Convergence analysis of multi-step one-shot methods for linear inverse problems

Marcella Bonazzoli, Houssem Haddar, Tuan Anh Vu

Convergence analysis of multi-step one-shot methods for linear inverse problems

Marcella Bonazzoli, Houssem Haddar, Tuan Anh Vu

RESEARCH REPORT
N° 9477
July 2022
Project-Teams IDEFIX

ISSN 0249-6399 ISRN INRIA/RR--9477--FR+ENG

Convergence analysis of multi-step one-shot methods for linear inverse problems

Marcella Bonazzoli*, Houssem Haddar*, Tuan Anh Vu*

Project-Teams IDEFIX

Research Report n° 9477 — July 2022 — 55 pages

Abstract: In this work we are interested in general linear inverse problems where the corresponding forward problem is solved iteratively using fixed point methods. Then one-shot methods, which iterate at the same time on the forward problem solution and on the inverse problem unknown, can be applied. We analyze two variants of the so-called multi-step one-shot methods and establish sufficient conditions on the descent step for their 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 usual and shifted gradient descent. In particular, we observe that very few inner iterations on the forward problem are enough to guarantee good convergence of the inversion algorithm.

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

* Inria, UMA, ENSTA Paris, Institut Polytechnique de Paris

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Analyse de convergence pour des méthodes d'inversion multi-étapes de type one-shot

Résumé : Dans ce travail nous nous intéressons à des problèmes inverses linéaires généraux où le problème direct correspondant est résolu de façon itérative en utilisant des méthodes de point fixe. Ainsi, les méthodes de type one-shot, qui itèrent en même temps sur la solution du problème direct et l'inconnue du problème inverse, peuvent être appliquées. Nous considérons deux variantes des méthodes multi-étapes de type one-shot et nous établissons des conditions suffisantes et nécessaires sur le pas de descente pour leur convergence, en étudiant les valeurs propres de la matrice par blocs des itérations couplées. Plusieurs tests numériques sont présentés pour illustrer la convergence de ces méthodes par rapport aux méthodes de descente de gradient usuelle et décentrée. En particulier, nous observons que très peu d'itérations internes pour le problème direct sont suffisantes pour garantir une bonne convergence de l'algorithme d'inversion.

Mots-clés : problèmes inverses, méthodes de type one-shot, analyse de convergence, identification de paramètres

Contents

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. Indeed, the corresponding linear systems could be too large to be handled with direct solvers (e.g. LU-type solvers), and iterative solvers are easier to parallelize on many cores. Naturally this leads to the idea of *one-step one-shot methods*, which 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 [22] 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 [23, 21, 11, 19, 18] and the literature review in the introduction of [19]. All-at-once approaches to inverse problems for parameter identification were studied in, e.g., [8, 2, 15]. An alternative method, called Wavefield Reconstruction Inversion (WRI), was introduced for seismic imaging in [25], as an improvement of the classical Full Waveform Inversion (FWI) [24]. 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 [26].

Few convergence proofs, especially for the multi-step one-shot methods, are available in literature. In particular, for non-linear design optimization problems, Griewank [6] 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 [9, 10, 4] 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. In [7] this particular one-step one-shot approach was extended to time-dependent problems.

In this work, we consider two variants of multi-step one-shot methods where the forward and adjoint problems are solved using fixed point methods and the inverse problem is solved using gradient descent methods. 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 inverse problem uniqueness 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 matrices. We prove that if the descent step is small enough then the considered multi-step one-shot methods converge. Moreover, the upper bounds for the descent step in these sufficient conditions are explicit in the number of inner iterations and in the norms of

the operators involved in the problem. In the particular scalar case (Appendix C), we establish 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. 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 both real and complex eigenvalues. In Section 5 we show that the previous analysis can be extended to the case where the state variable is complex. Finally, in Section 6 we test numerically the performance of the different algorithms on a toy 2D Helmholtz inverse problem.

Throughout this work, $\langle \cdot, \cdot \rangle$ indicates the usual Hermitian scalar product in $\mathbb{C}^n$, that is $\langle x, y \rangle \coloneqq \overline{y}^\intercal x, \forall x, y \in \mathbb{C}^n$, 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 $T \in \mathbb{C}^{n \times n}$ with $\rho(T) < 1$, we define

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

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: find $u \equiv u(\sigma)$ 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 of the direct problem, obtained after discretization, with parameter σ. Note that in the non-linear case B would be a function of σ. Equation (1) is also called *state equation* and u is called *state*. Given σ, we can solve for u by a fixed point iteration

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

which converges for any initial guess u_0 if and only if the spectral radius $\rho(B)$ is strictly less than 1 (see e.g. [5, Theorem 2.1.1]). Hence we assume $\rho(B) < 1$. Now, we measure $f = Hu(\sigma)$, where $H \in \mathbb{R}^{n_f \times n_u}$, and we are interested in the *linear inverse problem* of finding σ from f. In order to guarantee the uniqueness of the inverse problem, we assume that $H(I - B)^{-1}M$ is injective. In summary, we set

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

with the assumptions:

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

To solve the inverse problem we write its least squares formulation: given σ^{ex} the exact solution of the inverse problem and $f \coloneqq Hu(\sigma^{\text{ex}})$,

$$\sigma^{\text{ex}} = \operatorname{argmin}_{\sigma \in \mathbb{R}^{n_\sigma}} J(\sigma) \quad \text{where } J(\sigma) \coloneqq \frac{1}{2} \|Hu(\sigma) - f\|^2 .$$

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

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

and allows us to compute the gradient of the cost functional

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

The classical gradient descent algorithm then reads

$$\textbf{usual gradient descent:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n, \\ u^n = Bu^n + M\sigma^n + F, \\ p^n = B^* p^n + H^*(Hu^n - f), \end{cases} \tag{5}$$

where $\tau > 0$ is the descent step size, and the state and adjoint state equations are solved exactly by a direct solver. Here $\sigma^{n+1} = \sigma^n - \tau \nabla J(\sigma_n)$; if instead we update $\sigma^{n+1} = \sigma^n - \tau \nabla J(\sigma_{n-1})$, we obtain the

