the results of Tables 1 and 2 are a simple consequence of keeping the tolerances of the k_{pik} -algorithm constant even as the dimensions of the cases increase. However, by comparing the observed ranks with the numbers of degrees of freedom in Figs. 7 and 8 it is clear that this connection does not explain these results. The connection between the levels ℓ and the polynomial order p indicates that the key here is the accurate capturing of the effects due to the singular points. Nicely tying with the discussion above, again the $\ell = 4$ appears to be the level where the singularities are first captured.

It is clear that for the discontinuous loading (Dumbbell B) the observed ranks are slightly higher than those for Dumbbell A with smooth loading.

For a qualitative view of the reduction in rank, in both Figs. 9 and 10 two sets of eigenmodes from the hp -sequence corresponding to $\ell = 3$ and 4 are shown. Considering the 3D-plots of Fig. 2 for Dumbbell A, one can see that the superfluous modes (Fig. 9c, d) have features that are ultimately subsumed to the second mode. The explanation here is that the singularities pollute the solution and this results in modes with very small eigenvalues. For the Dumbbell B the situation is even more interesting. In fact, the both final two modes have ghost modes at $\ell = 3$. Here the eigenvalues are

$$\lambda_{\ell=3}^B = \{1.4 \times 10^{-3}, 1.1 \times 10^{-4}, 6.7 \times 10^{-6}, 1.2 \times 10^{-6}, 3.3 \times 10^{-7}, 8.5 \times 10^{-8}\},$$

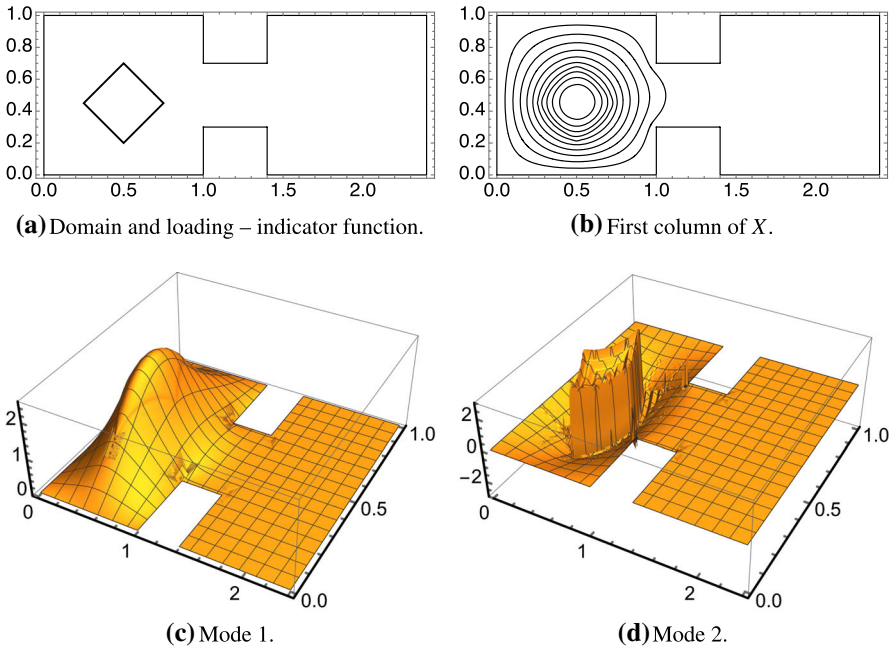


Fig. 6 A dumbbell domain with an indicator function loading. The columns of X can be plotted and the first one is illustrated. In this case the operator has numerical rank = 2, and the corresponding modes are shown

Fig. 7 Dumbbell A: Effect of the mesh grading. Relative error in the sum of eigenvalues versus the number of DOFs. For every indicated level ℓ the error is computed for a constant $p = 1, \dots, 8$. The reference solution (solid black line) is the hp -sequence with the proper p -vector \mathbf{p} for levels $\ell = 1, \dots, 11$

