On the convergence analysis of one-shot inversion methods

Marcella Bonazzoli, Houssem Haddar, Tuan Anh Vu

On the convergence analysis of one-shot inversion methods

Marcella Bonazzoli[1], Houssem Haddar[1], and Tuan-Anh Vu[1,2]

[1]Inria, UMA, ENSTA Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France
[2]Institute of Mathematics, Vietnam Academy of Science and Technology, Vietnam

Abstract

When an inverse problem is solved by a gradient-based optimization algorithm, the corresponding forward and adjoint problems, which are introduced to compute the gradient, can be also solved iteratively. The idea of iterating at the same time on the inverse problem unknown and on the forward and adjoint problem solutions yields to the concept of one-shot inversion methods. We are especially interested in the case where the inner iterations for the direct and adjoint problems are incomplete, that is, stopped before achieving a high accuracy on their solutions. Here, we focus on general linear inverse problems and generic fixed-point iterations for the associated forward problem. We analyze variants of the so-called multi-step one-shot methods, in particular semi-implicit schemes with a regularization parameter. We establish sufficient conditions on the descent step for convergence, by studying the eigenvalues of the block matrix of the coupled iterations. Several numerical experiments are provided to illustrate the convergence of these methods in comparison with the classical gradient descent, where the forward and adjoint problems are solved exactly by a direct solver instead. We observe that very few inner iterations are enough to guarantee good convergence of the inversion algorithm, even in the presence of noisy data.

Keywords: inverse problems, one-shot methods, convergence analysis, parameter identification

1 Introduction

For large-scale inverse problems, which often arise in real life applications, the solution of the corresponding forward and adjoint problems is generally computed using an iterative solver, such as preconditioned fixed point or Krylov subspace methods, rather than exactly by a direct solver, such as LU-type solvers (see e.g. [22, 1]). Indeed, the corresponding linear systems could be too large to be handled with direct solvers because of their high memory requirement. In addition, iterative solvers are easier to parallelize on many cores for time speed-up. By coupling the iterative solver with a gradient-based optimization iteration, the idea of *one-step one-shot methods* is to iterate at the same time on the forward problem solution (the state variable), the adjoint problem solution (the adjoint state) and on the inverse problem unknown (the parameter or design variable). If two or more inner iterations are performed on the state and adjoint state before updating the parameter (by starting from the previous iterates as initial guess for the state and adjoint state), we speak of *multi-step one-shot methods*. Our goal is to rigorously analyze the convergence of such inversion methods. In particular, we are interested in those schemes where the inner iterations on the direct and adjoint problems are incomplete, i.e. stopped before achieving convergence. Indeed, solving the forward and adjoint problems exactly by direct solvers or very accurately by iterative solvers could be very time-consuming with little improvement in the accuracy of the inverse problem solution.

The concept of one-shot methods was first introduced by Ta'asan [19] for optimal control problems. Based on this idea, a variety of related methods, such as the all-at-once methods, where

the state equation is included in the misfit functional, were developed for aerodynamic shape optimization, see for instance [20, 18, 12, 17, 16] and the literature review in the introduction of [17]. All-at-once approaches to inverse problems for parameter identification were studied in, e.g., [9, 3, 14]. An alternative method, called Wavefield Reconstruction Inversion (WRI), was introduced for seismic imaging in [23], as an improvement of the classical Full Waveform Inversion (FWI) [21]. WRI is a penalty method which combines the advantages of the all-at-once approach with those of the reduced approach (where the state equation represents a constraint and is enforced at each iteration, as in FWI), and was extended to more general inverse problems in [24].

Few convergence proofs, especially for the multi-step one-shot methods, are available in the literature. In particular, for non-linear design optimization problems, Griewank [7] proposed a version of one-step one-shot methods where a Hessian-based preconditioner is used in the design variable iteration. The author proved conditions to ensure that the real eigenvalues of the Jacobian of the coupled iterations are smaller than 1, but these are just necessary and not sufficient conditions to exclude real eigenvalues smaller than -1. In addition, no condition to also bound complex eigenvalues below 1 in modulus was found, and multi-step methods were not investigated. In [10, 11, 5] an exact penalty function of doubly augmented Lagrangian type was introduced to coordinate the coupled iterations, and global convergence of the proposed optimization approach was proved under some assumptions. This particular one-step one-shot approach was later extended to time-dependent problems in [8].

In this work, we consider variants of multi-step one-shot methods where the forward and adjoint problems are solved using fixed point iterations and the inverse problem is solved using gradient descent methods. In particular, we analyze semi-implicit schemes with a regularization parameter. This is a preparatory work where we focus on (discretized) linear inverse problems. Note that the present analysis in the linear case implies also local convergence in the non-linear case. The only basic assumptions we require are the uniqueness of the inverse problem solution and the convergence of the fixed point iteration for the forward problem. To analyze the convergence of the coupled iterations we study the real and complex eigenvalues of the block iteration matrix. We prove that if the descent step is small enough, then the considered one-shot methods converge. Moreover, the upper bounds for the descent step in this sufficient condition are explicit in the number of inner iterations, in the norms of the operators involved in the problem, and in the regularization parameter. Note that previously in our research report [2] we studied (shifted) explicit schemes with no regularization, while here we include a regularized cost functional and analyze semi-implicit schemes. Moreover, for the particular scalar case, in [2] we established sufficient and also necessary convergence conditions on the descent step.

This paper is structured as follows. In Section 2, we introduce the principle of multi-step one-shot methods and define two variants of these algorithms. Since their analysis is similar, we shall focus on one of them, namely the semi-implicit scheme. Then, in Section 3, respectively Section 4, we analyze the convergence of one-step one-shot methods, respectively multi-step one-shot methods: first, we establish eigenvalue equations for the block matrices of the coupled iterations, then we derive sufficient convergence conditions on the descent step by studying the location of the eigenvalues in the complex plane. Finally, in Section 5 we do several numerical tests on the performance of the different algorithms on a toy 2D Helmholtz inverse problem. These tests are carried out in two cases. In the first case, the measurements are noise-free and the parameter that we desire to reconstruct is discretized in a low-dimensional space, while in the second case, the measurements are affected by noise and the reconstructed parameter is discretized in a higher dimensional space. In particular, we observe that very few inner iterations are enough to guarantee good convergence of the inversion algorithms, even in the presence of noisy data.

Throughout this work, $\langle \cdot, \cdot \rangle$ indicates the usual Hermitian scalar product in $\mathbb{C}^n$, that is $\langle x, y \rangle := \overline{y}^\intercal x, \forall x, y \in \mathbb{C}^n$, and $\|\cdot\|$ the vector/matrix norms induced by $\langle \cdot, \cdot \rangle$. We denote by $A^* = \overline{A}^\intercal$ the adjoint operator of a matrix $A \in \mathbb{C}^{m \times n}$, and likewise by $z^* = \overline{z}$ the conjugate of a complex number

z. The identity matrix is always denoted by I, whose size is understood from context. Finally, for a matrix $T \in \mathbb{C}^{n\times n}$, we denote by $\rho(T)$ the spectral radius of T, and when $\rho(T) < 1$, we define

$$s(T) \coloneqq \sup_{z\in\mathbb{C},|z|\geq 1} \|(I - T/z)^{-1}\|,$$

which is further studied in Appendix A.

2 Multi-step one-shot inversion methods

We focus on (discretized) linear inverse problems, which correspond to a *direct (or forward) problem* of the form: seek $u \equiv u(\sigma)$ such that

$$u = Bu + M\sigma + F \tag{1}$$

where $u \in \mathbb{R}^{n_u}$, $\sigma \in \mathbb{R}^{n_\sigma}$, $B \in \mathbb{R}^{n_u\times n_u}$, $M \in \mathbb{R}^{n_u\times n_\sigma}$ and $F \in \mathbb{R}^{n_u}$. Here $I - B$ is the invertible matrix associated with the direct problem (e.g. obtained after discretization of a PDE model and applying a preconditioner to the linear system), with parameter σ. Equation (1) is also referred to as the *state equation* and u is the *state variable*. Given σ, we assume that one solves for u by a fixed point iteration

$$u_{\ell+1} = Bu_\ell + M\sigma + F, \quad \ell = 0, 1, \dots \tag{2}$$

We indeed assume $\rho(B) < 1$ so that the fixed point iteration (2) converges for any initial guess u_0 (see e.g. [6, Theorem 2.1.1]). Measuring $g = Hu(\sigma)$, where $H \in \mathbb{R}^{n_g\times n_u}$, we consider the *linear inverse problem* of retrieving σ from the knowledge of g. Let us set $A \coloneqq H(I-B)^{-1}M$. The inverse problem can be synthetically written as $A\sigma = g - H(I - B)^{-1}F$, which amounts to inverting the ill-conditioned matrix A. We shall assume in the following the uniqueness of the solution for this inverse problem, which is equivalent to the injectivity of A. In summary, we set

$$\begin{array}{ll} \text{direct problem:} & u = Bu + M\sigma + F, \\ \text{inverse problem:} & \text{measure } g = Hu(\sigma), \text{ retrieve } \sigma \end{array} \tag{3}$$

with the assumptions:

$$\rho(B) < 1, \quad H(I - B)^{-1}M \text{ is injective.} \tag{4}$$

Remark 2.1. Considering real-valued matrices B and M is not a restrictive assumption. Indeed, the case of complex-valued matrices can be rewritten as a system of real-valued equations by doubling the size of the linear system (see [2, Section 5]).

To solve the inverse problem we write its regularized least squares formulation: given σ^{ex} the exact solution of the inverse problem and $g \coloneqq Hu(\sigma^{\text{ex}})$ (g can also be a noisy version of $Hu(\sigma^{\text{ex}})$),

$$\sigma^{\text{ex}} = \text{argmin}_{\sigma\in\mathbb{R}^{n_\sigma}} J(\sigma) \quad \text{where } J(\sigma) \coloneqq \frac{1}{2}\|Hu(\sigma) - g\|^2 + \frac{\alpha}{2}\|\sigma\|^2, \ \alpha \geq 0. \tag{5}$$

Using the classical Lagrangian technique with real scalar products, we introduce the *adjoint state* $p \equiv p(\sigma)$, which is the solution of

$$p = B^*p + H^*(Hu - g)$$

and allows us to compute the gradient of the cost functional as

$$\nabla J(\sigma) = M^*p(\sigma) + \alpha\sigma.$$

The classical gradient descent algorithm then reads

$$\textbf{usual gradient descent:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^*p^n - \tau\alpha\sigma^n, \\ u^n = Bu^n + M\sigma^n + F, \\ p^n = B^*p^n + H^*(Hu^n - g), \end{cases} \tag{6}$$

where $\tau > 0$ is the descent step size, and the state and adjoint state equations are solved exactly at each iteration step for σ. Notice that when $F = 0$, (6) is equivalent to $\sigma^{n+1} = \sigma^n - \tau A^*(A\sigma^n - g) - \tau\alpha\sigma^n$. One can also consider a slightly different version where an implicit scheme is applied to the regularization term leading to the following semi-implicit gradient scheme

$$\textbf{semi-implicit gradient descent:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^{n+1}, \\ u^n = Bu^n + M\sigma^n + F, \\ p^n = B^* p^n + H^*(Hu^n - g). \end{cases} \tag{7}$$

It can be shown that both algorithms converge for sufficiently small $\tau > 0$: for any initial guess, (6) converges if and only if $\tau < \frac{2}{\rho(A^*A)+\alpha}$ and (7) converges if and only if $(\rho(A^*A) - \alpha)\tau < 2$. These results indicate in particular that we gain more stability with the semi-implicit scheme.

We are interested in methods where the direct and adjoint problems are rather solved iteratively as in (2), and where we iterate at the same time on the forward problem solution and the inverse problem unknown: such methods are called *one-shot methods*. More precisely, we are interested in two variants of *multi-step one-shot methods*, defined as follows. Let n be the index of the (outer) iteration on σ. We update $\sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^n$ as in gradient descent methods (or respectively, $\sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^{n+1}$ as in semi-implicit gradient descent methods), but the state and adjoint state equations are now solved by a fixed point iteration method, using just k *inner iterations* and as initial guess we naturally choose the information from the previous (outer) step. We then get the two following variants of multi-step one-shot algorithms

$$k\textbf{-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^n, \\ u_0^{n+1} = u^n, p_0^{n+1} = p^n. \text{ Repeat for } \ell = 0, 1, \dots, k-1, \\ \quad \left| \begin{array}{l} u_{\ell+1}^{n+1} = Bu_\ell^{n+1} + M\sigma^{n+1} + F, \\ p_{\ell+1}^{n+1} = B^* p_\ell^{n+1} + H^*(Hu_\ell^{n+1} - g), \end{array} \right. \\ u^{n+1} := u_k^{n+1}, p^{n+1} := p_k^{n+1} \end{cases} \tag{8}$$

and

$$\textbf{semi-implicit } k\textbf{-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^{n+1}, \\ u_0^{n+1} = u^n, p_0^{n+1} = p^n. \text{ Repeat for } \ell = 0, 1, \dots, k-1, \\ \quad \left| \begin{array}{l} u_{\ell+1}^{n+1} = Bu_\ell^{n+1} + M\sigma^{n+1} + F, \\ p_{\ell+1}^{n+1} = B^* p_\ell^{n+1} + H^*(Hu_\ell^{n+1} - g), \end{array} \right. \\ u^{n+1} := u_k^{n+1}, p^{n+1} := p_k^{n+1}. \end{cases} \tag{9}$$

In particular, when $k = 1$, we obtain the two following algorithms

$$\textbf{one-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^n, \\ u^{n+1} = Bu^n + M\sigma^{n+1} + F, \\ p^{n+1} = B^* p^n + H^*(Hu^n - g) \end{cases} \tag{10}$$

and

$$\textbf{semi-implicit one-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n - \tau\alpha\sigma^{n+1}, \\ u^{n+1} = Bu^n + M\sigma^{n+1} + F, \\ p^{n+1} = B^* p^n + H^*(Hu^n - g). \end{cases} \tag{11}$$

Note that when $k \to \infty$, the k-step one-shot method (8) formally converges to the usual gradient descent (6), while the semi-implicit k-step one-shot method (9) formally converges to the semi-implicit gradient descent (7). Since the analysis of the two schemes (8) and (9) can be done following

similar arguments, we choose to concentrate on only one of them, namely the semi-implicit scheme. We refer to [2] for the analysis in the case $\alpha = 0$ (for which the two schemes coincide).

We first analyze the one-step one-shot method ($k = 1$) in Section 3 and then the multi-step one-shot method ($k \geq 1$) in Section 4.

3 Convergence of the one-step one-shot method ($k = 1$)

3.1 Block iteration matrix and eigenvalue equation

To analyze the convergence of the semi-implicit one-step one-shot method (11), we first express $(\sigma^{n+1}, u^{n+1}, p^{n+1})$ in terms of (σ^n, u^n, p^n), by inserting the expression for σ^{n+1} into the iteration for u^{n+1} in (11), so that system (11) is rewritten as

$$\begin{cases} \sigma^{n+1} = \frac{1}{1+\tau\alpha}\sigma^n - \frac{\tau}{1+\tau\alpha}M^*p^n, \\ u^{n+1} = Bu^n + \frac{1}{1+\tau\alpha}M\sigma^n - \frac{\tau}{1+\tau\alpha}MM^*p^n + F, \\ p^{n+1} = B^*p^n + H^*Hu^n - H^*g. \end{cases} \tag{12}$$

Now, we consider the errors $(\sigma^n - \sigma^{\mathrm{ex}}, u^n - u(\sigma^{\mathrm{ex}}), p^n - p(\sigma^{\mathrm{ex}}))$ with respect to the exact solution at the n-th iteration, and, by abuse of notation, we designate them by (σ^n, u^n, p^n). We obtain that these errors satisfy

$$\begin{cases} \sigma^{n+1} = \frac{1}{1+\tau\alpha}\sigma^n - \frac{\tau}{1+\tau\alpha}M^*p^n, \\ u^{n+1} = Bu^n + \frac{1}{1+\tau\alpha}M\sigma^n - \frac{\tau}{1+\tau\alpha}MM^*p^n, \\ p^{n+1} = B^*p^n + H^*Hu^n, \end{cases} \tag{13}$$

or equivalently, by putting in evidence the block iteration matrix

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} B^* & H^*H & 0 \\ -\frac{\tau}{1+\tau\alpha}MM^* & B & \frac{1}{1+\tau\alpha}M \\ -\frac{\tau}{1+\tau\alpha}M^* & 0 & \frac{1}{1+\tau\alpha}I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix}. \tag{14}$$

Now recall that a fixed point iteration converges if and only if the spectral radius of its iteration matrix is less than 1. Therefore in the following proposition we establish an eigenvalue equation for the iteration matrix of the semi-implicit one-step one-shot method.

Proposition 3.1. *Assume that $\lambda \in \mathbb{C}, |\lambda| \geq 1$ is an eigenvalue of the iteration matrix in (14). If $\lambda \in \mathbb{C}$, $\lambda \notin \mathrm{Spec}(B)$ then $\exists\, y \in \mathbb{C}^{n_\sigma}, \|y\| = 1$ such that:*

$$(1+\tau\alpha)\lambda - 1 + \tau\lambda\langle M^*(\lambda I - B^*)^{-1}H^*H(\lambda I - B)^{-1}My, y\rangle = 0. \tag{15}$$

In particular, $\lambda = 1$ is not an eigenvalue of the iteration matrix.