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

Both algorithms converge for sufficiently small τ (see e.g. Appendix B): for any initial guess, (5) converges if

$$\tau < \frac{2}{\|H(I-B)^{-1}M\|^2}, \tag{7}$$

and (6) converges if

$$\tau < \frac{1}{\|H(I-B)^{-1}M\|^2}. \tag{8}$$

Here, 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, the solution to the inverse problem. We update $\sigma^{n+1} = \sigma^n - \tau M^* p^n$ as in 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 *coupled*:

$$\begin{cases} u_{\ell+1}^{n+1} = Bu_\ell^{n+1} + M\sigma + F, \\ p_{\ell+1}^{n+1} = B^* p_\ell^{n+1} + H^*(Hu_\ell^{n+1} - f), \end{cases} \qquad \ell = 0, 1, \dots, k, \qquad \begin{cases} u^{n+1} = u_k^{n+1}, \\ p^{n+1} = p_k^{n+1} \end{cases}$$

where σ depends on the considered variant ($\sigma = \sigma^{n+1}$ or, for the shifted methods, $\sigma = \sigma^n$). As initial guess we naturally choose $u_0^{n+1} = u^n$ and $p_0^{n+1} = p^n$, the information from the previous (outer) step. In summary, we have two multi-step one-shot algorithms

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

and

$$\textbf{shifted } k\textbf{-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n, \\ u_0^{n+1} = u^n, p_0^{n+1} = p^n, \\ \left| \begin{array}{l} u_{\ell+1}^{n+1} = B u_\ell^{n+1} + M\sigma^n + F, \\ p_{\ell+1}^{n+1} = B^* p_\ell^{n+1} + H^*(H u_\ell^{n+1} - f), \end{array} \right. \\ u^{n+1} = u_k^{n+1}, p^{n+1} = p_k^{n+1}, \end{cases} \tag{10}$$

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

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

and

$$\textbf{shifted one-step one-shot:} \quad \begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n, \\ u^{n+1} = B u^n + M\sigma^n + F \\ p^{n+1} = B^* p^n + H^*(H u^n - f). \end{cases} \tag{12}$$

The only difference for the shifted versions lies in the fact that σ^n is used in (10) and (12), instead of σ^{n+1} in (9) and (11), so that in (9) and (11) we need to wait for σ before updating u and p, while in (10) and (12) we can update σ, u, p at the same time. Also note that when $k \to \infty$, the k-step one-shot method (9) formally converges to the usual gradient descent (5), while the shifted k-step one-shot method (10) formally converges to the shifted gradient descent (6).

We first analyze the one-step one-shot methods ($k = 1$) in Section 3 and then the multi-step one-shot methods ($k \geq 2$) in Section 4.

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

3.1 Block iteration matrices and eigenvalue equations

To analyze the convergence of these methods, first we 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

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B u^n + M\sigma^n - \tau M M^* p^n + F \\ p^{n+1} = B^* p^n + H^* H u^n - H^* f. \end{cases} \tag{13}$$

System (12) is already in the form we need. In what follows we first study the shifted 1-step one-shot method, then the 1-step one-shot method.

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 the errors satisfy: for the shifted algorithm (12)

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B u^n + M\sigma^n \\ p^{n+1} = B^* p^n + H^* H u^n \end{cases} \tag{14}$$

and for algorithm (13)

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = Bu^n + M\sigma^n - \tau M M^* p^n \\ p^{n+1} = B^* p^n + H^* H u^n, \end{cases} \tag{15}$$

or equivalently, by putting in evidence the block iteration matrices

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} B^* & H^*H & 0 \\ 0 & B & M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix} \tag{16}$$

and

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} B^* & H^*H & 0 \\ -\tau M M^* & B & M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix}. \tag{17}$$

Now recall that a fixed point iteration converges if and only if the spectral radius of its iteration matrix is strictly less than 1. Therefore in the following propositions we establish eigenvalue equations for the iteration matrix of the two methods.

Proposition 3.1 (Eigenvalue equation for the shifted 1-step one-shot method)**.** *Assume that $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (16).*

(i) If $\lambda \in \mathbb{C}$, $\lambda \notin \mathrm{Spec}(B)$, then $\exists\, y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that

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

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix.

Remark 3.2. Since $\rho(B)$ is strictly less than 1, so is $\rho(B^*)$.

Proof. Since $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (16), there exists a non-zero vector $(\tilde{p}, \tilde{u}, y) \in \mathbb{C}^{n_u + n_u + n_\sigma}$ such that

$$\begin{cases} \lambda y = y - \tau M^* \tilde{p} \\ \lambda \tilde{u} = B\tilde{u} + My \\ \lambda \tilde{p} = B^* \tilde{p} + H^* H \tilde{u}. \end{cases} \tag{19}$$

By the second equation in (19) $\tilde{u} = (\lambda I - B)^{-1} M y$, so together with the third equation

$$\tilde{p} = (\lambda I - B^*)^{-1} H^* H \tilde{u} = (\lambda I - B^*)^{-1} H^* H (\lambda I - B)^{-1} M y,$$

and by inserting this result into the first equation we obtain

$$(\lambda - 1) y = -\tau M^* (\lambda I - B^*)^{-1} H^* H (\lambda I - B)^{-1} M y, \tag{20}$$

that gives (18) by taking the scalar product with y. We also see that if $y = 0$ then the above formulas for $\tilde{u}, \tilde{p}$ immediately give $\tilde{u} = \tilde{p} = 0$, that is a contradiction.
(ii) Assume that $\lambda = 1$ is an eigenvalue of the iteration matrix, then (20) gives us

$$M^*(I - B^*)^{-1} H^* H (I - B)^{-1} M y = 0,$$

but this cannot happen for $y \neq 0$ due to the injectivity of $H(I - B)^{-1} M$. □

Proposition 3.3 (Eigenvalue equation for the 1-step one-shot method)**.** *Assume that $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (17).*

(i) If $\lambda \in \mathbb{C}$, $\lambda \notin \mathrm{Spec}(B)$ then $\exists\, y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that:

$$(\lambda - 1)\,\|y\|^2 + \tau\lambda\langle M^*(\lambda I - B^*)^{-1}H^*H(\lambda I - B)^{-1}My, y\rangle = 0. \tag{21}$$

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix.