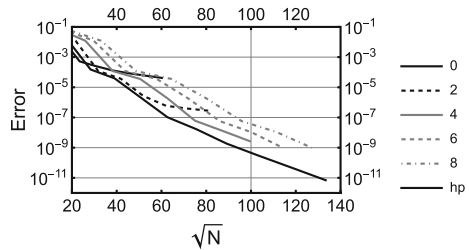


Fig. 8 Dumbbell B: Effect of the mesh grading. Relative error in the sum of eigenvalues versus the number of DOFs. For every indicated level ℓ the error is computed for a constant $p = 1, \dots, 8$. The reference solution (solid black line) is the hp -sequence with the proper p -vector \mathbf{p} for levels $\ell = 1, \dots, 11$

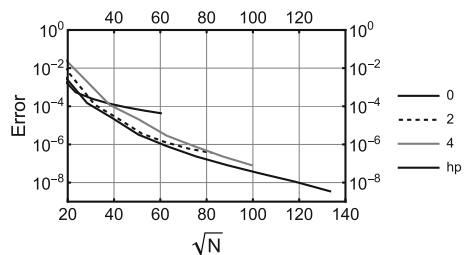


Table 1 Dumbbell A: Observed ranks

	p								P
	1	2	3	4	5	6	7	8	–
(a) p version									(b) hp version
ℓ									ℓ
0	5	6	6	6	6	5	5	4	0 5
1	5	6	6	5	5	5	2	2	1 6
2	5	6	4	2	2	2	2	2	1 6
3	5	3	2	2	2	2	2	2	3 4
4	6	2	2	2	2	2	2	2	4 2
5	5	2	2	2	2	2	2	2	5 2
6	3	2	2	2	2	2	2	2	6 2
7	2	2	2	2	2	2	2	2	7 2
8	2	2	2	2	2	2	2	2	8 2

The row index is the refinement level ℓ and the column index is the (constant) polynomial order p or the p -vector \mathbf{p}

Table 2 Dumbbell B: Observed ranks

	p								P
	1	2	3	4	5	6	7	8	–
(a) p version									(b) hp version
ℓ									ℓ
0	5	8	8	9	7	4	2	2	0 5
1	7	8	8	8	7	3	2	2	1 8
2	7	7	6	3	2	2	2	2	2 8
3	7	5	3	2	2	2	2	2	3 6
4	6	3	2	2	2	2	2	2	4 2
5	5	2	2	2	2	2	2	2	5 2
6	2	2	2	2	2	2	2	2	6 2
7	3	2	2	2	2	2	2	2	7 2
8	2	2	2	2	2	2	2	2	8 2

The row index is the refinement level ℓ and the column index is the (constant) polynomial order p or the p -vector \mathbf{p}

and

$$\lambda_{\ell=4}^B = \{1.5 \times 10^{-3}, 2.1 \times 10^{-5}\},$$

indicating the this “summation of modes” is also reflected in the eigenvalues as well.

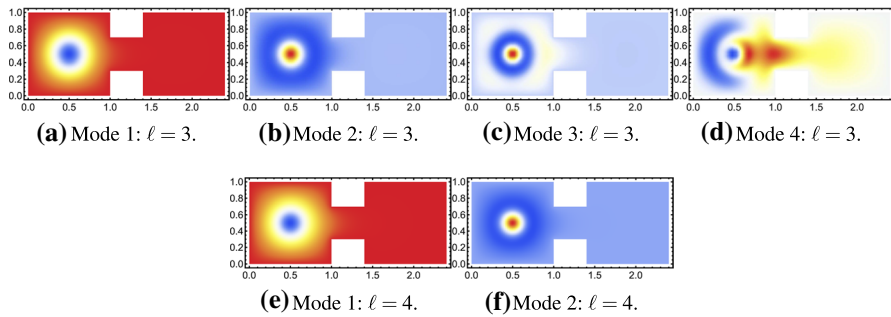


Fig. 9 Dumbbell A: Decreasing rank: Qualitative view via plots from the hp -sequence. As indicated in Table 1b, the numerical rank changes from 4 to 2 as the level ℓ changes from 3 to 4. As the corner singularities are better captured the two superfluous modes are subsumed into the second mode