Proof. Since $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (14), there exists a non-zero vector $(\tilde{p}, \tilde{u}, y) \in \mathbb{C}^{n_u+n_u+n_\sigma}$ such that

$$\begin{cases} \lambda y = \frac{1}{1+\tau\alpha}y - \frac{\tau}{1+\tau\alpha}M^*\tilde{p}, \\ \lambda\tilde{u} = B\tilde{u} + \frac{1}{1+\tau\alpha}My - \frac{\tau}{1+\tau\alpha}MM^*\tilde{p}, \\ \lambda\tilde{p} = B^*\tilde{p} + H^*H\tilde{u}. \end{cases} \tag{16}$$

By the second equation in (16),

$$\tilde{u} = \frac{1}{1+\tau\alpha}(\lambda I - B)^{-1}My - \frac{\tau}{1+\tau\alpha}(\lambda I - B)^{-1}MM^*\tilde{p}.$$

Inserting this result into the third equation we obtain

$$\lambda\tilde{p} = \left(B^* - \frac{\tau}{1+\tau\alpha}H^*H(\lambda I - B)^{-1}MM^*\right)\tilde{p} + \frac{1}{1+\tau\alpha}H^*H(\lambda I - B)^{-1}My,$$

or equivalently, with the short notation $D = (\lambda I - B^*)^{-1}H^*H(\lambda I - B)^{-1}$,

$$\tilde{p} + \frac{\tau}{1+\tau\alpha}DMM^*\tilde{p} = \frac{1}{1+\tau\alpha}DMy.$$

Multiplying both sides of this equation with M^* gives

$$\left(I + \frac{\tau}{1+\tau\alpha}M^*DM\right)M^*\tilde{p} = \frac{1}{1+\tau\alpha}M^*DMy.$$

By the first equation in (16), $M^*\tilde{p} = \dfrac{1-(1+\tau\alpha)\lambda}{\tau}y$, therefore

$$\frac{1-(1+\tau\alpha)\lambda}{\tau}\left(I + \frac{\tau}{1+\tau\alpha}M^*DM\right)y = \frac{1}{1+\tau\alpha}M^*DMy,$$

or equivalently, replacing D by its expression,

$$[(1+\alpha\tau)\lambda - 1]\,y + \tau\lambda M^*(\lambda I - B^*)^{-1}H^*H(\lambda I - B)^{-1}My = 0. \tag{17}$$

We prove that $y \neq 0$. Indeed if $y = 0$ then $M^*\tilde{p} = \dfrac{1-(1+\tau\alpha)\lambda}{\tau}y = 0$ and

$$\tilde{u} = \frac{1}{1+\tau\alpha}(\lambda I - B)^{-1}My - \frac{\tau}{1+\tau\alpha}(\lambda I - B)^{-1}MM^*\tilde{p} = 0.$$

Inserting these results into the third equation in (16) we obtain $\lambda\tilde{p} = B^*\tilde{p}$, which immediately implies $\tilde{p} = 0$ and gives a contradiction. Finally, by taking the scalar product of (17) with y, then normalizing by $\|y\|$, we obtain (15).
(ii) Now assume that $\lambda = 1$ is an eigenvalue of the iteration matrix, then (15) yields

$$\alpha + \|H(I - B)^{-1}My\|^2 = 0,$$

which cannot be true due to the injectivity of $H(I - B)^{-1}M$. □

In the following sections we will show that, for sufficiently small τ, equation (15) cannot hold if $|\lambda| \geq 1$, thus algorithm (11) converges. It is convenient to rewrite (15) as

$$(1+\tau\alpha)\lambda^2 - \lambda + \tau\langle M^*\left(I - B^*/\lambda\right)^{-1}H^*H\left(I - B/\lambda\right)^{-1}My, y\rangle = 0. \tag{18}$$

For the analysis we use auxiliary technical results proved in Appendix A.

First, we study separately the very particular case where $B = 0$.

Proposition 3.2. *When $B = 0$, the eigenvalue equation (18) cannot hold for $\lambda \in \mathbb{C}, |\lambda| \geq 1$ if $\tau > 0$ and*

$$(\|H\|^2\|M\|^2 - \alpha)\tau < 1.$$

Proof. When $B = 0$, equation (18) becomes $(1+\tau\alpha)\lambda^2 - \lambda + \tau\|HMy\|^2 = 0$. Then, the conclusion can be obtained by computing the roots of this second order polynomial. The result can also be immediately obtained by applying the following lemma, which is mainly based on Marden's works [15]. □

Lemma 3.3. *Let $a_0, a_1 \in \mathbb{R}$, then all roots of $\mathcal{P}(z) = a_0 + a_1 z + z^2$ stay (strictly) inside the unit circle of the complex plane if and only if*

$$|a_0| < 1 \quad \textit{and} \quad (a_0 - a_1 + 1)(a_0 + a_1 + 1) > 0.$$

The proof of this lemma can be deduced from Appendix C of [2].

3.2 Location of the eigenvalues in the complex plane

We now turn our attention to the eigenvalues λ for which (18) holds. We would like to derive conditions on the descent step τ such that all the eigenvalues lie inside the unit circle which would ensure the convergence for the scheme (11). We start with the simple case of real eigenvalues.

Proposition 3.4 (Real eigenvalues)**.** *Equation* (18) *admits no solution* $\lambda \in \mathbb{R}, \lambda \neq 1, |\lambda| \geq 1$ *for all* $\tau > 0$.

Proof. When $\lambda \in \mathbb{R}\backslash\{0\}$ equation (18) becomes

$$(1+\tau\alpha)\lambda^2 - \lambda + \tau\|H(I - B/\lambda)^{-1}My\|^2 = 0.$$

If $\lambda \in \mathbb{R}, \lambda \neq 1, |\lambda| \geq 1$ then $(1+\tau\alpha)\lambda^2 - \lambda \geq \lambda^2 - \lambda > 0$, thus the left-hand side of the above equation is positive for any $\tau > 0$. □

For the general case of complex eigenvalues, the study is much more complicated and technical. The following proposition summarizes the results we obtained.

Proposition 3.5 (Complex eigenvalues)**.** *If* $0 \neq \rho(B) < 1$, *there exists* $\tau > 0$ *sufficiently small such that equation* (18) *admits no solution* $\lambda \in \mathbb{C}\backslash\mathbb{R}, |\lambda| \geq 1$. *In particular, if* $\|B\| < 1$, *given any* $\delta_0 > 0$ *and* $0 < \theta_0 \leq \frac{\pi}{4}$, *one can choose*

$$\tau < \min_{1\leq i\leq 3} \left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^4}\varphi_i(\|B\|) + C_i\alpha \right)^{-1},$$

where

$$\varphi_1(b) := 4b^2, \quad \varphi_2(b) := \frac{1}{2\sin\frac{\theta_0}{2}}(1+b)^2(1-b)^2, \quad \varphi_3(b) := \frac{2c}{\delta_0}b^2,$$

$$C_1 := \sqrt{2} - 1, \quad C_2 := \sqrt{2} + \frac{1}{2\sin\frac{\theta_0}{2}} - 1, \quad C_3 := \frac{\sqrt{c}}{\delta_0} - 1 \quad \text{and} \quad c := \frac{1 + 2\delta_0\sin\frac{3\theta_0}{2} + \delta_0^2}{\cos^2\frac{3\theta_0}{2}}.$$

Proof. **Step 1. Rewrite equation** (18) **by separating real and imaginary parts.**

Let $\lambda = R(\cos\theta + \mathrm{i}\sin\theta)$ in polar form where $R = |\lambda| \geq 1$ and $\theta \in (-\pi, \pi)$, $\theta \neq 0$. Write $1/\lambda = r(\cos\phi + \mathrm{i}\sin\phi)$ in polar form where $r = 1/|\lambda| = 1/R \leq 1$ and $\phi = -\theta \in (-\pi, \pi)$. By Lemma A.3, we have

$$\left(I - \frac{B}{\lambda}\right)^{-1} = P(\lambda) + \mathrm{i}Q(\lambda), \quad \left(I - \frac{B^*}{\lambda}\right)^{-1} = P(\lambda)^* + \mathrm{i}Q(\lambda)^*$$

where $P(\lambda)$ and $Q(\lambda)$ are $\mathbb{C}^{n_u\times n_u}$ matrices that satisfy the following bounds in the case $\|B\| < 1$ for all $|\lambda| \geq 1$:

$$\|P(\lambda)\| \leq p := \frac{1}{1-\|B\|}, \tag{19}$$

$$\|Q(\lambda)\| \leq q_1 := \frac{\|B\|}{1-\|B\|} \quad \text{and} \quad \|Q(\lambda)\| \leq |\sin\theta| q_2 \quad \text{with} \quad q_2 := \frac{\|B\|}{(1-\|B\|)^2}. \tag{20}$$

These bounds still hold in the case $0 \neq \rho(B) < 1$ with

$$p = (1+\|B\|)s(B)^2, \quad q_1 = \|B\|s(B)^2 \quad \text{and} \quad q_2 = \|B\|s(B)^2. \tag{21}$$

To simplify the notation, we will not explicitly write the dependence of P and Q on λ. Now we rewrite (18) as

$$(1+\tau\alpha)(\lambda^2 - \lambda) + \tau\alpha(\lambda - 1) + \tau\alpha + \tau G(P^* + \mathrm{i}Q^*, P + \mathrm{i}Q) = 0 \tag{22}$$

where

$$G(X,Y) \coloneqq \langle M^*XH^*HYMy, y\rangle \in \mathbb{C}, \quad X,Y \in \mathbb{C}^{n_u\times n_u}.$$

Notice that G is a bilinear form and $G(X,Y) = G(Y^*,X^*)^*$ so that $G(X,Y)+G(X^*,Y^*)$ is real. With these properties of G, we expand (22) and take its real and imaginary parts, which yields

$$(1+\tau\alpha)\Re(\lambda^2-\lambda)+\tau\alpha\Re(\lambda-1)+\tau\alpha+\tau[G(P^*,P)-G(Q^*,Q)]=0, \tag{23}$$

and

$$(1+\tau\alpha)\Im(\lambda^2-\lambda)+\tau\alpha\Im(\lambda-1)+\tau[G(P^*,Q)+G(Q^*,P)]=0. \tag{24}$$

Step 2. Use a suitable combination of equations (23) and (24).

Let $\gamma\in\mathbb{R}$. Multiplying equation (24) with γ then summing it with equation (23), we obtain:

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma\Im(\lambda-1)]\\ +\tau\alpha+\tau[G(P^*,P)-G(Q^*,Q)+\gamma G(P^*,Q)+\gamma G(Q^*,P)]=0,\end{aligned}$$

or equivalently,

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma\Im(\lambda-1)]\\ +\tau\alpha+\tau G(P^*+\gamma Q^*,P+\gamma Q)-(1+\gamma^2)\tau G(Q^*,Q)=0.\end{aligned} \tag{25}$$

Now we consider four cases for λ as in Lemma A.4:

- *Case 1.* $\Re(\lambda^2-\lambda)\geq 0$;
- *Case 2.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in[\theta_0,\pi-\theta_0]\cup[-\pi+\theta_0,-\theta_0]$ for fixed $0<\theta_0\leq\frac{\pi}{4}$;
- *Case 3.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in(-\theta_0,\theta_0)$ for fixed $0<\theta_0\leq\frac{\pi}{4}$;
- *Case 4.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in(\pi-\theta_0,\pi)\cup(-\pi,-\pi+\theta_0)$ for fixed $0<\theta_0\leq\frac{\pi}{4}$.

Cases 1, 2 and 3 are respectively treated in Lemmas 3.6, 3.7 and 3.8 below. Notice that case 4 corresponds to an empty set, according to the Lemma A.4 (iv). The statement of the proposition easily follows from the combination of Lemmas 3.6, 3.7 and 3.8. □

In the lemmas below we shall make use of the following obvious property where $\|y\|=1$:

$$0\leq G(X^*,X)=\|HXMy\|^2\leq(\|H\|\|M\|\|X\|)^2, \quad \forall X\in\mathbb{C}^{n_u\times n_u}. \tag{26}$$

Lemma 3.6 (Case 1)**.** *Let $0\neq\rho(B)<1$ and let $|\lambda|\geq 1$ with $\Re(\lambda^2-\lambda)\geq 0$. Then equation (18) cannot hold if one chooses*

$$\tau<\left(\|H\|^2\|M\|^2 s(B)^4\cdot 4\|B\|^2+(\sqrt{2}-1)\alpha\right)^{-1}.$$

Moreover, if $\|B\|<1$, the result is also true if

$$\tau<\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^4}\cdot 4\|B\|^2+(\sqrt{2}-1)\alpha\right)^{-1}.$$

Proof. Define

$$\gamma_1=\gamma_1(\lambda)\coloneqq\begin{cases}1 & \text{if } \Im(\lambda^2-\lambda)\geq 0,\\ -1 & \text{if } \Im(\lambda^2-\lambda)<0\end{cases}$$

as in Lemma A.4 (i). Writing (25) for $\gamma = \gamma_1$ and using $\gamma_1^2 = 1$ we obtain

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)] \\ +\tau\alpha+\tau G(P^*+\gamma_1 Q^*, P+\gamma_1 Q)-2\tau G(Q^*,Q)=0. \quad (27)$$

Since $G(P^*+\gamma_1 Q^*, P+\gamma_1 Q)\geq 0$ by (26) and $\tau\alpha\geq 0$, then the left-hand side of (27) is positive if τ satisfies

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)]-\tau\alpha|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|-2\tau G(Q^*,Q)>0,$$

or equivalently,

$$1+\tau\alpha-\tau\alpha\frac{|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}-2\tau\frac{G(Q^*,Q)}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}>0. \quad (28)$$

Notice that the choice of γ_1 ensures $\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)>0$. By Lemma A.4 (i) we have

$$\frac{|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}\leq\frac{\sqrt{1+\gamma_1^2}|\lambda-1|}{|\lambda(\lambda-1)|}=\frac{\sqrt{2}}{|\lambda|}\leq\sqrt{2}.$$

Using again Lemma A.4 (i) and (26), we have

$$\frac{G(Q^*,Q)}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}\leq\frac{(\|H\|\|M\|\sin\theta|q_2)^2}{2|\sin(\theta/2)|}=2\left|\sin\frac{\theta}{2}\right|\cos^2\frac{\theta}{2}\|H\|^2\|M\|^2q_2^2 \\ \leq 2\|H\|^2\|M\|^2q_2^2.$$

Inserting the two previous inequalities in (28) gives the desired results using definitions (20) and (21) of q_2.