Proof. Since $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (17), there exists a non-zero vector $(\tilde{p}, \tilde{u}, y) \in \mathbb{C}^{n_u+n_u+n_\sigma}$ such that

$$\begin{cases} \lambda y = y - \tau M^*\tilde{p} \\ \lambda\tilde{u} = B\tilde{u} + My - \tau MM^*\tilde{p} \\ \lambda\tilde{p} = B^*\tilde{p} + H^*H\tilde{u}. \end{cases} \tag{22}$$

By the third equation in (22) $\tilde{p} = (\lambda I - B^*)^{-1}H^*H\tilde{u}$, and inserting this result into the second equation we obtain

$$\lambda\tilde{u} = B\tilde{u} + My - \tau MM^*(\lambda I - B^*)^{-1}H^*H\tilde{u},$$

or equivalently,

$$[I + \tau MM^*A](\lambda I - B)\tilde{u} = My$$

where $A = (\lambda I - B^*)^{-1}H^*H(\lambda I - B)^{-1}$. Since $\tau > 0$, $I + \tau MM^*A$ is a positive definite matrix. Therefore

$$\tilde{u} = (\lambda I - B)^{-1}[I + \tau MM^*A]^{-1}My$$

and

$$\tilde{p} = (\lambda I - B^*)^{-1}H^*H\tilde{u} = A[I + \tau MM^*A]^{-1}My.$$

By inserting this result into the first equation in (22) we obtain

$$(\lambda - 1)y = -\tau M^*A[I + \tau MM^*A]^{-1}My.$$

Thanks to the fact that $[I + \tau MM^*A]^{-1}$ and MM^*A commute, we have

$$(\lambda - 1)My = -\tau MM^*A[I + \tau MM^*A]^{-1}My = -\tau[I + \tau MM^*A]^{-1}MM^*AMy$$

then

$$(\lambda - 1)[I + \tau MM^*A]My = -\tau MM^*AMy,$$

that leads to

$$(\lambda - 1)My + \tau\lambda MM^*AMy = 0.$$

Since $H(I - B)^{-1}M$ is injective, so is M. Therefore

$$(\lambda - 1)y + \tau\lambda M^*AMy = 0, \tag{23}$$

that gives (21) by taking scalar product with y. We also see that if $y = 0$ then the above formulas for $\tilde{u}, \tilde{p}$ immediately give $\tilde{u} = \tilde{p} = 0$, that is a contradiction.
(ii) Assume that $\lambda = 1$ is an eigenvalue of the iteration matrix, then (23) gives us

$$M^*(I - B^*)^{-1}H^*H(I - B)^{-1}My = 0,$$

but this cannot happen for $y \neq 0$ due to the injectivity of $H(I - B)^{-1}M$. □

In the following sections we will show that, for sufficiently small τ, equations (18) and (21) admit no solution $|\lambda| \geq 1$, thus algorithms (12) and (11) converge. When $\lambda \neq 0$, it is convenient to rewrite (18) and (21) respectively as

$$\lambda^2(\lambda-1)\left\|y\right\|^2 + \tau\langle M^*\left(I-B^*/\lambda\right)^{-1}H^*H\left(I-B/\lambda\right)^{-1}My, y\rangle = 0 \tag{24}$$

and

$$\lambda(\lambda-1)\left\|y\right\|^2 + \tau\langle M^*\left(I-B^*/\lambda\right)^{-1}H^*H\left(I-B/\lambda\right)^{-1}My, y\rangle = 0. \tag{25}$$

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

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

Proposition 3.4 (shifted 1-step one-shot method)**.** *When $B = 0$, the eigenvalue equation (24) admits no solution $\lambda \in \mathbb{C}, |\lambda| \geq 1$ if $\tau < \frac{-1+\sqrt{5}}{2\|H\|^2\|M\|^2}$.*

Proof. When $B = 0$, equation (24) becomes $\lambda^2(\lambda-1)\left\|y\right\|^2 + \tau\left\|HMy\right\|^2 = 0$ which is equivalent to $\lambda^3 - \lambda^2 + \frac{\|HMy\|^2}{\|y\|^2}\tau = 0$. Then, the conclusion can be obtained by Lemma C.1. □

Proposition 3.5 (1-step one-shot method)**.** *When $B = 0$, the eigenvalue equation (25) admits no solution $\lambda \in \mathbb{C}, |\lambda| \leq 1$ if $\tau < \frac{1}{\|H\|^2\|M\|^2}$.*

Proof. When $B = 0$, equation (25) becomes $\lambda(\lambda-1)\left\|y\right\|^2 + \tau\left\|HMy\right\|^2 = 0$ which yields $\lambda^3 - \lambda^2 + \frac{\|HMy\|^2}{\|y\|^2}\tau\lambda = 0$. Then, the conclusion can be obtained by Lemma C.1. □

3.2 Real eigenvalues

We now find conditions on the descent step τ such that the real eigenvalues stay inside the unit disk. Recall that we have already proved that $\lambda = 1$ is not an eigenvalue for both methods.

Proposition 3.6 (shifted 1-step one-shot method)**.** *Equation (24)*

(i) admits no solution $\lambda \in \mathbb{R}, \lambda > 1$ for all $\tau > 0$;

(ii) admits no solution $\lambda \in \mathbb{R}, \lambda \leq -1$ if we take

$$\tau < \frac{2}{\left\|H\right\|^2\left\|M\right\|^2 s(B)^2},$$

where $s(B)$ is defined in Lemma A.2; moreover if $0 < \|B\| < 1$, we can take

$$\tau < \frac{\chi_0(1,\|B\|)}{\left\|H\right\|^2\left\|M\right\|^2}, \quad \text{where } \chi_0(1,b) = 2(1-b)^2 \tag{26}$$

(here in the notation $\chi_0(1,b)$, 1 refers to $k = 1$).

Proof. When $\lambda \in \mathbb{R}\backslash\{0\}$ equation (24) becomes

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

The left-hand side of the above equation is strictly positive for any $\tau > 0$ if $\lambda > 1$; it is strictly negative for τ satisfying the inequality in (ii) if $\lambda \leq -1$, noting that $\lambda \mapsto \lambda^2(\lambda-1)$ is increasing for $\lambda \leq -1$. □

Proposition 3.7 (1-step one-shot method)**.** *Equation* (25) *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 (25) becomes

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

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

3.3 Complex eigenvalues

We now look for conditions on the descent step τ such that also the complex eigenvalues stay inside the unit disk. We first deal with the shifted 1-step one-shot method.