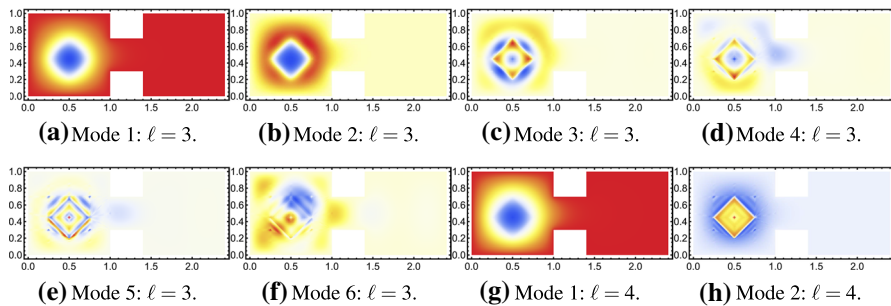


Fig. 10 Dumbbell B: Decreasing rank: Qualitative view via plots from the hp -sequence. As indicated in Table 2b, the numerical rank changes from 4 to 2 as the level ℓ changes from 3 to 4. As the corner singularities are better captured the first two modes and the following four at $\ell = 3$ are subsumed into the first and second mode at $\ell = 4$, respectively

7 Conclusions

In this paper we have presented the approximate, controlled by a threshold on the numerical rank, regularity structure of the solution operator to the operator Lyapunov equation. We point out the following consequence of the use of the threshold parameter to define the numerical rank—the dominant eigenvalues of the solution X can be seen to form the cluster. We are not computing the average of this cluster as is done in he work of Osborn [28]. We instead compute and track the asymptotic behavior of the sum of the clustered eigenvalues, as is done more generally in majorisation estimates from [17]. In this sense our high order greedy adaptivity strategy constructs a subspace which captures the trace better (not the multiplicity) than a standard approximation which is oblivious of the regularity of the eigenfunctions associated to the cluster of eigenvalues forming the trace. The subspace which we constructed is much smaller than the one which the Theorem 5.1 assumed. Yet, even with such a crude approach we still had exponential (but slower than optimal) convergence in the number of the finite element degrees of freedom due to the robust regularity structure of the approximate operators X_k . The future work will be focused on more singular right

hand sides (like boundary forcing) and in tighter analysis of the regularity structure of the eigenfunctions of the solution operator X .

Acknowledgements The preparatory research as well as the initial impetus and the problem suggestion were laid down at EPF-Lausanne where L.G. spent a part of his sabbatical leave in 2014 in the group of Prof. Dr. D. Kressner. The authors are thankful to D. Kressner for his hospitality and discussions on the mesh adaptivity in the context of low-rank solution methods for the Lyapunov equation. Also, we want to thank the anonymous referees whose careful reading and suggestions greatly improved the paper.

Funding Open Access funding provided by Aalto University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

A Sketch of the result in the general case

In this “Appendix” we will briefly outline the case in which the forcing $b \in \mathcal{X}'$ is such that $\|A^{-\alpha}b\| < \infty$ for some $0 \leq \alpha < 1/2$. The results of Opmeer from [27] allow for the treatment of the case of the sectorial, not necessarily self-adjoint, generator A . The low rank construction in this case depends on the angle of the sector where the resolvent is analytic, see [27]. In order to be explicit in the presentation of the results, we will concentrate on the case of the self-adjoint and positive definite A . For every $t > 0$ the vector

$$v_s = \exp(-tA)(A^{-\alpha}b)$$

is an A -analytic vector and so in particular for every $n \in \mathbb{N}$ the vector v_s verifies $v_s \in \text{dom}(A^n) \subset \text{dom}(A) \subset \text{dom}(A^\alpha)$ and we conclude $\|A^\alpha v_s\| < \infty$. Subsequently,

$$v = A^\alpha v_s = A^\alpha \exp(-tA)(A^{-\alpha}b) =: \exp(-tA)b$$

Furthermore, $\|A^n v\| < \infty$ for any $n \in \mathbb{N}$ and so the vector v is an infinite vector for the operator A .