□

Lemma 3.7 (Case 2). *Let $0\neq\rho(B)<1$ and let $|\lambda|\geq 1$, $\Re(\lambda^2-\lambda)<0$, $\theta\in[\theta_0,\pi-\theta_0]\cup[-\pi+\theta_0,-\theta_0]$ for given $0<\theta_0\leq\frac{\pi}{4}$. Then equation (18) cannot hold if one chooses*

$$\tau<\left(\|H\|^2\|M\|^2s(B)^4\cdot\frac{(1+2\|B\|)^2}{2\sin\frac{\theta_0}{2}}+\left(\sqrt{2}+\frac{1}{2\sin\frac{\theta_0}{2}}-1\right)\alpha\right)^{-1}.$$

Moreover, if $\|B\|<1$, the result is also true if

$$\tau<\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^2}\cdot\frac{(1+\|B\|)^2}{2\sin\frac{\theta_0}{2}}+\left(\sqrt{2}+\frac{1}{2\sin\frac{\theta_0}{2}}-1\right)\alpha\right)^{-1}.$$

Proof. Define

$$\gamma_2=\gamma_2(\lambda):=\begin{cases}-1 & \text{if } \Im(\lambda^2-\lambda)\geq 0,\\ 1 & \text{if } \Im(\lambda^2-\lambda)<0\end{cases}$$

as in Lemma A.4 (ii). Writing (25) for $\gamma=\gamma_2$ and using $\gamma_2^2=1$ we obtain

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma_2\Im(\lambda-1)] \\ +\tau\alpha+\tau G(P^*+\gamma_2 Q^*,P+\gamma_2 Q)-2\tau G(Q^*,Q)=0. \quad (29)$$

Since $G(Q^*,Q)\geq 0$ by (26), the left-hand side of (29) is negative if τ satisfies

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)]+\tau\alpha|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)| \\ +\tau\alpha+\tau G(P^*+\gamma_2 Q^*,P+\gamma_2 Q)<0,$$

or equivalently,

$$-1-\tau\alpha+\tau\alpha\frac{|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}+\frac{\tau\alpha}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} + \tau\frac{G(P^*+\gamma_2Q^*,P+\gamma_2Q)}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}<0. \quad (30)$$

Notice that the choice of γ_2 ensures $\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)<0$. In the following we derive upper bounds independent from λ for the terms appearing with the negative sign in (30). By Lemma A.4 (ii) we have

$$\frac{|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}\leq\frac{\sqrt{1+\gamma_2^2}|\lambda-1|}{|\lambda(\lambda-1)|}=\frac{\sqrt{2}}{|\lambda|}\leq\sqrt{2}$$

and

$$\frac{1}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}\leq\frac{1}{2\sin\frac{\theta_0}{2}}.$$

Using again Lemma A.4 (ii) and (26), we have

$$\frac{G(P^*+\gamma_2Q^*,P+\gamma_2Q)}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}\leq\frac{\|H\|^2\|M\|^2(p+q_1)^2}{2\sin\frac{\theta_0}{2}}.$$

Inserting these previous inequalities in (30) gives the desired results using definitions (19), (20) and (21) of p and q_1. □

Lemma 3.8 (Case 3). *Let $0\neq\rho(B)<1$ and let $|\lambda|\geq1$, $\Re(\lambda^2-\lambda)<0$, $\theta\in(-\theta_0,\theta_0)$ for given $0<\theta_0\leq\frac{\pi}{4}$. For any $\delta_0>0$, equation (18) cannot hold if one chooses*

$$\tau<\left(\|H\|^2\|M\|^2s(B)^4\cdot\frac{2c}{\delta_0}\|B\|^2+\left(\frac{\sqrt{c}}{\delta_0}-1\right)\alpha\right)^{-1}$$

where $c=c(\theta_0,\delta_0):=\left(1+2\delta_0\sin\frac{3\theta_0}{2}+\delta_0^2\right)/\cos^2\frac{3\theta_0}{2}$. Moreover, if $0<\|B\|<1$, the result is also true if

$$\tau<\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^{-4}}\cdot\frac{2c}{\delta_0}\|B\|^2+\left(\frac{\sqrt{c}}{\delta_0}-1\right)\alpha\right)^{-1}.$$

Proof. Define

$$\gamma_3=\gamma_3(\mathrm{sign}(\theta)):=\begin{cases}\left(\delta_0+\sin\frac{3\theta_0}{2}\right)/\cos\frac{3\theta_0}{2} & \text{if } \theta>0,\\ -\left(\delta_0+\sin\frac{3\theta_0}{2}\right)/\cos\frac{3\theta_0}{2} & \text{if } \theta<0\end{cases}$$

as in Lemma A.4 (iii). Writing (25) for $\gamma=\gamma_3$ we obtain

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma_3\Im(\lambda-1)] + \tau\alpha+\tau G(P^*+\gamma_3Q^*,P+\gamma_3Q)-\tau(1+\gamma_3^2)G(Q^*,Q)=0. \quad (31)$$

Since $G(P^*+\gamma_3Q^*,P+\gamma_3Q)\geq0$ by (26) and $\tau\alpha\geq0$, the left-hand side of (31) is positive if τ satisfies

$$(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)]-\tau\alpha|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|-\tau(1+\gamma_3^2)G(Q^*,Q)>0,$$

or equivalently,

$$1+\tau\alpha-\tau\alpha\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)}-\tau(1+\gamma_3^2)\frac{G(Q^*,Q)}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)}>0. \quad (32)$$

Notice that the choice of γ_3 ensures $\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)>0$, also

$$1+\gamma_3^2 = 1+\frac{\left(\delta_0+\sin\frac{3\theta_0}{2}\right)^2}{\cos^2\frac{3\theta_0}{2}} = \frac{1+2\delta_0\sin\frac{3\theta_0}{2}+\delta_0^2}{\cos^2\frac{3\theta_0}{2}} =: c$$

is a constant greater than δ_0^2. By Lemma A.4 (iii) we have

$$\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} \le \frac{\sqrt{1+\gamma_3^2}}{\delta_0} = \frac{\sqrt{c}}{\delta_0}.$$

Using again Lemma A.4 (iii) and (26), we have

$$\begin{aligned}\frac{G(Q^*,Q)}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} \le \frac{(\|H\|\|M\||\sin\theta|q_2)^2}{2\delta_0|\sin(\theta/2)|} &= \frac{2}{\delta_0}\left|\sin\frac{\theta}{2}\right|\cos^2\frac{\theta}{2}\|H\|^2\|M\|^2q_2^2 \\ &\le \frac{2}{\delta_0}\|H\|^2\|M\|^2q_2^2,\end{aligned}$$

Inserting the two previous inequalities in (32) gives the desired results using definitions (20) and (21) of q_2. □

3.3 Final result ($k=1$)

Considering Proposition 3.2 (for $B=0$), Proposition 3.4 (for real eigenvalues and $B\neq 0$) and taking the bound in Proposition 3.5 (for complex eigenvalues and $B\neq 0$), we obtain a sufficient condition on the descent step τ to ensure convergence of the semi-implicit one-step one-shot method.

Theorem 3.9 (Convergence of semi-implicit one-step one-shot)**.** *Under assumption (4), the one-step one-shot method (11) converges for sufficiently small τ. In particular, for $\|B\|<1$, there exists an explicit piecewise at-most-4th-order-polynomial function $\mathcal{P}_1$ and a constant $C>0$ such that $\mathcal{P}_1>0$ on $[0,1)$ and it is enough to take*

$$\tau < \left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^4}\mathcal{P}_1(\|B\|)+C\alpha\right)^{-1}.$$

4 Convergence of the multi-step one-shot method ($k\ge 1$)

We now tackle the general case of semi-implicit multi-step one-shot methods, that is algorithm (9) with $k\ge 1$. The procedure is quite similar to the case $k=1$ but with more involved technicalities.

4.1 Block iteration matrix and eigenvalue equation

Let $k\ge 1$ be the number of inner iterations for u and p. First we express $(\sigma^{n+1},u^{n+1},p^{n+1})$ in terms of (σ^n,u^n,p^n) in a matrix form as for the case $k=1$. More precisely, the system (9) can be equivalently written as

$$\begin{cases}\sigma^{n+1} = \sigma^n-\tau M^*p^n,\\ u^{n+1} = B^ku^n+T_kM\sigma^n-\tau T_kMM^*p^n+T_kF,\\ p^{n+1} = [(B^*)^k-\tau X_kMM^*]p^n+U_ku^n+X_kM\sigma^n+X_kF-T_k^*H^*g\end{cases} \tag{33}$$

where

$$T_k \coloneqq I+B+\ldots+B^{k-1} = (I-B)^{-1}(I-B^k), \tag{34}$$

$$U_k \coloneqq (B^*)^{k-1}H^*H + (B^*)^{k-2}H^*HB + ... + H^*HB^{k-1}, \tag{35}$$

$$X_k \coloneqq \begin{cases} (B^*)^{k-2}H^*HT_1 + (B^*)^{k-3}H^*HT_2 + ... + H^*HT_{k-1} & \text{if } k \geq 2, \\ 0 & \text{if } k = 1. \end{cases} \tag{36}$$

Before analyzing recursion (33), we gather in the following lemma some useful properties of T_k, U_k and X_k.

Lemma 4.1. *(i) The matrices U_k and X_k can be rewritten as*

$$U_k = \sum_{i+j=k-1} (B^*)^i H^* H B^j \quad \text{for } k \geq 1,$$
$$X_k = \sum_{l=0}^{k-2} \sum_{i+j=l} (B^*)^i H^* H B^j = \sum_{l=1}^{k-1} U_l \quad \text{for } k \geq 2.$$

(ii) The matrices U_k and X_k are self-adjoint: $U_k^* = U_k$, $X_k^* = X_k$.

(iii) We have the relation

$$U_k T_k - X_k B^k + X_k = T_k^* H^* H T_k, \quad \forall k \geq 1. \tag{37}$$

Proof. Property (i) is easy to check from the definitions (34), (35) and (36). Property (ii) straightforwardly follows from (i).
Now we prove (iii). For $k = 1$, we have $U_1 = H^*H$, $T_1 = I$ and $X_1 = 0$, hence the identity is verified. For $k \geq 2$, we remark that $X_{k+1} = B^*X_k + H^*HT_k$ by definition (36). Thanks to (ii) we know that X_{k+1} is self-adjoint, hence $X_{k+1} = X_{k+1}^* = X_k B + T_k^* H^* H$. On the other hand, from (i) we get that $X_{k+1} = X_k + U_k$. Thus,

$$X_k + U_k = X_k B + T_k^* H^* H, \quad \text{or equivalently,} \quad U_k = X_k(B - I) + T_k^* H^* H.$$

Finally,

$$U_k T_k = X_k(B - I)T_k + T_k^* H^* H T_k = X_k(B^k - I) + T_k^* H^* H T_k.$$

□

Now, we consider the errors $(\sigma^n - \sigma^{\text{ex}}, u^n - u(\sigma^{\text{ex}}), p^n - p(\sigma^{\text{ex}}))$ with respect to the exact solution at the n-th iteration, and, by abuse of notation, we denote them by (σ^n, u^n, p^n). We obtain that the errors satisfy

$$\begin{cases} \sigma^{n+1} = \frac{1}{1+\tau\alpha}\sigma^n - \frac{\tau}{1+\tau\alpha}M^*p^n, \\ u^{n+1} = B^k u^n + \frac{1}{1+\tau\alpha}T_k M\sigma^n - \frac{\tau}{1+\tau\alpha}T_k MM^*p^n, \\ p^{n+1} = \left[(B^*)^k - \frac{\tau}{1+\tau\alpha}X_k MM^*\right]p^n + U_k u^n + \frac{1}{1+\tau\alpha}X_k M\sigma^n, \end{cases} \tag{38}$$

or equivalently, by putting in evidence the block iteration matrix

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} (B^*)^k - \frac{\tau}{1+\tau\alpha}X_k MM^* & U_k & \frac{1}{1+\tau\alpha}X_k M \\ -\frac{\tau}{1+\tau\alpha}T_k MM^* & B^k & \frac{1}{1+\tau\alpha}T_k M \\ -\frac{\tau}{1+\tau\alpha}M^* & 0 & \frac{1}{1+\tau\alpha}I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix}. \tag{39}$$

Proposition 4.2. *Assume that $\lambda \in \mathbb{C}, |\lambda| \geq 1$ is an eigenvalue of the iteration matrix in (39). If $\lambda \in \mathbb{C}$, $\lambda \notin \text{Spec}(B)$ then $\exists\, y \in \mathbb{C}^{n_\sigma}, \|y\| = 1$ such that:*

$$(1 + \tau\alpha)\lambda - 1 + \tau\lambda\langle M^*[\lambda I - (B^*)^k]^{-1}[(\lambda - 1)X_k + T_k^* H^* H T_k](\lambda I - B^k)^{-1}My, y\rangle = 0. \tag{40}$$

In particular, $\lambda = 1$ is not an eigenvalue of the iteration matrix.

The proof is similar to the proof of Proposition 3.1. The slight difference is that in the calculation we use (37) to simplify some terms and exploit the fact that T_k and $(\lambda I - B^k)^{-1}$ commute.

Proof. Since $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (14), there exists a non-zero vector $(\tilde{p}, \tilde{u}, y) \in \mathbb{C}^{n_u+n_u+n_\sigma}$ such that

$$\begin{cases} \lambda y = \frac{1}{1+\tau\alpha} y - \frac{\tau}{1+\tau\alpha} M^* \tilde{p}, \\ \lambda \tilde{u} = B^k \tilde{u} + \frac{1}{1+\tau\alpha} T_k M y - \frac{\tau}{1+\tau\alpha} T_k M M^* \tilde{p}, \\ \lambda \tilde{p} = \left[(B^*)^k - \frac{\tau}{1+\tau\alpha} X_k M M^* \right] \tilde{p} + U_k \tilde{u} + \frac{1}{1+\tau\alpha} X_k M y. \end{cases} \tag{41}$$

By the second equation in (41),

$$\tilde{u} = \frac{1}{1+\tau\alpha} T_k (\lambda I - B^k)^{-1} M y - \frac{\tau}{1+\tau\alpha} T_k (\lambda I - B^k)^{-1} M M^* \tilde{p}.$$

Inserting this result into the third equation we obtain

$$\begin{aligned} \lambda \tilde{p} = &\left((B^*)^k - \frac{\tau}{1+\tau\alpha} X_k M M^* - \frac{\tau}{1+\tau\alpha} U_k T_k (\lambda I - B^k)^{-1} M M^* \right) \tilde{p} \\ &+ \frac{1}{1+\tau\alpha} \left(X_k + U_k T_k (\lambda I - B^k)^{-1} \right) M y, \end{aligned}$$

or equivalently,

$$\tilde{p} + \frac{\tau}{1+\tau\alpha} [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M M^* \tilde{p} = \frac{1}{1+\tau\alpha} [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M y$$

where

$$V \coloneqq X_k (\lambda I - B^k) + U_k T_k = (\lambda - 1) X_k + T_k^* H^* H T_k$$

thanks to Lemma 4.1. Multiplying both sides of this equation with M^* gives

$$\begin{aligned} \left(I + \frac{\tau}{1+\tau\alpha} M^* [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M \right) M^* \tilde{p} = \\ \frac{1}{1+\tau\alpha} M^* [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M y. \end{aligned}$$

By the first equation in (16), $M^* \tilde{p} = \dfrac{1 - (1+\tau\alpha)\lambda}{\tau} y$, therefore

$$\begin{aligned} \frac{1 - (1+\tau\alpha)\lambda}{\tau} \left(I + \frac{\tau}{1+\alpha\tau} M^* [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M \right) y = \\ \frac{1}{1+\tau\alpha} M^* [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M y, \end{aligned}$$

or equivalently,

$$[(1+\alpha\tau)\lambda - 1]\, y + \tau\lambda M^* [\lambda I - (B^*)^k]^{-1} V (\lambda I - B^k)^{-1} M y = 0, \tag{42}$$

We prove that $y \neq 0$. Indeed if $y = 0$ then $M^* \tilde{p} = \dfrac{1 - (1+\tau\alpha)\lambda}{\tau} y = 0$ and

$$\tilde{u} = \frac{1}{1+\tau\alpha} T_k (\lambda I - B^k)^{-1} M y - \frac{\tau}{1+\tau\alpha} T_k (\lambda I - B^k)^{-1} M M^* \tilde{p} = 0.$$

Inserting these results into the third equation in (41) we obtain $\lambda\tilde{p} = (B^*)^k\tilde{p}$, which immediately implies $\tilde{p} = 0$ and gives a contradiction. Finally, by taking scalar product of (42) with y, then dividing by $\|y\|$, we obtain (40).
(ii) Now assume that $\lambda = 1$ is an eigenvalue of the iteration matrix, then (40) yields

$$\alpha + \|H(I-B)^{-1}My\|^2 = 0,$$

which cannot be true due to the injectivity of $H(I-B)^{-1}M$. □

In the following sections we will show that, for sufficiently small τ, equation (40) admits no solution $|\lambda| \geq 1$, thus algorithm (9) converges. When $\lambda \neq 0$, it is convenient to rewrite (40) as

$$(1+\tau\alpha)\lambda^2 - \lambda + \tau\langle M^*\left[I-(B^*)^k/\lambda\right]^{-1}\left[(\lambda-1)X_k + T_k^*H^*HT_k\right]\left(I-B^k/\lambda\right)^{-1}My, y\rangle = 0. \quad (43)$$

Remark 4.3. The simple scalar case where $n_u, n_\sigma, n_g = 1$ and $\alpha = 0$ is analyzed in [2], for which necessary and sufficient conditions on τ are derived.

Remark 4.4. Note that when $B = 0$ and $k \geq 2$, the semi-implicit k-step one-shot (9) is equivalent to the semi-implicit gradient descent method (7), which converges if and only if $(\rho(A^*A) - \alpha)\tau < 2$.

For the analysis we use some auxiliary results proved in Appendix A, and the following bounds for $s(B^k), T_k, X_k$.

Lemma 4.5. *If $\|B\| < 1$ then for every $k \geq 1$:*

$$s(B^k) = s((B^*)^k) \leq \frac{1}{1-\|B\|^k}, \quad \|T_k\| \leq \frac{1-\|B\|^k}{1-\|B\|}$$

and

$$\|X_k\| \leq \frac{\|H\|^2(1-k\|B\|^{k-1}+(k-1)\|B\|^k)}{(1-\|B\|)^2}.$$

Proof. The bound for $s(B^k)$ is proved using Lemma A.2 and $\|B^k\| \leq \|B\|^k$. Next, from (34) we have

$$\|T_k\| \leq 1 + \|B\| + ... + \|B\|^{k-1} = \frac{1-\|B\|^k}{1-\|B\|}.$$

From (36), if $k \geq 2$ we have

$$\begin{aligned}\|X_k\| &\leq \|H\|^2\big(\|B\|^{k-2} + \|B\|^{k-3}(1+\|B\|) + ... + (1+\|B\|+...+\|B\|^{k-2})\big)\\ &= \|H\|^2(1+2\|B\|+...+(k-1)\|B\|^{k-2}) = \frac{\|H\|^2(1-k\|B\|^{k-1}+(k-1)\|B\|^k)}{(1-\|B\|)^2}.\end{aligned}$$

□

4.2 Location of eigenvalues in the complex plane

We first establish conditions on the descent step $\tau > 0$ such that the real eigenvalues stay inside the unit disk. Recall that we have already proved that $\lambda = 1$ is not an eigenvalue for any k.