Proposition 3.8 (shifted 1-step one-shot method)**.** *If* $B \neq 0$, $\exists \tau > 0$ *sufficiently small such that equation* (24) *admits no solution* $\lambda \in \mathbb{C}\backslash\mathbb{R}, |\lambda| \geq 1$. *In particular, if* $0 < \|B\| < 1$, *given any* $\delta_0 > 0$ *and* $0 < \theta_0 \leq \frac{\pi}{6}$, *take*

$$\tau < \frac{\min\{\chi_1(1, \|B\|),\ \chi_2(1, \|B\|),\ \chi_3(1, \|B\|),\ \chi_4(1, \|B\|)\}}{\|H\|^2 \|M\|^2},$$

where

$$\chi_1(1, b) = \frac{(1-b)^4}{4b^2}, \quad \chi_2(1, b) = \frac{2 \sin\frac{\theta_0}{2}(1-b)^2}{(1+b)^2},$$

$$\chi_3(1, b) = \frac{\delta_0 \cos^2\frac{5\theta_0}{2}}{2\left(1 + 2\delta_0 \sin\frac{5\theta_0}{2} + \delta_0^2\right)} \cdot \frac{(1-b)^4}{b^2}, \quad \chi_4(1, b) = \left[\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0\right](1-b)^2$$

(here in the notation $\chi_i(1, b), i = 1, \ldots, 4$, 1 *refers to* $k = 1$*).*

Proof. **Step 1. Rewrite equation** (24) **so that we can study its real and imaginary parts.**

Let $\lambda = R(\cos\theta + \mathrm{i}\sin\theta)$ in polar form where $R = |\lambda| \geq 1$ and $\theta \in (-\pi, \pi)$. 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 $\mathbb{C}^{n_u \times n_u}$-valued functions, and, by omitting the dependence on λ,

$$\|P\| \leq p := \begin{cases} (1 + \|B\|)s(B)^2 & \text{for general } B \neq 0, \\ \dfrac{1}{1 - \|B\|} & \text{when } \|B\| < 1; \end{cases} \tag{27}$$

$$\|Q\| \leq q_1 := \begin{cases} \|B\|\, s(B)^2 & \text{for general } B \neq 0, \\ \dfrac{\|B\|}{1 - \|B\|} & \text{when } 0 < \|B\| < 1; \end{cases} \tag{28}$$

$$\|Q\| \leq |\sin\theta| q_2, \quad q_2 := \begin{cases} \|B\|\, s(B)^2 & \text{for general } B \neq 0, \\ \dfrac{\|B\|}{(1 - \|B\|)^2} & \text{when } 0 < \|B\| < 1. \end{cases} \tag{29}$$

Now we rewrite (24) as

$$\lambda^2(\lambda-1)\left\|y\right\|^2+\tau G(P^*+\mathrm{i}Q^*,P+\mathrm{i}Q)=0 \tag{30}$$

where

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

G satisfies the following properties:

- $\forall X,Y_1,Y_2\in\mathbb{C}^{n_u\times n_u},\forall z_1,z_2\in\mathbb{C}$: $\quad G(X,z_1Y_1+z_2Y_2)=z_1G(X,Y_1)+z_2G(X,Y_2)$.
- $\forall X_1,X_2,Y\in\mathbb{C}^{n_u\times n_u},\forall z_1,z_2\in\mathbb{C}$: $\quad G(z_1X_1+z_2X_2,Y)=z_1G(X_1,Y)+z_2G(X_2,Y)$.
- $\forall X\in\mathbb{C}^{n_u\times n_u}$: $\quad 0\le G(X^*,X)=\left\|HXMy\right\|^2\le(\left\|H\right\|\left\|M\right\|\left\|X\right\|)^2\left\|y\right\|^2$.
- $\forall X,Y\in\mathbb{C}^{n_u\times n_u}$: $\quad G(X,Y)+G(Y^*,X^*)\in\mathbb{R}$, indeed

$$\begin{aligned}G(X,Y)&=\langle M^*XH^*HYMy,y\rangle=\langle y,M^*Y^*H^*HX^*My\rangle\\&=\langle M^*Y^*H^*HX^*My,y\rangle^*=G(Y^*,X^*)^*.\end{aligned}$$

With these properties of G, we expand (30) and take its real and imaginary parts, so we respectively obtain:

$$\Re(\lambda^3-\lambda^2)\left\|y\right\|^2+\tau[G(P^*,P)-G(Q^*,Q)]=0 \tag{31}$$

and

$$\Im(\lambda^3-\lambda^2)\left\|y\right\|^2+\tau[G(P^*,Q)+G(Q^*,P)]=0 \tag{32}$$

Step 2. Find a suitable combination of equations (31) and (32), choose τ so that we obtain a new equation with a left-hand side which is strictly positive/negative.

Let $\gamma=\gamma(\lambda)\in\mathbb{R}$, defined by cases as in Lemma A.4. Multiplying equation (32) with γ then summing it with equation (31), we obtain:

$$[\Re(\lambda^3-\lambda^2)+\gamma\Im(\lambda^3-\lambda^2)]\left\|y\right\|^2+\tau[G(P^*,P)-G(Q^*,Q)+\gamma G(P^*,Q)+\gamma G(Q^*,P)]=0,$$

or equivalently,

$$[\Re(\lambda^3-\lambda^2)+\gamma\Im(\lambda^3-\lambda^2)]\left\|y\right\|^2+\tau G(P^*+\gamma Q^*,P+\gamma Q)-(1+\gamma^2)\tau G(Q^*,Q)=0. \tag{33}$$

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

- *Case 1.* $\Re(\lambda^3-\lambda^2)\ge 0$;
- *Case 2.* $\Re(\lambda^3-\lambda^2)<0$ and $\theta\in[\theta_0,\pi-\theta_0]\cup[-\pi+\theta_0,-\theta_0]$ for fixed $0<\theta_0\le\frac{\pi}{6}$;
- *Case 3.* $\Re(\lambda^3-\lambda^2)<0$ and $\theta\in(-\theta_0,\theta_0)$ for fixed $0<\theta_0\le\frac{\pi}{6}$;
- *Case 4.* $\Re(\lambda^3-\lambda^2)<0$ and $\theta\in(\pi-\theta_0,\pi)\cup(-\pi,-\pi+\theta_0)$ for fixed $0<\theta_0\le\frac{\pi}{6}$.