Now set

$$h = \pi \sqrt{\frac{\alpha}{1 - 2\alpha}}, \quad \tilde{t}_p = \tilde{\omega}_p := \exp(ph),$$

then

$$X_k = h \sum_{p=-k}^{\lceil \frac{1}{h} \log\left(\frac{1-2\alpha}{2\lambda_1} kh\right) + 1 \rceil} \tilde{\omega}_p (\exp(-\tilde{t}_p A)b)(\exp(-\tilde{t}_p A)b, \cdot)$$

is a low rank approximation to the unique positive self-adjoint solution X of the Lyapunov equation and the eigenvectors of X_k are at least infinite vectors for the operator A . In fact it can be show, using the similar argument as in [30], that the vector v is A -analytic. We have now outlined the proof of the following corollary

Corollary A.1 *Assume A is self-adjoint and positive definite and let $\tau > 0$ be given and let the forcing $b \in \mathcal{X}'$ be such that $A^{-\alpha}b$, for $0 \leq \alpha < 1/2$. By X denote the unique self-adjoint and positive solution of (3.2). Then there exist numbers r and k such that*

$$|\operatorname{tr}(X) - \sum_{i=1}^r \mu_i(X_k)| \leq \tau \operatorname{tr}(X)$$

and eigenvectors $\hat{u}_i \neq 0$, $X_k \hat{u}_i = \mu_i(X_k) \hat{u}_i$ are A -analytic.

The corollary holds even when A is not self-adjoint, but it is a generator of an exponentially stable analytic semigroup. We did not go to this generality in this paper, since we were not planning to present any examples with such a singular forcing b and non self-adjoint A . Doing so would have only made the paper less readable. However, all of the subsequent results can be generalised to this level. We plan to take this avenue in a subsequent paper where we will present appropriate set of examples.

References

1. Antoulas, A.C., Sorensen, D.C.: Approximation of large-scale dynamical systems: an overview. *Int. J. Appl. Math. Comput. Sci.* **11**(5), 1093–1121 (2001). Numerical analysis and systems theory (Perpignan, 2000)
2. Babuška, I., Guo, B.Q.: Approximation properties of the h - p version of the finite element method. *Comput. Methods Appl. Mech. Eng.* **133**(3–4), 319–346 (1996)
3. Bank, R.E., Owall, J.S.: Some remarks on interpolation and best approximation. *Numer. Math.* **137**(2), 289–302 (2017)
4. Bonito, A., Lei, W., Pasciak, J.E.: Numerical approximation of space-time fractional parabolic equations. *Comput. Methods Appl. Math.* **17**(4), 679–705 (2017)
5. Bonito, A., Lei, W., Pasciak, J.E.: On sinc quadrature approximations of fractional powers of regularly accretive operators. *J. Numer. Math.* **27**(2), 57–68 (2019)
6. Chatelin, F.: Spectral approximation of linear operators, vol. 65. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM) (2011)
7. Curtain, R.F., Sasane, A.J.: Hankel norm approximation for well-posed linear systems. *Syst. Control Lett.* **48**(5), 407–414 (2003). [https://doi.org/10.1016/S0167-6911\(02\)00301-8](https://doi.org/10.1016/S0167-6911(02)00301-8)
8. Davis, C., Kahan, W.M.: The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7**, 1–46 (1970). <https://doi.org/10.1137/0707001>
9. Gavriljuk, I.P., Hackbusch, W., Khoromskij, B.N.: Data-sparse approximation to a class of operator-valued functions. *Math. Comput.* **74**(250), 681–708 (2005)
10. Giani, S., Grubišić, L., Hakula, H., Owall, J.S.: A posteriori error estimates for elliptic eigenvalue problems using auxiliary subspace techniques. *J. Sci. Comput.* **88**(3), 55 (2021). <https://doi.org/10.1007/s10915-021-01572-2>
11. Grubišić, L., Kressner, D.: On the eigenvalue decay of solutions to operator Lyapunov equations. *Syst. Control Lett.* **73**, 42–47 (2014)
12. Hakula, H., Neilan, M., Owall, J.S.: A posteriori estimates using auxiliary subspace techniques. *J. Sci. Comput.* **72**(1), 97–127 (2017)
13. Hakula, H., Tuominen, T.: Mathematica implementation of the high order finite element method applied to eigenproblems. *Computing* **95**(1), 277–301 (2013)