Proposition 4.6 (Real eigenvalues)**.** *Let $0 \neq \rho(B) < 1$ and $\lambda \in \mathbb{R}, \lambda \neq 1, |\lambda| \geq 1$. Then equation (43) cannot hold if $\tau > 0$ and*

$$\left(\|M\|^2\|X_k\|s(B^k)^2 - \frac{1}{2}\alpha\right)\tau < 1.$$

Moreover, if $\|B\| < 1$, the result is also true if $\tau > 0$ and

$$\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^2(1-\|B\|^k)^2}\left(1-k\|B\|^{k-1}+(k-1)\|B\|^k\right) - \frac{1}{2}\alpha\right)\tau < 1.$$

Proof. When $\lambda \in \mathbb{R}$ equation (43) can be rewritten as

$$(1+\tau\alpha)\lambda^2 - \lambda + \tau\|HT_k\left(I - \frac{B^k}{\lambda}\right)^{-1} My\|^2 \\ + \tau(\lambda-1)\langle M^*\left[I - \frac{(B^*)^k}{\lambda}\right]^{-1} X_k \left(I - \frac{B^k}{\lambda}\right)^{-1} My, y\rangle = 0. \quad (44)$$

We show that we can choose τ so that the left-hand side of the above equation is positive. First, we note that

$$\left|\langle M^*\left[I - \frac{(B^*)^k}{\lambda}\right]^{-1} X_k \left(I - \frac{B^k}{\lambda}\right)^{-1} My, y\rangle\right| \leq \|M\|^2\|X_k\|s(B^k)^2.$$

If $\lambda > 1$, we rewrite equation (44) again as

$$(1+\tau\alpha)\lambda(\lambda-1) + \tau\alpha + \tau\|HT_k\left(I - \frac{B^k}{\lambda}\right)^{-1} My\|^2 \\ + \tau(\lambda-1)\langle M^*\left[I - \frac{(B^*)^k}{\lambda}\right]^{-1} X_k \left(I - \frac{B^k}{\lambda}\right)^{-1} My, y\rangle = 0.$$

Since $\lambda(\lambda-1) \geq \lambda - 1$, $\|HT_k\left(I - \frac{B^k}{\lambda}\right)^{-1} My\|^2 \geq 0$ and $\tau\alpha \geq 0$, we choose τ such that

$$(1+\tau\alpha)(\lambda-1) - \tau(\lambda-1)\|M\|^2\|X_k\|s(B^k)^2 > 0,$$

or equivalently,

$$1 + \tau\alpha - \tau\|M\|^2\|X_k\|s(B^k) > 0.$$

If $\lambda \leq -1$, we consider equation (44). Since $\|HT_k\left(I - \frac{B^k}{\lambda}\right)^{-1} My\|^2 \geq 0$, we choose τ such that

$$(1+\tau\alpha)\lambda^2 - \lambda - \tau(1-\lambda)\|M\|^2\|X_k\|s(B^k)^2 > 0,$$

or equivalently,

$$\frac{(1+\tau\alpha)\lambda^2 - \lambda}{1-\lambda} - \tau\|M\|^2\|X_k\|s(B^k)^2 > 0,$$

Since the function $\lambda \mapsto \frac{(1+\tau\alpha)\lambda^2-\lambda}{\lambda-1}$ is decreasing on $(-\infty, -1]$, it suffices to choose τ such that

$$1 + \frac{\alpha}{2}\tau - \tau\|M\|^2\|X_k\|s(B^k)^2 > 0,$$

which proves the first statement of the proposition. Finally, the case $\|B\| < 1$ can be deduced using Lemma 4.5. □

For the general case of complex eigenvalues, the study is much more complicated and technical. The following proposition summarizes the results obtained.

Proposition 4.7 (Complex eigenvalues)**.** *If $0 \neq \rho(B) < 1$, there exists $\tau > 0$ sufficiently small such that equation (43) admits no solution $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $|\lambda| \geq 1$. In particular, if $\|B\| < 1$, given any $\delta_0 > 0$ and $0 < \theta_0 < \frac{\pi}{4}$, one can chooses*

$$\tau < \min_{1\leq i\leq 3}\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^2(1-\|B\|^k)^2}\psi_i(k, \|B\|) + C_i\alpha\right)^{-1},$$

where

$$\psi_1(k,b) \coloneqq 4b^{2k} + \sqrt{2}[1 - kb^{k-1} + (k-1)b^k](1+b^k),$$
$$\psi_2(k,b) \coloneqq \left(\frac{1}{2\sin\frac{\theta_0}{2}}(1-b^k)^2 + \sqrt{2}(1 - kb^{k-1} + (k-1)b^k)\right)(1+b^k)^2,$$
$$\psi_3(k,b) \coloneqq \frac{2c\sin\frac{\theta_0}{2}}{\delta_0}b^{2k} + \frac{\sqrt{c}}{\delta_0}[1 - kb^{k-1} + (k-1)b^k](1+b^{2k}) + 2\max\left(\frac{\sqrt{c}}{\delta_0}, \frac{\sqrt{c}}{\cos 2\theta_0}\right)[1 - kb^{k-1} + (k-1)b^k]b^k,$$

$$C_1 \coloneqq \sqrt{2} - 1, \quad C_2 \coloneqq \sqrt{2} + \frac{1}{2\sin\frac{\theta_0}{2}} - 1, \quad C_3 \coloneqq \frac{\sqrt{c}}{\delta_0} - 1, \quad c \coloneqq \frac{1 + 2\delta_0\sin\frac{3\theta_0}{2} + \delta_0^2}{\cos^2\frac{3\theta_0}{2}}.$$

Proof. **Step 1. Rewrite equation (43) by separating real and imaginary parts.**

Let $\lambda = R(\cos\theta + \mathrm{i}\sin\theta)$ in polar form where $R = |\lambda| \geq 1$ and $\theta \in (-\pi, \pi)$, $\theta \neq 0$. Write $1/\lambda = r(\cos\phi + \mathrm{i}\sin\phi)$ in polar form where $r = 1/|\lambda| = 1/R \leq 1$ and $\phi = -\theta \in (-\pi, \pi)$. By Lemma A.3 applied to $T = B^k$, we have

$$\left(I - \frac{B^k}{\lambda}\right)^{-1} = P_k(\lambda) + \mathrm{i}Q_k(\lambda), \quad \left(I - \frac{(B^*)^k}{\lambda}\right)^{-1} = P_k(\lambda)^* + \mathrm{i}Q_k(\lambda)^*$$

where $P_k(\lambda)$ and $Q_k(\lambda)$ are $\mathbb{C}^{n_u \times n_u}$ matrices that satisfy the following bounds in the case $\|B\| < 1$ for all $|\lambda| \geq 1$:

$$\|P_k(\lambda)\| \leq p \coloneqq \frac{1}{1 - \|B\|^k}, \tag{45}$$

$$\|Q_k(\lambda)\| \leq q_1 \coloneqq \frac{\|B\|^k}{1 - \|B\|^k} \quad \text{and} \quad \|Q_k(\lambda)\| \leq q_2|\sin\theta| \quad \text{with} \quad q_2 \coloneqq \frac{\|B\|^k}{(1 - \|B\|^k)^2} \tag{46}$$

These bounds still hold in the case $0 \neq \rho(B) < 1$ with

$$p \coloneqq (1 + \|B^k\|)s(B^k)^2, \quad q_1 \coloneqq \|B^k\|s(B^k)^2, \quad \text{and} \quad q_2 \coloneqq \frac{\|B^k\|}{1 - \|B^k\|}. \tag{47}$$

To simplify the notation, we will not explicitly write the dependence of P_k and Q_k on λ. Now we rewrite (43) as

$$(1+\tau\alpha)(\lambda^2 - \lambda) + \tau\alpha(\lambda - 1) + \tau\alpha + \tau G_k(P_k^* + \mathrm{i}Q_k^*, P_k + \mathrm{i}Q_k) + \tau(\lambda - 1)L_k(P_k^* + \mathrm{i}Q_k^*, P_k + \mathrm{i}Q_k) = 0 \tag{48}$$

where

$$G_k(X,Y) = \langle M^*XT_k^*H^*HT_kYMy, y\rangle, \quad L_k(X,Y) = \langle M^*XX_kYMy, y\rangle, \quad X, Y \in \mathbb{C}^{n_u \times n_u}.$$

Notice that G_k is a bilinear form and $G_k(X,Y) = G_k(Y^*, X^*)^*$ so that $G_k(X,Y) + G_k(X^*, Y^*)$ is real. Similarly, L_k has the same properties as G_k (note that $X_k^* = X_k$ by Lemma 4.1). With these properties of G_k and L_k, we expand (4.2) and take its real and imaginary parts, which yields

$$(1+\tau\alpha)\Re(\lambda^2 - \lambda) + \tau\alpha\Re(\lambda - 1) + \tau\alpha + \tau G_{1,k} + \tau[\Re(\lambda - 1)L_{1,k} - \Im(\lambda - 1)L_{2,k}] = 0 \tag{49}$$

and

$$(1+\tau\alpha)\Im(\lambda^2-\lambda)+\tau\alpha\Im(\lambda-1)+\tau G_{2,k}+\tau[\Im(\lambda-1)L_{1,k}+\Re(\lambda-1)L_{2,k}]=0 \tag{50}$$

where

$$G_{1,k}\coloneqq G_k(P_k^*,P_k)-G_k(Q_k^*,Q_k),\quad G_{2,k}\coloneqq G_k(P_k^*,Q_k)+G_k(Q_k^*,P_k),$$
$$L_{1,k}\coloneqq L_k(P_k^*,P_k)-L_k(Q_k^*,Q_k)\quad\text{and}\quad L_{2,k}\coloneqq L_k(P_k^*,Q_k)+L_k(Q_k^*,P_k).$$

Step 2. Use a suitable combination of equations (49) and (50).
Let $\gamma\in\mathbb{R}$. Multiplying equation (50) with γ then summing it with equation (49), we obtain:

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma\Im(\lambda^2-\lambda)]+\tau\alpha[\Re(\lambda-1)+\gamma\Im(\lambda-1)]+\tau\alpha\\ +\tau G_k(P_k^*+\gamma Q_k^*,P_k+\gamma Q_k)-\tau(1+\gamma^2)G_k(Q_k^*,Q_k)\\ +\tau[\Re(\lambda-1)+\gamma\Im(\lambda-1)]L_{1,k}+\tau[\gamma\Re(\lambda-1)-\Im(\lambda-1)]L_{2,k}=0.\end{aligned} \tag{51}$$

Now we consider four cases of λ as in the proof for $k=1$ namely:

- *Case 1.* $\Re(\lambda^2-\lambda)\geq 0$;
- *Case 2.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in[\theta_0,\pi-\theta_0]\cup[-\pi+\theta_0,-\theta_0]$ for fixed $0<\theta_0<\frac{\pi}{4}$;
- *Case 3.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in(-\theta_0,\theta_0)$ for fixed $0<\theta_0<\frac{\pi}{4}$;
- *Case 4.* $\Re(\lambda^2-\lambda)<0$ and $\theta\in(\pi-\theta_0,\pi)\cup(-\pi,-\pi+\theta_0)$ for fixed $0<\theta_0<\frac{\pi}{4}$.

Cases 1, 2 and 3 are respectively treated in Lemmas 4.8, 4.9 and 4.10 below. Notice that case 4 corresponds to an empty set, according to the Lemma A.4 (iv). The statement of the proposition easily follows from the combination of Lemmas 4.8, 4.9 and 4.10. □

In the lemmas below we shall make use of the following obvious properties where $\|y\|=1$:

$$0\leq G_k(X^*,X)=\|HT_kXMy\|^2\leq(\|H\|\|M\|\|T_k\|\|X\|)^2,\quad\forall X\in\mathbb{C}^{n_u\times n_u}. \tag{52}$$

$$\begin{aligned}|L_{1,k}|=|L_k(P_k^*,P_k)-L_k(Q_k^*,Q_k)|&\leq|L_k(P_k^*,P_k)|+|L_k(Q_k^*,Q_k)|\\ &\leq\|X_k\|\|M\|^2(\|P_k\|^2+\|Q_k\|^2)\leq\|X_k\|\|M\|^2(p^2+q_1^2)\end{aligned} \tag{53}$$

and

$$\begin{aligned}|L_{2,k}|=|L_k(P_k^*,Q_k)+L_k(Q_k^*,P_k)|&\leq|L_k(P_k^*,Q_k)|+|L_k(Q_k^*,P_k)|\\ &\leq 2\|X_k\|\|M\|^2\|P_k\|\|Q_k\|\leq 2\|X_k\|\|M\|^2pq_1.\end{aligned} \tag{54}$$

Lemma 4.8 (Case 1). *Let $\rho(B)<1$ and let $|\lambda|\geq 1$ with $\Re(\lambda^2-\lambda)\geq 0$. Then equation (43) cannot hold if $\tau>0$ and*

$$\left(4\|H\|^2\|M\|^2\|T_k\|^2\|B^k\|^2s(B^k)^4+\sqrt{2}\|M\|^2\|X_k\|(1+2\|B^k\|)^2s(B^k)^4+(\sqrt{2}-1)\alpha\right)\tau<1.$$

Moreover, if $\|B\|<1$, the result is also true if $\tau>0$ and

$$\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^2(1-\|B\|^k)^2}\psi_1(k,\|B\|)+(\sqrt{2}-1)\alpha\right)\tau<1.$$

where $\psi_1(k,b)\coloneqq 4b^{2k}+\sqrt{2}(1-kb^{k-1}+(k-1)b^k)(1+b^k)$.

Proof. Define

$$\gamma_1 = \gamma_1(\lambda) := \begin{cases} 1 & \text{if } \Im(\lambda^2 - \lambda) \geq 0, \\ -1 & \text{if } \Im(\lambda^2 - \lambda) < 0 \end{cases}$$

as in Lemma A.4 (i). Writing (51) for $\gamma = \gamma_1$ as in Lemma A.4 and using $\gamma_1^2 = 1$ we obtain

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)] + \tau\alpha[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)] + \tau\alpha \\ + \tau G_k(P_k^* + \gamma_1 Q_k^*, P_k + \gamma_1 Q_k) - 2\tau G_k(Q_k^*, Q_k) \\ + \tau[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)]L_{1,k} + \tau[\gamma_1\Re(\lambda-1) - \Im(\lambda-1)]L_{2,k} = 0. \quad (55)\end{aligned}$$

Since $G_k(P_k^* + \gamma_1 Q_k^*, P_k + \gamma_1 Q_k) \geq 0$ by (52) and $\tau\alpha \geq 0$, the left-hand side of (55) is positive if τ satisfies

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)] - \tau\alpha[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)] - 2\tau G_k(Q_k^*, Q_k) \\ - \tau[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)]|L_{1,k}| - \tau[\gamma_1\Re(\lambda-1) - \Im(\lambda-1)]|L_{2,k}| > 0,\end{aligned}$$

or equivalently,

$$\begin{aligned}1 + \tau\alpha - \tau\alpha\frac{|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} - 2\tau\frac{G_k(Q_k^*, Q_k)}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} \\ - \tau\frac{|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}|L_{1,k}| - \tau\frac{|\gamma_1\Re(\lambda-1) - \Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)}|L_{2,k}| > 0. \quad (56)\end{aligned}$$

Notice that the choice of γ_1 ensures $\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda) > 0$. In the following we derive upper bounds independent from λ for the terms appearing with the negative sign in (56). By Lemma A.4 (i) we have

$$\frac{|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} \leq \frac{\sqrt{1+\gamma_1^2}|\lambda-1|}{|\lambda(\lambda-1)|} = \frac{\sqrt{2}}{|\lambda|} \leq \sqrt{2}$$

and

$$\frac{|\gamma_1\Re(\lambda-1) - \Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} \leq \frac{\sqrt{1+\gamma_1^2}|\lambda-1|}{|\lambda(\lambda-1)|} = \frac{\sqrt{2}}{|\lambda|} \leq \sqrt{2}.$$

Using again Lemma A.4 (i) and (54), we have

$$\begin{aligned}\frac{G_k(Q_k^*, Q_k)}{\Re(\lambda^2-\lambda)+\gamma_1\Im(\lambda^2-\lambda)} &\leq \frac{(\|H\|\|M\|\|T_k\|\sin\theta|q_2)^2}{2|\sin(\theta/2)|} \\ &= 2\left|\sin\frac{\theta}{2}\right|\cos^2\frac{\theta}{2}\|H\|^2\|M\|^2\|T_k\|^2 q_2^2 \leq 2\|H\|^2\|M\|^2\|T_k\|^2 q_2^2.\end{aligned}$$

Recall from (53) and (54) that

$$|L_{1,k}| \leq \|X_k\|\|M\|^2(p^2+q_1^2), \quad |L_{2,k}| \leq 2\|X_k\|\|M\|^2 pq_1.$$

Inserting these previous inequalities in (56) gives the desired result thanks to expressions (45), (46) and (47) of respectively p, q_1 and q_2, and thanks to Lemma 4.5. □

Lemma 4.9 (Case 2). *Let $\rho(B) < 1$ and let $|\lambda| \geq 1$, $\Re(\lambda^2-\lambda) < 0$, $\theta \in [\theta_0, \pi-\theta_0] \cup [-\pi+\theta_0, -\theta_0]$ for given $0 < \theta_0 \leq \frac{\pi}{4}$. Then equation (43) cannot hold if $\tau > 0$ and*

$$\left(\left(\frac{1}{2\sin\frac{\theta_0}{2}}\|H\|^2\|M\|^2\|T_k\|^2 + \sqrt{2}\|M\|^2\|X_k\|\right)(1+2\|B^k\|)^2 s(B^k)^4 + (\sqrt{2}-1)\alpha\right)\tau < 1.$$

Moreover, if $\|B\| < 1$, *the result is also true if*

$$\tau < \left(\frac{\|H\|^2 \|M\|^2}{(1-\|B\|)^2(1-\|B\|^k)^2} \psi_2(k, \|B\|) + (\sqrt{2}-1)\alpha \right)^{-1}$$

where $\psi_2(k,b) = \left(\frac{1}{2\sin\frac{\theta_0}{2}}(1-b^k)^2 + \sqrt{2}(1-kb^{k-1}+(k-1)b^k) \right)(1+b^k)^2$.