The four cases will be treated in the following four lemmas (Lemmas 3.9–3.12), which together give the statement of this proposition. □

Lemma 3.9 (Case 1)**.** *Equation* (24) *admits no solutions* λ *in Case 1 if we take*

$$\tau<\frac{1}{4\left\|H\right\|^2\left\|M\right\|^2\left\|B\right\|^2 s(B)^4}.$$

Moreover, if $0<\|B\|<1$*, we can take*

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

Proof. Writing (33) for $\gamma = \gamma_1$ as in Lemma A.4 (i) (in particular $\gamma_1^2 = 1$), we have

$$[\Re(\lambda^3 - \lambda^2) + \gamma_1 \Im(\lambda^3 - \lambda^2)] \|y\|^2 + \tau G(P^* + \gamma_1 Q^*, P + \gamma_1 Q) - 2\tau G(Q^*, Q) = 0. \tag{34}$$

By the properties of G we have

$$G(P^* + \gamma_1 Q^*, P + \gamma_1 Q) \geq 0$$

and

$$G(Q^*, Q) \leq (\|H\| \|M\| \|Q\|)^2 \|y\|^2 \leq (\|H\| \|M\| |\sin\theta| q_2)^2 \|y\|^2,$$

therefore the left-hand side of (34) will be strictly positive if τ satisfies

$$\tau < \frac{\Re(\lambda^3 - \lambda^2) + \gamma_1 \Im(\lambda^3 - \lambda^2)}{2 (\|H\| \|M\| |\sin\theta| q_2)^2}.$$

Since $\Re(\lambda^3 - \lambda^2) + \gamma_1 \Im(\lambda^3 - \lambda^2) \geq 2|\sin(\theta/2)|$ by Lemma A.4 (i), it is enough to choose

$$\tau < \frac{1}{4 \left|\sin\frac{\theta}{2}\right| \cos^2\frac{\theta}{2} \|H\|^2 \|M\|^2 q_2^2}.$$

Since $\left|\sin\frac{\theta}{2}\right| \cos^2\frac{\theta}{2} \leq 1$, it is sufficient to choose $\tau < \frac{1}{4\|H\|^2\|M\|^2 q_2^2}$ and we use definition (29) of q_2. □

Lemma 3.10 (Case 2)**.** *Equation* (24) *admits no solutions* λ *in Case 2 if we take*

$$\tau < \frac{2 \sin\frac{\theta_0}{2}}{\|H\|^2 \|M\|^2 (1 + 2\|B\|)^2 s(B)^4}.$$

Moreover, if $0 < \|B\| < 1$*, we can take*

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

Proof. Writing (33) for $\gamma = \gamma_2$ as in Lemma A.4 (ii) (in particular $\gamma_2^2 = 1$), we have

$$[\Re(\lambda^3 - \lambda^2) + \gamma_2 \Im(\lambda^3 - \lambda^2)] \|y\|^2 + \tau G(P^* + \gamma_2 Q^*, P + \gamma_2 Q) - 2\tau G(Q^*, Q) = 0. \tag{35}$$

By the properties of G

$$G(Q^*, Q) \geq 0, \quad G(P^* + \gamma_2 Q^*, P + \gamma_2 Q) \leq (\|H\| \|M\| \|P + \gamma_2 Q\|)^2 \|y\|^2$$

and the estimate $\|P + \gamma_2 Q\| \leq \|P\| + |\gamma_2| \|Q\| = \|P\| + \|Q\| \leq p + q_1$, the left-hand side of (35) will be strictly negative if τ satisfies:

$$\tau < \frac{-\Re(\lambda^3 - \lambda^2) - \gamma_2 \Im(\lambda^3 - \lambda^2)}{[\|H\| \|M\| (p + q_1)]^2}.$$

Thanks to Lemma A.4 (ii), it is sufficient to choose

$$\tau < \frac{2 \sin\frac{\theta_0}{2}}{\|H\|^2 \|M\|^2 (p + q_1)^2}$$

and we use definitions (27) and (28) of p and q_1. □

Lemma 3.11 (Case 3). *Let $\delta_0 > 0$ be fixed. Equation (24) admits no solutions λ in Case 3 if we take*

$$\tau < \frac{\delta_0 \cos^2 \frac{5\theta_0}{2}}{2\left(1 + 2\delta_0 \sin \frac{5\theta_0}{2} + \delta_0^2\right)} \cdot \frac{1}{\|H\|^2 \|M\|^2 \|B\|^2 s(B)^4}.$$

Moreover, if $0 < \|B\| < 1$, we can take

$$\tau < \frac{\delta_0 \cos^2 \frac{5\theta_0}{2}}{2\left(1 + 2\delta_0 \sin \frac{5\theta_0}{2} + \delta_0^2\right)} \cdot \frac{(1 - \|B\|)^4}{\|H\|^2 \|M\|^2 \|B\|^2}.$$

Proof. Writing (33) for $\gamma = \gamma_3$ as in Lemma A.4 (iii), we have

$$[\Re(\lambda^3 - \lambda^2) + \gamma_3 \Im(\lambda^3 - \lambda^2)] \|y\|^2 + \tau G(P^* + \gamma_3 Q^*, P + \gamma_3 Q) - (1 + \gamma_3^2)\tau G(Q^*, Q) = 0. \quad (36)$$

By the properties of G

$$G(P^* + \gamma_3 Q^*, P + \gamma_3 Q) \geq 0, \quad G(Q^*, Q) \leq (\|H\| \|M\| \|Q\|)^2 \|y\|^2$$

and by the estimate $\|Q\| \leq |\sin\theta| q_2$, the left-hand side of (36) will be strictly positive if τ satisfies:

$$\tau < \frac{\Re(\lambda^2 - \lambda) + \gamma_3 \Im(\lambda^2 - \lambda)}{(1 + \gamma_3^2)\left(\|H\| \|M\| |\sin\theta| q_2\right)^2}.$$