14. Hansen, S., Weiss, G.: New results on the operator Carleson measure criterion. *IMA J. Math. Control Inform.* **14**(1), 3–32 (1997). Distributed parameter systems: analysis, synthesis and applications, Part 1
15. Knyazev, A., Jujunashvili, A., Argentati, M.: Angles between infinite dimensional subspaces with applications to the Rayleigh–Ritz and alternating projectors methods. *J. Funct. Anal.* **259**(6), 1323–1345 (2010)
16. Knyazev, A.V., Argentati, M.E.: Principal angles between subspaces in an a -based scalar product: algorithms and perturbation estimates. *SIAM J. Sci. Comput.* **23**(6), 2008–2040 (2002). <https://doi.org/10.1137/S1064827500377332>
17. Knyazev, A.V., Argentati, M.E.: Majorization for changes in angles between subspaces, Ritz values, and graph Laplacian spectra. *SIAM J. Matrix Anal. Appl.* **29**(1), 15–32 (2006/07)
18. Knyazev, A.V., Argentati, M.E.: Rayleigh–Ritz majorization error bounds with applications to FEM. *SIAM J. Matrix Anal. Appl.* **31**(3), 1521–1537 (2009)
19. Kürschner, P., Freitag, M.A.: Inexact methods for the low rank solution to large scale Lyapunov equations. *BIT* **60**(4), 1221–1259 (2020)
20. Lasiecka, I., Triggiani, R.: Control theory for partial differential equations: continuous and approximation theories. I. *Encyclopedia of Mathematics and its Applications*, vol. 74. Cambridge University Press, Cambridge (2000). Abstract parabolic systems
21. Målqvist, A., Persson, A., Stillfjord, T.: Multiscale differential Riccati equations for linear quadratic regulator problems. *SIAM J. Sci. Comput.* **40**(4), A2406–A2426 (2018). <https://doi.org/10.1137/17M1134500>
22. Mehrmann, V., Stykel, T.: Balanced truncation model reduction for large-scale systems in descriptor form. In: Benner, P., Sorensen, D.C., Mehrmann, V. (eds.) *Dimension Reduction of Large-Scale Systems*, pp. 83–115. Springer, Berlin (2005)
23. Melenk, J.M.: hp -finite element methods for singular perturbations, vol. 1796. Springer, Berlin (2002)
24. Nelson, E.: Analytic vectors. *Ann. Math.* **2**(70), 572–615 (1959)
25. Neymeyr, K.: A posteriori error estimation for elliptic eigenproblems. *Numer. Linear Algebra Appl.* **9**(4), 263–279 (2002)
26. Opmeer, M.R.: Decay of Hankel singular values of analytic control systems. *Syst. Control Lett.* **59**(10), 635–638 (2010)
27. Opmeer, M.R.: Decay of singular values of the Gramians of infinite-dimensional systems. In: 2015 European Control Conference (ECC), pp. 1183–1188 (2015). <https://doi.org/10.1109/ECC.2015.7330700>
28. Osborn, J.E.: Spectral approximation for compact operators. *Math. Comput.* **29**, 712–725 (1975)
29. Salamon, D.: Infinite-dimensional linear systems with unbounded control and observation: a functional analytic approach. *Trans. Am. Math. Soc.* **300**(2), 383–431 (1987). <https://doi.org/10.2307/2000351>
30. Schmüdgen, K.: *Unbounded Self-Adjoint Operators on Hilbert Space*, vol. 265. Springer, Dordrecht (2012)
31. Schwab, C.: *p - and hp -Finite Element Methods*. Oxford University Press, Oxford (1998)
32. Simoncini, V.: A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput.* **29**(3), 1268–1288 (2007). <https://doi.org/10.1137/06066120X>
33. Stillfjord, T.: Singular value decay of operator-valued differential Lyapunov and Riccati equations. *SIAM J. Control Optim.* **56**(5), 3598–3618 (2018). <https://doi.org/10.1137/18M1178815>
34. Szabo, B., Babuska, I.: *Finite Element Analysis*. Wiley, Hoboken (1991)
35. Wedin, P.A.: On angles between subspaces of a finite dimensional inner product space. *Matrix pencils, Proc. Conf., Pite Havsbad/Swed. 1982, Lect. Notes Math.* 973, pp. 263–285 (1983)
36. Weyl, H.: Über beschränkte quadratische Formen, deren Differenz vollstetig ist. *Rend. Circ. Mat. Palermo* **27**, 373–392, 402 (1909)
37. Zhu, P., Argentati, M.E., Knyazev, A.V.: Bounds for the Rayleigh quotient and the spectrum of self-adjoint operators. *SIAM J. Matrix Anal. Appl.* **34**(1), 244–256 (2013)