Proof. Define

$$\gamma_2 = \gamma_2(\lambda) := \begin{cases} -1 & \text{if } \Im(\lambda^2-\lambda) \geq 0, \\ 1 & \text{if } \Im(\lambda^2-\lambda) < 0 \end{cases}$$

as in Lemma A.4 (ii). Writing (51) for $\gamma = \gamma_2$ as in Lemma A.4 (ii) and using $\gamma_2^2 = 1$, we obtain

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)] + \tau\alpha[\Re(\lambda-1)+\gamma_2\Im(\lambda-1)] + \tau\alpha \\ + \tau G_k(P_k^* + \gamma_2 Q_k^*, P_k + \gamma_2 Q_k) - 2\tau G_k(Q_k^*, Q_k) \\ + \tau[\Re(\lambda-1)+\gamma_2\Im(\lambda-1)]L_{1,k} + \tau[\gamma_2\Re(\lambda-1)-\Im(\lambda-1)]L_{2,k} = 0. \quad (57)\end{aligned}$$

Since $G_k(Q_k^*, Q_k) \geq 0$ by (52), the left-hand side of (57) is negative if τ satisfies

$$\begin{aligned}(1+\tau\alpha)[\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)] + \tau\alpha|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)| + \tau\alpha \\ + \tau G_k(P_k^* + \gamma_2 Q_k^*, P_k + \gamma_2 Q_k) \\ + \tau|\Re(\lambda-1)+\gamma_1\Im(\lambda-1)||L_{1,k}| + \tau|\gamma_1\Re(\lambda-1)-\Im(\lambda-1)||L_{2,k}| < 0.\end{aligned}$$

or equivalently,

$$\begin{aligned}-1-\tau\alpha + \tau\alpha\frac{|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \\ + \frac{\tau\alpha}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} + \tau\frac{G_k(P_k^* + \gamma_2 Q_k^*, P_k + \gamma_2 Q_k)}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \\ + \tau\frac{|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}|L_{1,k}| + \tau\frac{|\gamma_2\Re(\lambda-1)-\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|}|L_{2,k}| < 0. \quad (58)\end{aligned}$$

Notice that the choice of γ_2 ensures $\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda) < 0$. In the following we derive upper bounds independent from λ for the terms appearing with the positive sign in (58). By Lemma A.4 (ii) we have

$$\frac{|\Re(\lambda-1)+\gamma_2\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \leq \frac{\sqrt{1+\gamma_2^2}|\lambda-1|}{|\lambda(\lambda-1)|} = \frac{\sqrt{2}}{|\lambda|} \leq \sqrt{2},$$

$$\frac{|\gamma_2\Re(\lambda-1)-\Im(\lambda-1)|}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \leq \frac{\sqrt{1+\gamma_2^2}|\lambda-1|}{|\lambda(\lambda-1)|} = \frac{\sqrt{2}}{|\lambda|} \leq \sqrt{2},$$

and

$$\frac{1}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \leq \frac{1}{2\sin\frac{\theta_0}{2}}.$$

Using again Lemma A.4 (ii) and (52), we have

$$\frac{G_k(P_k^* + \gamma_2 Q_k^*, P_k + \gamma_2 Q_k)}{|\Re(\lambda^2-\lambda)+\gamma_2\Im(\lambda^2-\lambda)|} \leq \frac{\|H\|^2\|M\|^2\|T_k\|^2(p+q_1)^2}{2\sin\frac{\theta_0}{2}}.$$

Recall from (53) and (54) that

$$|L_{1,k}| \leq \|X_k\|\|M\|^2(p^2+q_1^2), \quad |L_{2,k}| \leq 2\|X_k\|\|M\|^2 pq_1.$$

Inserting these previous inequalities in (58) gives the desired result thanks to expressions (45), (46) and (47) of respectively p and q_1, and thanks to Lemma 4.5. □

Lemma 4.10 (Case 3). *Let $\rho(B) < 1$ and let $|\lambda| \geq 1$, $\Re(\lambda^2 - \lambda) < 0$, $\theta \in (-\theta_0, \theta_0)$ for given $0 < \theta_0 \leq \frac{\pi}{4}$. For any $\delta_0 > 0$, equation (43) cannot hold if $\tau > 0$ and*

$$\left(\left[\frac{2c \sin\frac{\theta_0}{2}}{\delta_0} \|H\|^2 \|M\|^2 \|T_k\|^2 \|B^k\|^2 + \frac{\sqrt{c}}{\delta_0} \|M\|^2 \|X_k\| (1 + 2\|B^k\| + 2\|B^k\|^2) \right.\right.$$
$$\left.\left. + 2 \max\left(\frac{\sqrt{c}}{\delta_0}, \frac{\sqrt{c}}{\cos 2\theta_0} \right) \|M\|^2 \|X_k\| (\|B^k\| + \|B^k\|^2) \right] s(B^k)^2 + (\sqrt{2} - 1)\alpha \right) \tau < 1$$

where $c = c(\theta_0, \delta_0) := \left(1 + 2\delta_0 \sin\frac{3\theta_0}{2} + \delta_0^2\right) / \cos^2\frac{3\theta_0}{2}$. Moreover, if $\|B\| < 1$, the result is also true if $\tau > 0$ and

$$\left(\frac{\|H\|^2 \|M\|^2}{(1 - \|B\|)^2 (1 - \|B\|^k)^2} \psi_3(k, \|B\|) + \left(\frac{\sqrt{c}}{\delta_0} - 1 \right) \alpha \right) \tau < 1$$

where

$$\psi_3(k, b) := \frac{2c \sin\frac{\theta_0}{2}}{\delta_0} b^{2k}$$
$$+ \left(\frac{\sqrt{c}}{\delta_0}(1 + b^{2k}) + 2 \max\left(\frac{\sqrt{c}}{\delta_0}, \frac{\sqrt{c}}{\cos 2\theta_0} \right) b^k \right) (1 - kb^{k-1} + (k-1)b^k) b^k.$$

Proof. Define

$$\gamma_3 = \gamma_3(\mathrm{sign}(\theta)) := \begin{cases} \left(\delta_0 + \sin\frac{3\theta_0}{2}\right) / \cos\frac{3\theta_0}{2} & \text{if } \theta > 0, \\ -\left(\delta_0 + \sin\frac{3\theta_0}{2}\right) / \cos\frac{3\theta_0}{2} & \text{if } \theta < 0 \end{cases}$$

as in Lemma A.4 (iii). Writing (51) for $\gamma = \gamma_3$ we obtain

$$(1 + \tau\alpha)[\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)] + \tau\alpha[\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)] + \tau\alpha$$
$$+ \tau G_k(P_k^* + \gamma_3 Q_k^*, P_k + \gamma_3 Q_k) - (1 + \gamma_3^2)\tau G_k(Q_k^*, Q_k)$$
$$+ \tau\left([\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)] L_{1,k} + [\gamma_3 \Re(\lambda - 1) - \Im(\lambda - 1)] L_{2,k}\right) = 0. \quad (59)$$

Since $G(P^* + \gamma_3 Q^*, P + \gamma_3 Q) \geq 0$ and $\tau\alpha \geq 0$, the left-hand side of (59) is positive if τ satisfies

$$(1 + \tau\alpha)[\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)] - \tau\alpha |\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)| - (1 + \gamma_3^2)\tau G_k(Q_k^*, Q_k)$$
$$- \tau |\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)| |L_{1,k}| + |\gamma_3 \Re(\lambda - 1) - \Im(\lambda - 1)| |L_{2,k}| > 0,$$

or equivalently,

$$1 + \tau\alpha - \tau\alpha \frac{|\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} - \tau(1 + \gamma_3^2) \frac{G_k(Q_k^*, Q_k)}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)}$$
$$- \tau \frac{|\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} |L_{1,k}| - \tau \frac{|\gamma_3 \Re(\lambda - 1) - \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} |L_{2,k}| > 0. \quad (60)$$

Notice that the choice of γ_3 ensures $\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda) > 0$, also

$$1 + \gamma_3^2 = 1 + \frac{\left(\delta_0 + \sin\frac{3\theta_0}{2}\right)^2}{\cos^2\frac{3\theta_0}{2}} = \frac{1 + 2\delta_0 \sin\frac{3\theta_0}{2} + \delta_0^2}{\cos^2\frac{3\theta_0}{2}} =: c$$

is a constant greater than δ_0^2. In the following we derive upper bounds independent from λ for the terms appearing with the negative sign in (60). By Lemma A.4 (iii) we have

$$\frac{|\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} \leq \frac{\sqrt{1 + \gamma_3^2}}{\delta_0} = \frac{\sqrt{c}}{\delta_0}$$

and

$$\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} \leq \max\left(\frac{\sqrt{1+\gamma_3^2}}{\delta_0}, \frac{\sqrt{1+\gamma_3^2}}{\cos 2\theta_0}\right) = \max\left(\frac{\sqrt{c}}{\delta_0}, \frac{\sqrt{c}}{\cos 2\theta_0}\right).$$

Using again Lemma A.4 (iii) and (52), we have

$$\begin{aligned}\frac{G_k(Q_k^*, Q_k)}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} &\leq \frac{(\|H\|\|M\|\|T_k\|\sin\theta|q_2)^2}{2\delta_0|\sin(\theta/2)|} \\ &= \frac{2}{\delta_0}\left|\sin\frac{\theta}{2}\right|\cos^2\frac{\theta}{2}\|H\|^2\|M\|^2\|T_k\|^2 q_2^2 \leq \frac{2}{\delta_0}\|H\|^2\|M\|^2\|T_k\|^2 q_2^2.\end{aligned}$$

Recall from (53) and (54) that

$$|L_{1,k}| \leq \|X_k\|\|M\|^2(p^2+q_1^2), \quad |L_{2,k}| \leq 2\|X_k\|\|M\|^2 pq_1.$$

Inserting these previous inequalities in (60) gives the desired result thanks to expressions (45), (46) and (47) of respectively p, q_1 and q_2, and thanks to Lemma 4.5. □

4.3 Final result ($k \geq 1$)

Considering Proposition 4.6 (for real eigenvalues) and taking the bound in Proposition 4.7 (for complex eigenvalues), we finally obtain a sufficient condition on the descent step τ to ensure convergence of the multi-step one-shot method.

Theorem 4.11 (Convergence of semi-implicit k-step one-shot, $k \geq 1$)**.** *Under assumption (4), the k-step one-shot method (9), $k \geq 1$, converges for sufficiently small τ. In particular, for $\|B\| < 1$, there exists an explicit piecewise at-most-$(4k)$th-order-polynomial function $\mathcal{P}_k$ and a constant $C > 0$ independent of k such that $\mathcal{P}_k > 0$ on $[0,1)$ and it is enough to take $\tau > 0$ and*

$$\left(\frac{\|H\|^2\|M\|^2}{(1-\|B\|)^2(1-\|B\|^k)^2}\mathcal{P}_k(\|B\|) + C\alpha\right)\tau < 1.$$

Theorem 4.11 includes the case $B = 0$, but the bound for τ in this case is quite far from optimal. Note that the optimal bound for τ in the case $B = 0$ can be found in Proposition 3.2 (for $k = 1$) and Remark 4.4 (for $k \geq 2$).

5 Numerical experiments on a toy problem

In order to compare the performance of classical gradient algorithm (7) with the k-step one-shot algorithms (9), we propose the following toy model related to inverse conductivity problem in a cavity. Given $\Omega \subset \mathbb{R}^2$ an open bounded regular domain, we consider the direct problem for the linearized scattered field $u \in \mathrm{H}^1(\Omega)$ given by the Helmholtz equation

$$\begin{cases} \operatorname{div}(\sigma_0\nabla u) + \omega^2 u = \operatorname{div}(\sigma\nabla u_0), & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{61}$$

where the incident field $u_0 \in \mathrm{H}^1(\Omega)$ satisfies

$$\begin{cases} \operatorname{div}(\sigma_0\nabla u_0) + \omega^2 u_0 = 0, & \text{in } \Omega, \\ u_0 = f, & \text{on } \partial\Omega, \end{cases} \tag{62}$$

with the boundary data $f \in \mathrm{H}^{1/2}(\partial\Omega)$. Here σ and $\sigma_0 \in \mathrm{L}^\infty(\Omega)$ are supposed to be positive-definite functions such that the support $\overline{\Omega}_0$ of σ is strictly included inside Ω. The variational formulation of (61) is: find $u \in \mathrm{H}^1_0(\Omega)$ such that

$$\int_\Omega \sigma_0 \nabla u \cdot \nabla v \,\mathrm{d}x - \int_\Omega \omega^2 uv \,\mathrm{d}x = \int_\Omega \sigma \nabla u_0 \cdot \nabla v \,\mathrm{d}x, \quad \forall v \in \mathrm{H}^1_0(\Omega). \tag{63}$$

We consider the inverse problem of retrieving σ from the measurement $g \coloneqq \sigma_0 \frac{\partial u}{\partial \nu}\big|_{\partial\Omega}$.
By discretizing u using $\mathbb{P}^1$-Lagrange finite elements on a mesh $\mathcal{T}_h(\Omega)$ of Ω, and σ by $\mathbb{P}^0$-Lagrange finite elements on a coarser mesh $\widetilde{\mathcal{T}}_{h'}(\Omega_0)$ of Ω_0, the discretization of (63) leads to a linear system of the form

$$A_1 \vec{u} = A_2 \vec{\sigma}, \tag{64}$$

where $\vec{u} \in \mathbb{R}^{n_u}$, $\vec{\sigma} \in \mathbb{R}^{n_\sigma}$ with n_u denoting the number of inner nodes of $\mathcal{T}_h(\Omega)$ and n_σ denoting the number of triangles in $\widetilde{\mathcal{T}}_{h'}(\Omega_0)$. In order to rewrite the linear system in the form (1) with a controllable norm for the matrix B, we choose to parameterize σ_0 as $\sigma_0 = \tilde{\sigma}_0 + \delta\sigma_{\mathrm{r}}$ where $\delta > 0$ is a small parameter and $\sigma_{\mathrm{r}} \leq 1$ is a random function. With this choice of σ_0, we can write the matrix A_1 as $A_1 = A_{11} + \delta A_{12}$ where A_{11} corresponds to the discretization of the bilinear form $\int_\Omega (\tilde{\sigma}_0 \nabla u \cdot \nabla v - \omega^2 uv) \,\mathrm{d}x$ on $\mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega)$ and A_{12} corresponds to the discretization of the bilinear form $\int_\Omega \sigma_{\mathrm{r}} \nabla u \cdot \nabla v \,\mathrm{d}x$ on $\mathrm{H}^1_0(\Omega) \times \mathrm{H}^1_0(\Omega)$. For ω^2 not a Dirichlet eigenvalue of $-\operatorname{div}(\tilde{\sigma}_0 \nabla u)$ in Ω, the matrix A_{11} is invertible and we can equivalently write (64) as

$$\vec{u} = A_{11}^{-1}(-\delta A_{12}\vec{u} + A_2 \vec{\sigma})$$

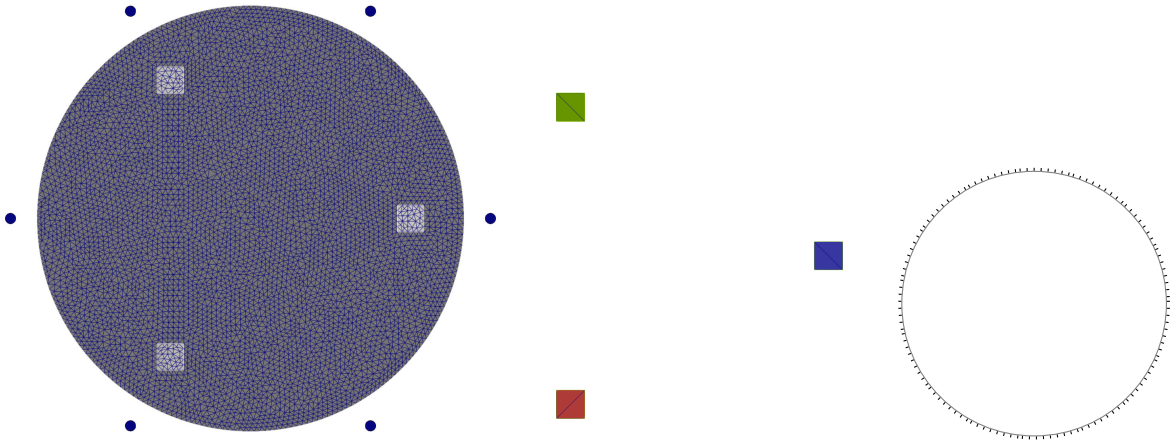
which is in the form (1) with $B = \delta A_{11}^{-1} A_{12}$, $M = A_{11}^{-1} A_2$ and $F = 0$. Indeed $\|B\| < 1$ for sufficiently small δ. We employed the finite element library FreeFEM [13] to generate the matrices A_{11}, A_{12}, A_2 and the measurement operator $H \in \mathbb{R}^{n_g \times n_u}$, which is the discretization of the normal trace operator $\sigma_0 \frac{\partial u}{\partial \nu}\big|_{\partial\Omega}$, where n_g denotes the number of nodes on $\partial\Omega$.