Since by Lemma A.4 (iii) $\Re(\lambda^3 - \lambda^2) + \gamma_3 \Im(\lambda^3 - \lambda^2) > 2\delta_0 \left|\sin \frac{\theta}{2}\right|$, it is sufficient to choose

$$\tau < \frac{\delta_0}{2(1 + \gamma_3^2) \|H\|^2 \|M\|^2 q_2^2} = \frac{1}{2 \|H\|^2 \|M\|^2 q_2^2} \cdot \frac{\delta_0 \cos^2 \frac{5\theta_0}{2}}{1 + 2\delta_0 \sin \frac{5\theta_0}{2} + \delta_0^2},$$

where we have used the definition of γ_3. To conclude we use definition (29) of q_2. □

Lemma 3.12 (Case 4). *Equation (24) admits no solutions λ in Case 4 if we take*

$$\tau < \frac{\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0}{\|H\|^2 \|M\|^2 (1 + \|B\|)^2 s(B)^2}.$$

Moreover, if $0 < \|B\| < 1$, we can take

$$\tau < \left[\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0\right] \frac{(1 - \|B\|)^2}{\|H\|^2 \|M\|^2}.$$

Proof. Here it is enough to consider (31). By the properties of G

$$G(Q^*, Q) \geq 0, \quad G(P^*, P) \leq (\|H\| \|M\| p)^2 \|y\|^2$$

we see that the left-hand side of (31) will be strictly negative if τ satisfies

$$\tau < \frac{-\Re(\lambda^3 - \lambda^2)}{(\|H\| \|M\| p)^2}.$$

Thanks to Lemma A.4 (iv), it is sufficient to choose

$$\tau < \frac{\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0}{\|H\|^2 \|M\|^2 p^2},$$

and definition (27) of p leads to the conclusion. □

Similarly, with the help of Lemma A.5, we prove for the 1-step one-shot method the analogue of Proposition 3.8. In particular, note that here just three cases of λ need to be considered, because the analogue of the fourth one is excluded by Lemma A.5 (iv).

Proposition 3.13 (1-step one-shot method)**.** *If $B \neq 0$, $\exists \tau > 0$ sufficiently small such that equation (25) admits no solution $\lambda \in \mathbb{C}\backslash\mathbb{R}, |\lambda| \geq 1$. In particular, if $0 < \|B\| < 1$, given any $\delta_0 > 0$ and $0 < \theta_0 \leq \frac{\pi}{4}$, take*

$$\tau < \frac{\min\{\psi_1(1,\|B\|),\ \psi_2(1,\|B\|),\ \psi_3(1,\|B\|)\}}{\|H\|^2 \|M\|^2},$$

where

$$\psi_1(1,b) = \frac{(1-b)^4}{4b^2}, \quad \psi_2(1,b) = \frac{2\sin\frac{\theta_0}{2}(1-b)^2}{(1+b)^2}, \quad \psi_3(1,b) = \frac{\delta_0 \cos^2\frac{3\theta_0}{2}(1-b)^4}{2\left(1+2\delta_0\sin\frac{3\theta_0}{2}+\delta_0^2\right)b^2}$$

(here in the notation $\psi_i(1,b), i = 1,2,3$, 1 refers to $k = 1$).

3.4 Final result ($k = 1$)

Considering Proposition 3.4, and taking the minimum between the bound (26) in Proposition 3.6 for real eigenvalues and the bound in Proposition 3.8 for complex eigenvalues, we obtain a sufficient condition on the descent step τ to ensure convergence of the shifted 1-step one-shot method.

Theorem 3.14 (Convergence of shifted 1-step one-shot)**.** *Under assumption (4), the shifted 1-step one-shot method (12) converges for sufficiently small τ. In particular, for $\|B\| < 1$, it is enough to take*

$$\tau < \frac{\chi(1,\|B\|)}{\|H\|^2 \|M\|^2},$$

where $\chi(1,\|B\|)$ is an explicit function of $\|B\|$ (in this notation 1 refers to $k = 1$).

Remark 3.15. Set $b = \|B\|$. For $0 < b < 1$, a practical (but not optimal) bound for τ is

$$\tau < \frac{1}{\|H\|^2 \|M\|^2} \cdot \min\left\{\frac{1}{2}\cdot\frac{(1-b)^2}{(1+b)^2},\ \frac{1-\sin\frac{5\pi}{12}}{4}\cdot\frac{(1-b)^4}{b^2}\right\}.$$

Indeed, using the notation in Proposition 3.6 and 3.8, it is easy to show that $\chi_2(1,b) \leq \chi_0(1,b)$ and $\chi_3(1,b) \leq \chi_1(1,b)$. By studying $\chi_3(1,b)$ and noting that $\delta_0^2 + 1 \geq 2\delta_0$, we see that we should take $\delta_0 = 1$. Finally, we can take for instance $\theta_0 = \frac{\pi}{6}$, then compare $\chi_2(1,b)$, $\chi_3(1,b)$ and $\chi_4(1,b)$.

Putting together Propositions 3.5, 3.7, 3.13, we obtain a sufficient condition on the descent step τ to ensure convergence of the 1-step one-shot method.

Theorem 3.16 (Convergence of 1-step one-shot)**.** *Under assumption (4), the 1-step one-shot method (11) converges for sufficiently small τ. In particular, for $\|B\| < 1$, it is enough to take*

$$\tau < \frac{\psi(1,\|B\|)}{\|H\|^2 \|M\|^2},$$

where $\psi(1,\|B\|)$ is an explicit function of $\|B\|$ (in this notation 1 refers to $k = 1$).

Remark 3.17. Similarly as above, for $0 < b < 1$, a practical (but not optimal) bound for τ is

$$\tau < \frac{1}{\|H\|^2 \|M\|^2} \cdot \min\left\{2\sin\frac{\pi}{8}\cdot\frac{(1-b)^2}{(1+b)^2},\ \frac{1-\sin\frac{3\pi}{8}}{4}\cdot\frac{(1-b)^4}{b^2}\right\}.$$

4 Convergence of multi-step one-shot methods ($k \geq 2$)

We now tackle the multi-step case, that is the k-step one-shot methods with $k \geq 2$.