For the numerical tests below, we set $\omega = 2\pi$, $\tilde{\sigma}_0 = 1$, $\delta = 0.01$ and the mesh size $h = \lambda/20 = 0.05$ where $\lambda = \sqrt{\tilde{\sigma}_0} 2\pi/\omega = 1$. The domain Ω is the disk of radius $R = 2\lambda$ and Ω_0 is formed by three squares as in Figures 1 and 4. To generate measurements g, we use 6 different incident fields u_0 corresponding to imposing boundary data f given by the zero-order Bessel functions of the second kind centered at six different source points outside Ω (located on the circle of radius $R + 0.25\lambda$). The cost functional is then the normalized sum of the cost functionals associated with each of the incident fields. We take as exact solution $\sigma^{\mathrm{ex}} = 10$ in each square, and as initial guess $\sigma^0 = 12$ in each square.

The case of noise-free data

We first consider the case of noise-free data, without regularization ($\alpha = 0$). Here, the domain Ω_0 is formed by three squares of size $\lambda/4$ distributed as shown in Figure 1. We set the mesh size $h' = \lambda/4$ so that each square is divided into two triangles (see Figure 1b), which gives $n_\sigma = 6$. The mesh used for Ω (see Figure 1a) leads to $n_u = 5849$. The boundary mesh used for generating the data g (see Figure 1c) is two times coarser than the mesh for u and this gives $n_g = 125$. We compare the performances of k-step one-shot methods (9) (which coincide with (8) in the present case $\alpha = 0$). Recall that k is the number of inner iterations on the direct and adjoint problems. We consider two series of experiments.

In the first one, we study the dependence on the descent step τ. In Figure 2a–2b and 2c–2d we respectively fix $k = 1$ and $k = 2$ and compare k-step one-shot methods with the gradient descent method. We plot in semi-log scale the value of the cost functional in terms of the (outer) iteration number n in (6) and (9). We can verify that for sufficiently small τ, the k-step one-shot methods converge. This is not always the case for larger value of τ. In particular, for $\tau = 2$, while



(a) Mesh for u. (b) Mesh for σ. (c) Mesh for g.

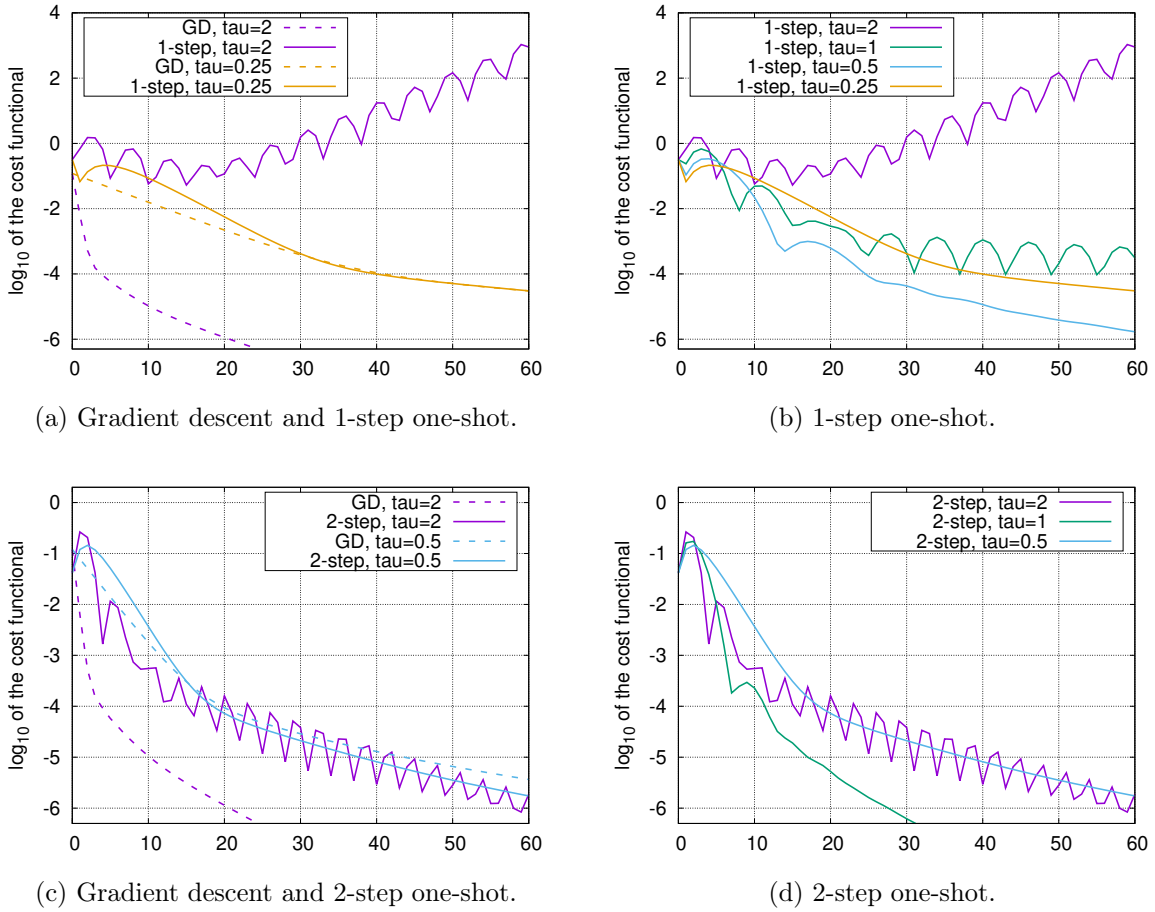
Figure 1: Meshes used to generate the matrices A_1, A_2 and H in the case of noise-free data.

gradient descent and 2-step one-shot converge, 1-step one-shot diverges. Oscillations may appear on the convergence curve for certain values of τ, but they gradually disappear when τ gets smaller. For sufficiently small τ, the convergence curves of one-shot methods are comparable to the one of gradient descent.

In the second series of experiments, we study the dependence on the number of inner iterations k, for fixed τ. First (Figures 3a and 3c), we investigate for which k the convergence curve of k-step one-shot is comparable with the one of gradient descent, and, as in the previous figures, on the horizontal axis we indicate the (outer) iteration number n. For $\tau = 2$ (see Figure 3a), we observe that for $k = 3, 4$ the convergence curves of k-step one-shot are close to the one of gradient descent. Note that with 3 inner iterations the L^2 error between u^n and the exact solution to the forward problem ranges between $59 \cdot 10^{-9}$ and 0.48334 for different n in (9); in fact, this error is rather significant at the beginning of the iteration, then it reduces as we get closer to the convergence for the parameter σ. Therefore, incomplete inner iterations on the forward problem are enough to have good precision on the solution of the inverse problem. In the particular case $\tau = 2.5$ (see Figure 3c), we observe an interesting phenomenon: when $k = 3, 4, 10$, with k-step one-shot the cost functional decreases even faster than with gradient descent. For larger values of k, for example $k = 14$, the convergence curve of one-shot method gets closer to the one of gradient descent as one may expect. Next, in Figures 3b and 3d, we display the results of the same experiment as in Figures 3a and 3c, but this time on the horizontal axis we indicate the accumulated inner iteration number, which is equal to kn where n is the number of outer iterations. This would allow us to compare the overall speed of convergence among k-step one-shot methods. For $\tau = 2$ (respectively for $\tau = 2.5$), $k = 3$ and $k = 4$ (respectively $k = 3$) appear to provide the fastest rate of convergence. This confirms the potential interest that this type of method would have in solving large-scale inverse problem since few inner iterations are capable of providing fast convergence.

The case of noisy data

We consider now the case where the measurements g are corrupted with noise. More specifically, we replace the vector $g = Hu(\sigma^{\mathrm{ex}})$ in the cost functional by the vector $g^\varepsilon \in \mathbb{R}^{n_g}$ with $g_i^\varepsilon := g_i + \varepsilon_i g_i$, where the ε_i are random numbers uniformly distributed between $-\varepsilon$ and ε for a noise level ε. In order to hit the ill-posedness of the inverse problem, we artificially increase the size of the discretization space for the parameter σ to $n_\sigma = 698$ by enlarging the size of the squares (now the size of their edges equals $2\lambda/3$ and the distance of their center from the boundary equals λ, see Figure 4b). We also show in Figure 4a the mesh for u, which is about 20 times finer than the boundary mesh used



(a) Gradient descent and 1-step one-shot.

(b) 1-step one-shot.

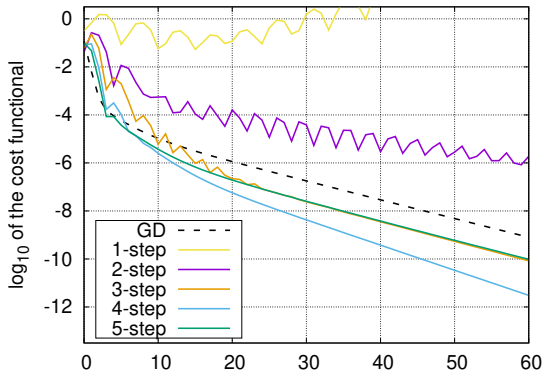
(c) Gradient descent and 2-step one-shot.

(d) 2-step one-shot.

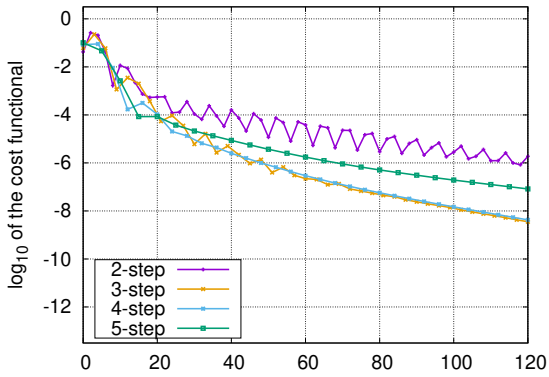
Figure 2: Convergence curves of gradient descent and k-step one-shot: dependence on the descent step τ.

for generating the 6 data g^{ε} (see Figure 4c). In this new configuration, $n_u = 5582$ and $n_g = 12$. Notice also that the meshes of the squares in Figure 4a and 4b do not coincide.

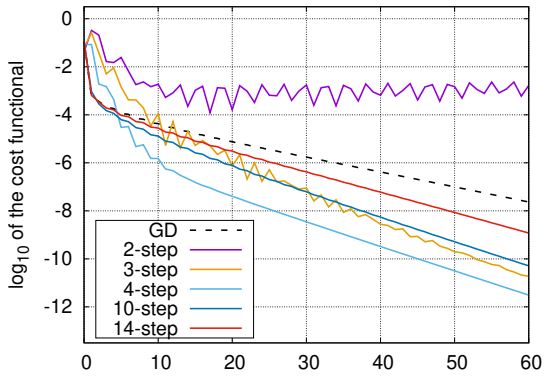
We perform several numerical tests with different noise levels: $\varepsilon = 1\%, 3\%$ and 5%. We choose the regularization parameter α depending on the noise level ε in order to optimize the accuracy of the reconstruction. This choice, which is made by trial and error, does not affect much the convergence of the algorithms (see Figures 5 and 6) and mainly reduces the size of the oscillations for the reconstructed σ. The convergence curves displayed in Figure 5 for different noise levels show that the semi-implicit k-step one-shot methods ($k = 3, 4$) require a similar number of outer iterations as the semi-implicit gradient descent algorithm to achieve the same precision. We also see from Figure 5 that the convergence curves with $\alpha \neq 0$ look less steep than those with $\alpha = 0$. Since in the case of noisy data the convergence for the cost functional does not imply in general the accuracy of the reconstructed parameter σ, we also plot convergence curves for the relative error on σ in Figure 6 to check the quality of the reconstruction. Also in these plots we see that the semi-implicit k-step one-shot methods ($k = 3, 4$) require a similar number of outer iterations as the semi-implicit gradient descent algorithm to achieve the same accuracy. Indeed, this proves the potential of these methods since only a few inner iterations are used. Moreover, the chosen values of $\alpha \neq 0$ adapted to the noise level give fairly better relative error curves. Finally, to better show the effect of the regularization, Figures 7a–7f display the final reconstructions for σ by semi-implicit 3-step one-shot, whose relative error curves were presented in Figures 6a–6f. On the left of the color scale we also indicate the actual maximum and minimum values attained for each reconstruction; note that this range gets wider for higher noise level. These plots confirm the benefit of regularization for treating



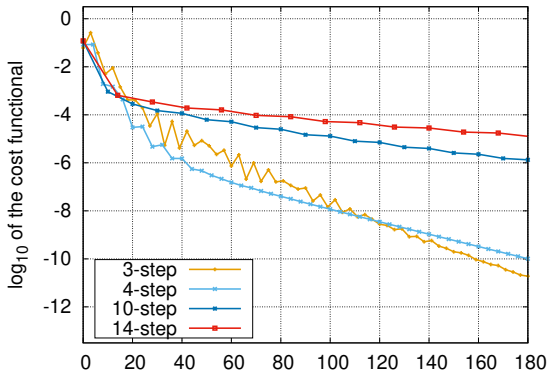
(a) Gradient descent and k-step one-shot with $\tau = 2$.



(b) k-step one-shot with $\tau = 2$ in terms of the accumulated inner iteration number.



(c) Gradient descent and k-step one-shot with $\tau = 2.5$.



(d) k-step one-shot with $\tau = 2.5$ in terms of the accumulated inner iteration number.

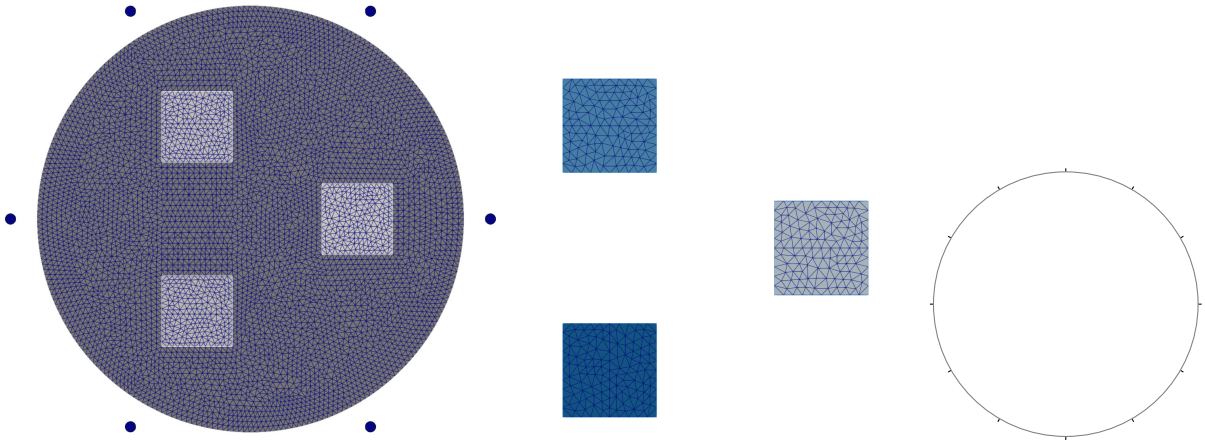
Figure 3: Convergence curves of gradient descent and k-step one-shot: dependence on the number of inner iterations k.

noisy data.