4.1 Block iteration matrices and eigenvalue equations

Once again, to analyze the convergence of these methods, first we express $(\sigma^{n+1}, u^{n+1}, p^{n+1})$ in terms of (σ^n, u^n, p^n), by rewriting the recursions for u and p: systems (9) and (10) are respectively rewritten as

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B^k u^n + T_k M \sigma^n - \tau T_k M M^* p^n + T_k F \\ p^{n+1} = [(B^*)^k - \tau X_k M M^*] p^n + U_k u^n + X_k M \sigma^n + X_k F - T_k^* H^* f \end{cases} \tag{37}$$

and

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B^k u^n + T_k M \sigma^n + T_k F \\ p^{n+1} = (B^*)^k p^n + U_k u^n + X_k M \sigma^n + X_k F - T_k^* H^* f \end{cases} \tag{38}$$

where

$$T_k = I + B + ... + B^{k-1} = (I - B)^{-1}(I - B^k), \quad k \geq 1, \tag{39}$$
$$U_k = (B^*)^{k-1} H^* H + (B^*)^{k-2} H^* H B + ... + H^* H B^{k-1}, \quad k \geq 1,$$

$$X_k = \begin{cases} (B^*)^{k-2} H^* H T_1 + (B^*)^{k-3} H^* H T_2 + ... + H^* H T_{k-1} & \text{if } k \geq 2, \\ 0 & \text{if } k = 1. \end{cases} \tag{40}$$

Note that (37) (k-step one-shot) can be obtained from (38) (shifted k-step one-shot) by replacing σ^n with $\sigma^{n+1} = \sigma^n - \tau M^* p^n$ in the equations for u and p, which yields two extra terms in (37). In what follows we first study the shifted k-step one-shot method then the k-step one-shot method. The following lemma gathers 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

$$U_k T_k - X_k B^k + X_k = T_k^* H^* H T_k, \quad \forall k \geq 1. \tag{41}$$

Proof. (i) is easy to check by the definitions. (ii) follows from (i).
(iii) For $k = 1$, we have $U_1 = H^* H$, $T_1 = I$ and $X_1 = 0$, hence the identity is verified. For $k \geq 2$, note that $X_{k+1} = B^* X_k + H^* H T_k$, then by (ii) $X_{k+1} = X_{k+1}^* = X_k B + T_k^* H^* H$. 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 designate them by (σ^n, u^n, p^n). We obtain that the errors satisfy: for the shifted algorithm (38)

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B^k u^n + T_k M \sigma^n \\ p^{n+1} = (B^*)^k p^n + U_k u^n + X_k M \sigma^n \end{cases} \tag{42}$$

and for algorithm (37)

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau M^* p^n \\ u^{n+1} = B^k u^n + T_k M \sigma^n - \tau T_k M M^* p^n \\ p^{n+1} = [(B^*)^k - \tau X_k M M^*] p^n + U_k u^n + X_k M \sigma^n, \end{cases} \tag{43}$$

or equivalently, by putting in evidence the block iteration matrices

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} (B^*)^k & U_k & X_k M \\ 0 & B^k & T_k M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix} \tag{44}$$

and

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} (B^*)^k - \tau X_k M M^* & U_k & X_k M \\ -\tau T_k M M^* & B^k & T_k M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix}. \tag{45}$$

Now recall that a fixed point iteration converges if and only if the spectral radius of its iteration matrix is strictly less than 1. Therefore in the following propositions we establish eigenvalue equations for the iteration matrix of the two methods.

Proposition 4.2 (Eigenvalue equation for the shifted k-step one-shot method)**.** *Assume that $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (44).*

(i) If $\lambda \in \mathbb{C}$, $\lambda \notin \text{Spec}(B^k)$, then $\exists\, y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that

$$(\lambda - 1)\|y\|^2 + \tau \langle M^*[\lambda I - (B^*)^k]^{-1}[(\lambda-1)X_k + T_k^* H^* H T_k](\lambda I - B^k)^{-1} M y, y\rangle = 0. \tag{46}$$

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix.

Proposition 4.3 (Eigenvalue equation for the k-step one-shot method)**.** *Assume that $\lambda \in \mathbb{C}$ is an eigenvalue of the iteration matrix in (45).*

(i) If $\lambda \in \mathbb{C}$, $\lambda \notin \text{Spec}(B^k)$ then $\exists\, y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that:

$$(\lambda - 1)\|y\|^2 + \tau\lambda \langle M^*[\lambda I - (B^*)^k]^{-1}[(\lambda-1)X_k + T_k^* H^* H T_k](\lambda I - B^k)^{-1} M y, y\rangle = 0. \tag{47}$$

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix.

Remark 4.4. Since $\rho(B)$ is strictly less than 1, so are $\rho(B^*), \rho(B^k)$ and $\rho((B^*)^k)$.

The proofs for Propositions 4.2 and 4.3 are respectively similar to the ones of Propositions 3.1 and 3.3, the slight difference is that in the calculation we use (41) to simplify some terms.

In the following sections we will show that, for sufficiently small τ, equations (46) and (47) admit no solution $|\lambda| \geq 1$, thus algorithms (10) and (9) converge. When $\lambda \neq 0$, it is convenient to rewrite (46) and (47) respectively as

$$\lambda^2(\lambda-1)\left\|y\right\|^2 + \tau\langle M^*\left[I-(B^*)^k/\lambda\right]^{-1}\left[(\lambda-1)X_k + T_k^*H^*HT_k\right]\left(I-B^k/\lambda\right)^{-1}My, y\rangle = 0 \quad (48)$$

and

$$\lambda(\lambda-1)\left\|y\right\|^2 + \tau\langle M^*\left[I-(B^*)^k/\lambda\right]^{-1}\left[(\lambda-1)X_k + T_k^*H^*HT_k\right]\left(I-B^k/\lambda\right)^{-1}My, y\rangle = 0 \quad (49)$$

The scalar case where $n_u, n_\sigma, n_f = 1$ is analyzed in Appendix C.

Remark 4.5. Note that when $B = 0$ and $k \geq 2$, the shifted k-step one-shot and k-step one-shot are respectively equivalent to the shifted and usual gradient descent methods, therefore we retrieve the same bounds (8)–(7) for the descent step τ as for those methods.