6 Conclusion

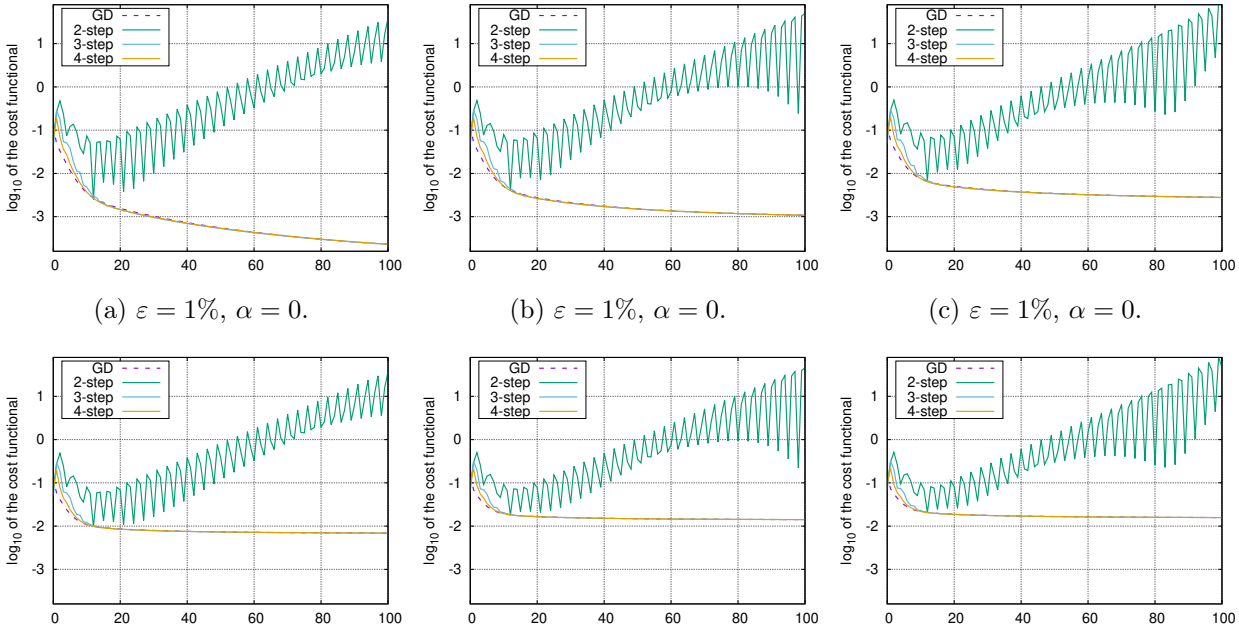
We have proved sufficient conditions on the descent step for the convergence of semi-implicit multi-step one-shot methods, with a regularization parameter $\alpha \geq 0$. This complement the results obtained in our research report [2] for explicit schemes with no regularization. Although these bounds on the descent step are not optimal, to our knowledge no other bounds, explicit in the number of inner iterations, are available in the literature for multi-step one-shot methods. Furthermore, we have shown in the numerical experiments that very few inner iterations on the forward and adjoint problems may be enough to guarantee similar results as for classical gradient descent algorithm.

These preliminary numerical results are just proof of concept since the size of the direct problem is not very large. In our future work, we shall carry out more realistic numerical investigation where the iterative solvers are based on domain decomposition methods (see e.g. [4]), which are well adapted to large-scale problems. In addition, the inner fixed point iterations could be replaced by more efficient Krylov subspace methods, such as conjugate gradient or GMRES, and one could use L-BFGS instead of gradient descent as optimization algorithm. Another interesting issue is how to adapt the number of inner iterations in the course of the outer iterations. Moreover, based on this linear inverse problem study, we plan to tackle the challenging case of non-linear inverse problems.



(a) Mesh for u. (b) Mesh for σ. (c) Mesh for g^{ε}.

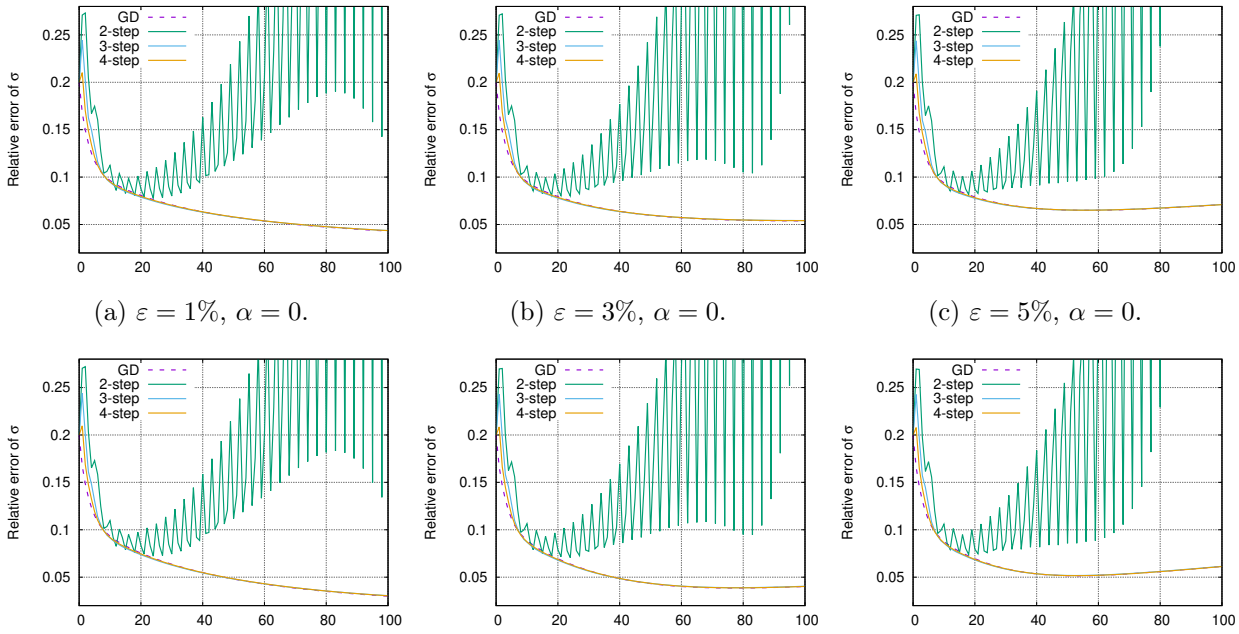
Figure 4: Meshes used to generate the matrices A_1, A_2 and H in the case of noisy data.



(a) $\varepsilon = 1\%$, $\alpha = 0$. (b) $\varepsilon = 1\%$, $\alpha = 0$. (c) $\varepsilon = 1\%$, $\alpha = 0$.

(d) $\varepsilon = 1\%$, $\alpha = 10^{-4}$. (e) $\varepsilon = 3\%$, $\alpha = 2 \cdot 10^{-4}$. (f) $\varepsilon = 5\%$, $\alpha = 2 \cdot 10^{-4}$.

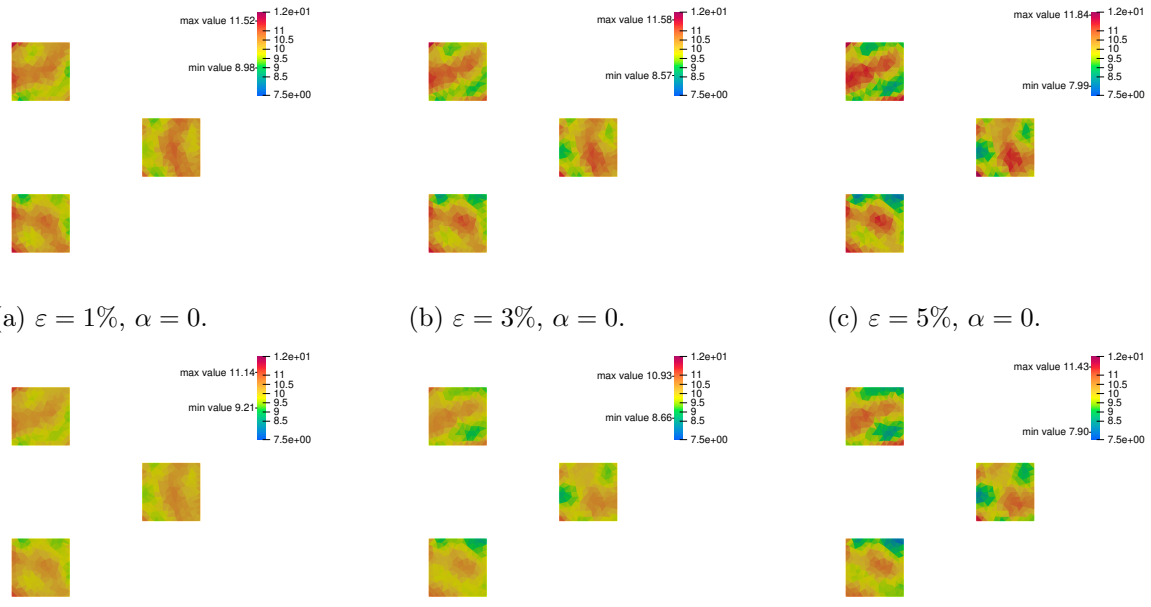
Figure 5: Convergence curves of semi-implicit gradient descent and k-step one-shot with different noise levels ε and regularization parameters α.



(a) $\varepsilon = 1\%$, $\alpha = 0$. (b) $\varepsilon = 3\%$, $\alpha = 0$. (c) $\varepsilon = 5\%$, $\alpha = 0$.

(d) $\varepsilon = 1\%$, $\alpha = 10^{-4}$. (e) $\varepsilon = 3\%$, $\alpha = 2 \cdot 10^{-4}$. (f) $\varepsilon = 5\%$, $\alpha = 2 \cdot 10^{-4}$.

Figure 6: Convergence curves for the relative error on the parameter σ of semi-implicit gradient descent and k-step one-shot with different noise levels ε and regularization parameters α.



(a) $\varepsilon = 1\%$, $\alpha = 0$. (b) $\varepsilon = 3\%$, $\alpha = 0$. (c) $\varepsilon = 5\%$, $\alpha = 0$.

(d) $\varepsilon = 1\%$, $\alpha = 10^{-4}$. (e) $\varepsilon = 3\%$, $\alpha = 2 \cdot 10^{-4}$. (f) $\varepsilon = 5\%$, $\alpha = 2 \cdot 10^{-4}$.

Figure 7: Reconstructed σ by semi-implicit 3-step one-shot with different noise levels ε.

References

[1] L. Audibert, H. Girardon, H. Haddar, and P. Jolivet. Inversion of eddy-current signals using a level-set method and block Krylov solvers. *Accepted for publication in SIAM Journal on Scientific Computing*, 2023.

[2] M. Bonazzoli, H. Haddar, and T.A. Vu. Convergence analysis of multi-step one-shot methods for linear inverse problems. Research Report RR-9477, Inria Saclay; ENSTA ParisTech, July 2022.

[3] M. Burger and W. Mühlhuber. Iterative regularization of parameter identification problems by sequential quadratic programming methods. *Inverse Problems*, 18:943–969, 2002.

[4] V. Dolean, P. Jolivet, and F. Nataf. *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015.

[5] N. Gauger, A. Griewank, A. Hamdi, C. Kratzenstein, E.Özkaya, and T. Slawig. Automated extension of fixed point PDE solvers for optimal design with bounded retardation. In *Constrained Optimization and Optimal Control for Partial Differential Equations, International Series of Numerical Mathematics*, pages 99–122. Springer Basel, 2012.

[6] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. Number 17 in Frontiers in Applied Mathematics. Soc. for Industrial and Applied Math, Philadelphia, 1997.

[7] A. Griewank. Projected Hessians for Preconditioning in One-Step One-Shot Design Optimization. In *Large-Scale Nonlinear Optimization*, volume 83, pages 151–171. Springer US, Boston, MA, 2006. Series Title: Nonconvex Optimization and Its Applications.

[8] S. Günther, N. R. Gauger, and Q. Wang. Simultaneous single-step one-shot optimization with unsteady PDEs. *Journal of Computational and Applied Mathematics*, 294:12–22, 2016.

[9] E. Haber and U. M. Ascher. Preconditioned all-at-once methods for large, sparse parameter estimation problems. *Inverse Problems*, 17(6):1847–1864, 2001.

[10] A. Hamdi and A. Griewank. Reduced quasi-Newton method for simultaneous design and optimization. *Computational Optimization and Applications*, 49(3):521–548, 2009.

[11] A. Hamdi and A. Griewank. Properties of an augmented Lagrangian for design optimization. *Optimization Methods and Software*, 25(4):645–664, 2010.

[12] S.B. Hazra, V. Schulz, J. Brezillon, and N.R. Gauger. Aerodynamic shape optimization using simultaneous pseudo-timestepping. *Journal of Computational Physics*, 204(1):46–64, 2005.

[13] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[14] B. Kaltenbacher, A. Kirchner, and B. Vexler. Goal oriented adaptivity in the IRGNM for parameter identification in PDEs II: all-at-once formulations. *Inverse Problems*, 30:045002, 2014.

[15] M. Marden. *Geometry of Polynomials*. Number 3 in Mathematical Surveys and Monographs. American Math. Soc, Providence, RI, 2nd edition, 1966.

[16] E. Özkaya and N. R. Gauger. Single-step One-shot Aerodynamic Shape Optimization. In *Optimal Control of Coupled Systems of Partial Differential Equations*, volume 158, pages 191–204. Birkhäuser Basel, Basel, 2009. Series Title: International Series of Numerical Mathematics.

[17] V. Schulz and I. Gherman. One-Shot Methods for Aerodynamic Shape Optimization. In *MEGADESIGN and MegaOpt - German Initiatives for Aerodynamic Simulation and Optimization in Aircraft Design*, volume 107, pages 207–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. Series Title: Notes on Numerical Fluid Mechanics and Multidisciplinary Design.

[18] A. Shenoy, M. Heinkenschloss, and E. M. Cliff. Airfoil design by an all-at-once method. *International Journal of Computational Fluid Dynamics*, 11(1-2):3–25, 1998.

[19] S. Ta'asan. "One Shot" Methods for Optimal Control of Distributed Parameter Systems I: Finite Dimensional Control. Technical Report 91-2, ICASE, Hampton, 1991.

[20] S. Ta'asan, G. Kuruvila, and M. Salas. Aerodynamic design and optimization in one shot. In *30th Aerospace Sciences Meeting and Exhibit*, Reno, NV, U.S.A., 1992. American Institute of Aeronautics and Astronautics.

[21] A. Tarantola and B. Valette. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics*, 20(2):219–232, 1982.

[22] P.-H. Tournier, I. Aliferis, M. Bonazzoli, M. de Buhan, M. Darbas, V. Dolean, F. Hecht, P. Jolivet, I. El Kanfoud, C. Migliaccio, F. Nataf, C. Pichot, and S. Semenov. Microwave tomographic imaging of cerebrovascular accidents by using high-performance computing. *Parallel Computing*, 85:88–97, 2019.

[23] T. van Leeuwen and F. J. Herrmann. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(1):661–667, 2013.

[24] T. van Leeuwen and F. J. Herrmann. A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, 32(1):015007, 2015.

A Some useful lemmas

We state auxiliary results about matrices like those appearing in the eigenvalue equations (18) and (43).

Lemma A.1. *Let* $(\mathbb{C}^{n\times n}, \|\cdot\|)$ *be a normed space and* $T \in \mathbb{C}^{n\times n}$. *If* $\rho(T) < 1$, *then*

$$\sum_{k=0}^{\infty} T^k \text{ converges and } \sum_{k=0}^{\infty} T^k = (I-T)^{-1}.$$

Moreover, if $\|T\| < 1$, $\|(I-T)^{-1}\| \leq \frac{1}{1-\|T\|}$.

Lemma A.2. *Let* $T \in \mathbb{C}^{n\times n}$ *such that* $\rho(T) < 1$. *Set*

$$s(T) \coloneqq \sup_{z\in\mathbb{C}, |z|\geq 1} \|(I - T/z)^{-1}\| \tag{65}$$

then $0 < \|(I-T)^{-1}\| \leq s(T) = s(T^*) < +\infty$. *Moreover, if* $\|T\| < 1$, $0 < s(T) \leq \dfrac{1}{1-\|T\|}$.

Proof. The existence of $s(T)$ (and also $s(T^*)$) is deduced from the fact that the functional $z \mapsto \|(I - T/z)^{-1}\|$, with $z \in \mathbb{C}, |z| \geq 1$, is well-defined and continuous. For every $z \in \mathbb{C}, |z| \geq 1$ we have

$$\|(I - T/z)^{-1}\| = \|\big((I - T/z)^{-1}\big)^*\| = \|(I - T^*/z^*)^{-1}\| \leq s(T^*)$$

and

$$\|(I - T^*/z)^{-1}\| = \|\big((I - T^*/z)^{-1}\big)^*\| = \|(I - T/z^*)^{-1}\| \le s(T),$$

thus $s(T) = s(T^*)$. The second part of conclusion is obtained by Lemma A.1. □

The following lemma says that, for $T \in \mathbb{C}^{n\times n}$ and $\lambda \in \mathbb{C}, |\lambda| \ge 1$, we can decompose

$$\left(I - \frac{T}{\lambda}\right)^{-1} = P(\lambda) + \mathrm{i}Q(\lambda) \quad \text{and} \quad \left(I - \frac{T^*}{\lambda}\right)^{-1} = P(\lambda)^* + \mathrm{i}Q(\lambda)^*$$

and gives bounds for $P(\lambda)$ and $Q(\lambda)$.

Lemma A.3. *Let $T \in \mathbb{C}^{n\times n}$ such that $\rho(T) < 1$ and $\lambda \in \mathbb{C}, |\lambda| \ge 1$. Write $\frac{1}{\lambda} = r(\cos\phi + \mathrm{i}\sin\phi)$ in polar form, where $0 < r \le 1$ and $\phi \in [-\pi, \pi]$. Then*

$$\left(I - \frac{T}{\lambda}\right)^{-1} = P(\lambda) + \mathrm{i}Q(\lambda) \quad \textit{and} \quad \left(I - \frac{T^*}{\lambda}\right)^{-1} = P(\lambda)^* + \mathrm{i}Q(\lambda)^*$$

where

$$P(\lambda) = (I - r\cos\phi\, T)(I - 2r\cos\phi\, T + r^2T^2)^{-1}, \quad Q(\lambda) = r\sin\phi\, T(I - 2r\cos\phi\, T + r^2T^2)^{-1}$$

are $\mathbb{C}^{n\times n}$-valued functions. We also have the following properties:

(i) $\|P(\lambda)\| \le (1 + \|T\|)\, s(T)^2$ *and* $\|Q(\lambda)\| \le |\sin\phi|\|T\|s(T)^2 \le \|T\|s(T)^2$.

(ii) Moreover if $\|T\| < 1$ *then*

$$\|P(\lambda)\| \le \frac{1}{1 - \|T\|} \quad \textit{and} \quad \|Q(\lambda)\| \le \frac{\|T\|}{1 - \|T\|}.$$

Proof. The first part of the lemma is verified by direct computation, using

$$(I - T/\lambda)^{-1} = (I - T/\lambda^*)\left[(I - T/\lambda)\,(I - T/\lambda^*)\right]^{-1},$$

$$(I - T^*/\lambda)^{-1} = \left[(I - T^*/\lambda^*)\,(I - T^*/\lambda)\right]^{-1}(I - T^*/\lambda^*)$$

and

$$(I - T/\lambda)\,(I - T/\lambda^*) = I - 2r\cos\phi\, T + r^2T^2.$$

After that, with the help of Lemma A.2, it is not difficult to show the inequalities in (i). To prove (ii), first observe that the two series

$$\sum_{k=0}^{\infty} r^k\cos(k\phi)T^k \quad \text{and} \quad \sum_{k=1}^{\infty} r^k\sin(k\phi)T^k$$

converge. Then, by expanding and simplifying the left-hand sides, we can show that

$$\left(\sum_{k=0}^{\infty} r^k\cos(k\phi)T^k\right)(I - 2r\cos\phi\, T + r^2T^2) = I - r\cos\phi\, T$$

and

$$\left(\sum_{k=1}^{\infty} r^k\sin(k\phi)T^k\right)(I - 2r\cos\phi\, T + r^2T^2) = r\sin\phi\, T$$

so $P(\lambda)$ and $Q(\lambda)$ can be expressed as the series above, and the inequalities in (ii) follow. □

In Sections 3.2 and 4.2 we identify different cases of $\lambda \in \mathbb{C}$ and we need corresponding estimations, given in the following lemma.

Lemma A.4. *For $\lambda \in \mathbb{C} \setminus \mathbb{R}, |\lambda| \geq 1$ we write $\lambda = R(\cos\theta + \mathrm{i}\sin\theta)$ in polar form where $R \geq 1$, $\theta \in (-\pi, \pi)$, $\theta \neq 0$.*

(i) For λ satisfying $\Re(\lambda^2 - \lambda) \geq 0$, let $\gamma_1 = \gamma_1(\lambda) := \begin{cases} 1 & \text{if } \Im(\lambda^2 - \lambda) \geq 0, \\ -1 & \text{if } \Im(\lambda^2 - \lambda) < 0 \end{cases}$ then

$$\Re(\lambda^2 - \lambda) + \gamma_1 \Im(\lambda^2 - \lambda) \geq |\lambda(\lambda - 1)| \geq 2|\sin(\theta/2)|.$$

(ii) Let $0 < \theta_0 \leq \frac{\pi}{4}$. For λ satisfying $\Re(\lambda^2 - \lambda) < 0$ and $\theta \in [\theta_0, \pi - \theta_0] \cup [-\pi + \theta_0, -\theta_0]$, let $\gamma_2 = \gamma_2(\lambda) := \begin{cases} -1 & \text{if } \Im(\lambda^2 - \lambda) \geq 0, \\ 1 & \text{if } \Im(\lambda^2 - \lambda) < 0 \end{cases}$ then

$$|\Re(\lambda^2 - \lambda) + \gamma_2 \Im(\lambda^2 - \lambda)| \geq |\lambda(\lambda - 1)| \geq 2\sin(\theta_0/2).$$

(iii) Let $0 < \theta_0 \leq \frac{\pi}{4}$ and $\delta_0 > 0$. For λ satisfying $\Re(\lambda^2 - \lambda) < 0$ and $\theta \in (-\theta_0, \theta_0)\backslash\{0\}$, let $\gamma_3 = \gamma_3(\mathrm{sign}(\theta)) := \begin{cases} \left(\delta_0 + \sin\frac{3\theta_0}{2}\right) / \cos\frac{3\theta_0}{2} & \text{if } \theta > 0, \\ -\left(\delta_0 + \sin\frac{3\theta_0}{2}\right) / \cos\frac{3\theta_0}{2} & \text{if } \theta < 0 \end{cases}$ then

$$\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda) \geq 2\delta_0 |\sin(\theta/2)|.$$

Moreover, if $0 < \theta_0 < \frac{\pi}{4}$ then

$$\frac{|\Re(\lambda - 1) + \gamma_3 \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} \leq \frac{\sqrt{1 + \gamma_3^2}}{\delta_0}$$

and

$$\frac{|\gamma_3 \Re(\lambda - 1) - \Im(\lambda - 1)|}{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)} \leq \max\left(\frac{\sqrt{1 + \gamma_3^2}}{\delta_0}, \frac{\sqrt{1 + \gamma_3^2}}{\cos 2\theta_0}\right).$$

(iv) Let $0 < \theta_0 \leq \frac{\pi}{4}$. There exists no λ satisfying $\Re(\lambda^2 - \lambda) < 0$ and $\theta \in (\pi - \theta_0, \pi) \cup (-\pi, -\pi + \theta_0)$.

Proof. (i) Notice that $\gamma_1^2 = 1$, $\gamma_1 \Im(\lambda^2 - \lambda) \geq 0$. We have

$$\begin{aligned} \left[\Re(\lambda^2 - \lambda) + \gamma_1 \Im(\lambda^2 - \lambda)\right]^2 &= \left[\Re(\lambda^2 - \lambda)\right]^2 + \left[\Im(\lambda^2 - \lambda)\right]^2 + 2\gamma_1 \Re(\lambda^2 - \lambda)\Im(\lambda^2 - \lambda) \\ &\geq \left[\Re(\lambda^2 - \lambda)\right]^2 + \left[\Im(\lambda^2 - \lambda)\right]^2 \\ &= |\lambda^2 - \lambda|^2, \end{aligned}$$

which yields $\Re(\lambda^2 - \lambda) + \gamma_1 \Im(\lambda^2 - \lambda) \geq |\lambda(\lambda - 1)|$. Finally,

$$|\lambda - 1| = |R\cos\theta - 1 + \mathrm{i}R\sin\theta| = \sqrt{R^2 + 1 - 2R\cos\theta} \geq \sqrt{2 - 2\cos\theta} = 2\left|\sin\frac{\theta}{2}\right|$$

since the function $R \mapsto R^2 + 1 - 2R\cos\theta$, for $R \geq 1$, is increasing.
(ii) In this case we have $\frac{\theta}{2} \in \left[\frac{\theta_0}{2}, \frac{\pi}{2} - \frac{\theta_0}{2}\right] \cup \left[-\frac{\pi}{2} + \frac{\theta_0}{2}, -\frac{\theta_0}{2}\right]$ so $\left|\sin\frac{\theta}{2}\right| \geq \sin\frac{\theta_0}{2}$. We also notice that $\gamma_2^2 = 1$ and $\gamma_2 \Im(\lambda^2 - \lambda) \leq 0$. Similar to (i), we have $|\Re(\lambda^2 - \lambda) + \gamma_2 \Im(\lambda^2 - \lambda)| = -\Re(\lambda^2 - \lambda) - \gamma_2 \Im(\lambda^2 - \lambda) \geq |\lambda(\lambda - 1)| \geq 2|\sin(\theta/2)|$, that implies the conclusion.

(iii) Note that $\cos 2\theta > 0, -\frac{\pi}{2} < 2\theta < \frac{\pi}{2}$, and $\sin 2\theta$ has the same sign as θ and γ_3, so we have

$$\begin{aligned}\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda) &= R(R\cos 2\theta-\cos\theta+\gamma_3 R\sin 2\theta-\gamma_3\sin\theta)\\ &\geq \cos 2\theta-\cos\theta+\gamma_3\sin 2\theta-\gamma_3\sin\theta\\ &= -2\sin\tfrac{3\theta}{2}\sin\tfrac{\theta}{2}+2\gamma_3\cos\tfrac{3\theta}{2}\sin\tfrac{\theta}{2}\\ &= 2\sin\tfrac{\theta}{2}\left(\gamma_3\cos\tfrac{3\theta}{2}-\sin\tfrac{3\theta}{2}\right).\end{aligned}$$

Then we consider two cases: if $0<\theta<\theta_0$ then $\gamma_3>0$, $\left|\sin\frac{\theta}{2}\right| = \sin\frac{\theta}{2} > 0$, $0 < \frac{3\theta}{2} < \frac{3\theta_0}{2} < \frac{\pi}{2}$ and $\gamma_3\cos\frac{3\theta}{2}-\sin\frac{3\theta}{2} > \gamma_3\cos\frac{3\theta_0}{2}-\sin\frac{3\theta_0}{2} = \delta_0$; if $-\theta_0<\theta<0$ then $-\gamma_3>0$, $\left|\sin\frac{\theta}{2}\right| = -\sin\frac{\theta}{2} > 0$, $-\frac{\pi}{2} < -\frac{3\theta_0}{2} < \frac{3\theta}{2} < 0$ and $-\gamma_3\cos\frac{3\theta}{2}+\sin\frac{3\theta}{2} > -\gamma_3\cos\frac{3\theta_0}{2}-\sin\frac{3\theta_0}{2} = \delta_0$.

Next, if $0<\theta_0<\frac{\pi}{4}$, we will show that $\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)}$ and $\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)}$ are both bounded. First,

$$\begin{aligned}\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} &= \frac{|(\cos\theta+\gamma_3\sin\theta)R-1|}{R[(\cos 2\theta+\gamma_3\sin 2\theta)R-(\cos\theta+\gamma_3\sin\theta)]}\\ &\leq \frac{|(\cos\theta+\gamma_3\sin\theta)R-1|}{(\cos 2\theta+\gamma_3\sin 2\theta)R-(\cos\theta+\gamma_3\sin\theta)}.\end{aligned}$$

Since γ_3 does not depend on R, let us study $f_1(R) \coloneqq \left(\frac{aR-1}{bR-a}\right)^2$ where $a \coloneqq \cos\theta+\gamma_3\sin\theta$ and $b \coloneqq \cos 2\theta+\gamma_3\sin 2\theta$. We observe that:

- $a>0$ and $b>0$. Indeed, $\cos\theta>0$, $\cos 2\theta>0$, and θ and γ_3 have the same sign.
- $bR-a>0$ since $\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)>0$, thus $R>\frac{a}{b}$.
- $a^2>b$ (equivalently $\frac{a}{b}>\frac{1}{a}$), since $a^2=\cos^2\theta+\gamma_3^2\sin^2\theta+\gamma_3\sin 2\theta > \cos^2\theta-\sin^2\theta+\gamma_3\sin 2\theta = b$.

Now, $f_1'(R) = 2\cdot\frac{aR-1}{bR-a}\cdot\frac{b-a^2}{(bR-a)^2} < 0$ for $R>\frac{a}{b}>\frac{1}{a}$ and we would like to have $\frac{a}{b}<1$ so that $f_1(R)\leq f_1(1), \forall R\geq 1$. Indeed $\frac{a}{b}<1$ is equivalent to

$$\cos\theta+\gamma_3\sin\theta < \cos 2\theta+\gamma_3\sin 2\theta \Leftrightarrow |\gamma_3| > \frac{\left|\sin\frac{3\theta}{2}\right|}{\cos\frac{3\theta}{2}},$$

which is true since

$$|\gamma_3| = \frac{\delta_0+\sin\frac{3\theta_0}{2}}{\cos\frac{3\theta_0}{2}} > \frac{\left|\sin\frac{3\theta}{2}\right|}{\cos\frac{3\theta}{2}}+\varepsilon_0 \quad \text{where} \quad \varepsilon_0 = \frac{\delta_0}{\cos\frac{3\theta_0}{2}}.$$

Then we study

$$f_1(1) = \left(\frac{\cos\theta-1+\gamma_3\sin\theta}{\cos 2\theta-\cos\theta+\gamma_3(\sin 2\theta-\sin\theta)}\right)^2 = \left(\frac{-\sin\frac{\theta}{2}+\gamma_3\cos\frac{\theta}{2}}{-\gamma_3\sin\frac{3\theta}{2}+\gamma_3^2\cos\frac{3\theta}{2}}\right)^2\gamma_3^2.$$

We have:

- $(-\sin\frac{\theta}{2}+\gamma_3\cos\frac{\theta}{2})^2 \leq 1+\gamma_3^2$ by Cauchy-Schwartz inequality;
- $\gamma_3^2 = |\gamma_3|^2 > \frac{\gamma_3\sin\frac{3\theta}{2}}{\cos\frac{3\theta}{2}}+\varepsilon_0|\gamma_3|$ that leads to $-\gamma_3\sin\frac{3\theta}{2}+\gamma_3^2\cos\frac{3\theta}{2} > \varepsilon_0\cos\frac{3\theta}{2}|\gamma_3| = \delta_0|\gamma_3|$;

hence $f_1(1) \leq \frac{1+\gamma_3^2}{\delta_0^2}$ and finally $\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} \leq \frac{\sqrt{1+\gamma_3^2}}{\delta_0}$. Next, we have

$$\begin{aligned}\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} &= \frac{|(\gamma_3\cos\theta-\sin\theta)R-\gamma_3|}{R[(\cos 2\theta+\gamma_3\sin 2\theta)R-(\cos\theta+\gamma_3\sin\theta)]}\\ &\leq \frac{|(\gamma_3\cos\theta-\sin\theta)R-\gamma_3|}{(\cos 2\theta+\gamma_3\sin 2\theta)R-(\cos\theta+\gamma_3\sin\theta)}.\end{aligned}$$

Since γ_3 does not depend on R, let us study $f_2(R) \coloneqq \left(\frac{cR-\gamma_3}{bR-a}\right)^2$ where $c \coloneqq \gamma_3\cos\theta-\sin\theta$ and a, b as above. We observe that:

- $\gamma_3 b - ca$ and θ have the same sign. Indeed, $\gamma_3 b - ca = (\gamma_3^2+1)\sin\theta\cos\theta$. Consequently, we always have $(\gamma_3 b - ca)\gamma_3 > 0$.
- We always have $\frac{\gamma_3}{c} > 1$. Indeed, if $\theta > 0$ then $c > 0$ since $\gamma_3 = \frac{\delta_0+\sin\frac{3\theta_0}{2}}{\cos\frac{3\theta_0}{2}} > \frac{\sin\theta}{\cos\theta}$, also $\frac{\gamma_3}{c} = \frac{\gamma_3}{\gamma_3\cos\theta-\sin\theta} > 1$; if $\theta < 0$ then $c < 0$ since $-\gamma_3 = \frac{\delta_0+\sin\frac{3\theta_0}{2}}{\cos\frac{3\theta_0}{2}} > -\frac{\sin\theta}{\cos\theta}$, also $\frac{\gamma_3}{c} = \frac{-\gamma_3}{-\gamma_3\cos\theta+\sin\theta} > 1$.

Now, $f_2'(R) = 2\cdot\frac{\frac{c}{\gamma_3}R-1}{bR-a}\cdot\frac{(\gamma_3 b-ca)\gamma_3}{(bR-a)^2}$, so, thanks to the above results, $f_2(R)$ decreases for $1 \leq R < \frac{\gamma_3}{c}$ and increases for $R > \frac{\gamma_3}{c}$. Moreover, like for $f_1(1)$, we can estimate

$$f_2(1) = \left(\frac{-\cos\frac{\theta}{2}-\gamma_3\sin\frac{\theta}{2}}{-\gamma_3\sin\frac{3\theta}{2}+\gamma_3^2\cos\frac{3\theta}{2}}\right)^2\gamma_3^2 \leq \frac{1+\gamma_3^2}{\delta_0^2}$$

and $\lim_{R\to+\infty} f_2(R) = \left(\frac{\gamma_3\cos\theta-\sin\theta}{\cos 2\theta+\gamma_3\sin 2\theta}\right)^2 \leq \frac{1+\gamma_3^2}{\cos 2\theta_0}$. Therefore

$$\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^2-\lambda)+\gamma_3\Im(\lambda^2-\lambda)} \leq \max\left(\frac{\sqrt{1+\gamma_3^2}}{\delta_0}, \frac{\sqrt{1+\gamma_3^2}}{\cos 2\theta_0}\right).$$

(iv) For $\theta \in (\pi-\theta_0, \pi)\cup(-\pi, -\pi+\theta_0)$, we have $\cos 2\theta > 0$ since $2\theta \in \left(\frac{3\pi}{2}, 2\pi\right)\cup\left(-2\pi, -\frac{3\pi}{2}\right)$, while $\cos\theta < 0$. Hence $\Re(\lambda^2-\lambda) = R(R\cos 2\theta-\cos\theta) > 0$. □