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

Lemma 4.6. *If* $\|B\| < 1$,

$$s(B^k) \leq \frac{1}{1-\|B\|^k}, \quad \|T_k\| \leq \frac{1-\|B\|^k}{1-\|B\|}, \quad \|X_k\| \leq \frac{\|H\|^2\left(1-k\|B\|^{k-1}+(k-1)\|B\|^k\right)}{(1-\|B\|)^2}.$$

Proof. The bound for $s(B^k)$ is proved using Lemma A.2 and $\left\|B^k\right\| \leq \|B\|^k$. Next, from (39) we have

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

From (40), if $k \geq 2$ we have

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

□

4.2 Real eigenvalues

We first find conditions on the descent step τ 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 4.7 (shifted k-step one-shot method)**.** *When $k \geq 2$, $\exists \tau > 0$ sufficiently small such that equation (48) admits no solution $\lambda \in \mathbb{R}, \lambda \neq 1, |\lambda| \geq 1$. More precisely, take*

- $\tau < \frac{2}{\|M\|^2\left(\|H\|^2\|T_k\|^2+2\|X_k\|\right)s(B^k)^2}$ *if the denominator of the right-hand side is not* 0*;*
- *any $\tau > 0$ otherwise.*

Moreover, if $\|B\| < 1$, we can take

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

Proof. When $\lambda \in \mathbb{R}$ equation (48) is rewritten as

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

We show that if $\lambda > 1$ (or respectively $\lambda \leq -1$) we can choose τ so that the left-hand side of the above equation is strictly positive (or respectively negative). Indeed, if $\lambda > 1$, we choose τ such that

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

and this can be done by taking τ such that

$$[\|X_k\|\|M\|^2 s(B^k)^2]\tau<1.$$

If $\lambda \leq -1$, we choose τ such that

$$\begin{aligned}&\lambda^2(\lambda-1)\|y\|^2+\tau\left\|HT_k\left(I-\tfrac{B^k}{\lambda}\right)^{-1}My\right\|^2\\&+\tau(1-\lambda)\left|\langle M^*\left[I-\tfrac{(B^*)^k}{\lambda}\right]^{-1}X_k\left(I-\tfrac{B^k}{\lambda}\right)^{-1}My,y\rangle\right| \quad<0\end{aligned}$$

and this can be done by taking τ such that

$$\left[\frac{\|H\|^2\|T_k\|^2\|M\|^2 s(B^k)^2}{2}+\|X_k\|\|M\|^2 s(B^k)^2\right]\tau<1,$$

so we obtain the first conclusion. Finally, the second conclusion in the case $\|B\| < 1$ can be obtained by Lemma 4.6. □

Proposition 4.8 (k-step one-shot method)**.** *When $k \geq 2$, $\exists \tau > 0$ sufficiently small such that equation (49) admits no solution $\lambda \in \mathbb{R}, \lambda \neq 1, |\lambda| \geq 1$. More precisely, take*

- $\tau < \dfrac{1}{\|X_k\|\|M\|^2 s(B^k)^2}$ *if the denominator of the right-hand side is not* 0*;*
- *any $\tau > 0$ otherwise.*

Moreover, if $\|B\| < 1$, we can take

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

Proof. When $\lambda \in \mathbb{R}$ equation (49) is rewritten as

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

We show that we can choose τ so that the left-hand side of the above equation is strictly positive. Indeed, if $\lambda > 1$, we choose τ such that

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

and this can be done by taking τ such that

$$\|X_k\| \|M\|^2 s(B^k)^2 \tau < 1.$$

If $\lambda \leq -1$, we choose τ such that

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

and this is also done by taking τ such that

$$\|X_k\| \|M\|^2 s(B^k)^2 \tau < 1.$$

so we obtain the first conclusion. Finally, the conclusion in the case $\|B\| < 1$ can be obtained by Lemma 4.6. □

4.3 Complex eigenvalues

We now look for conditions on the descent step τ such that also the complex eigenvalues stay inside the unit disk. We first deal with the shifted k-step one-shot method.

Proposition 4.9 (shifted k-step one-shot method)**.** *When $k \geq 2$, $\exists \tau > 0$ sufficiently small such that equation (48) admits no solution $\lambda \in \mathbb{C} \backslash \mathbb{R}$, $|\lambda| \geq 1$. In particular, if $\|B\| < 1$, given any $\delta_0 > 0$ and $0 < \theta_0 < \frac{\pi}{6}$, take*

$$\tau < \frac{\min\{\chi_1(k, \|B\|),\ \chi_2(k, \|B\|),\ \chi_3(k, \|B\|),\ \chi_4(k, \|B\|)\}}{\|H\|^2 \|M\|^2}$$

where

$$\chi_1(k,b) = \frac{(1-b)^2(1-b^k)^2}{4b^{2k} + \sqrt{2}(1 - kb^{k-1} + (k-1)b^k)(1+b^k)^2}$$

$$\chi_2(k,b) = \frac{(1-b)^2(1-b^k)^2}{\left[\frac{1}{2\sin(\theta_0/2)}(1-b^k)^2 + \sqrt{2}(1 - kb^{k-1} + (k-1)b^k)\right](1+b^k)^2}$$

$$\chi_3(k,b) = \frac{(1-b)^2(1-b^k)^2}{\frac{2c\sin(\theta_0/2)}{\delta_0} b^{2k} + (1 - kb^{k-1} + (k-1)b^k)\left[\frac{\sqrt{c}}{\delta_0}(1+b^{2k}) + 2\max\left(\frac{\sqrt{c}}{\delta_0}, \frac{\sqrt{c}}{\cos 3\theta_0}\right) b^k\right]}$$

$$\chi_4(k,b) = \frac{\left[\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0\right](1-b)^2(1-b^k)^2}{(1-b^k)^2 + 2(1 - kb^{k-1} + (k-1)b^k)(1+b^k)^2}$$

and $c = \frac{1 + 2\delta_0 \sin\frac{5\theta_0}{2} + \delta_0^2}{\cos^2\frac{5\theta_0}{2}}$.

Proof. **Step 1. Rewrite equation (48) so that we can study its real and imaginary parts.**

Let $\lambda = R(\cos\theta + \mathrm{i}\sin\theta)$ in polar form where $R = |\lambda| \geq 1$ and $\theta \in (-\pi, \pi)$. 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

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

where $P(\lambda)$ and $Q(\lambda)$ are $\mathbb{C}^{n_u \times n_u}$-valued functions, and, by omitting the dependence on λ,

$$\|P\| \leq p \coloneqq \begin{cases} (1 + \left\|B^k\right\|)s(B^k)^2 & \text{for general } B, \\ \dfrac{1}{1 - \|B\|^k} & \text{when } \|B\| < 1; \end{cases} \tag{50}$$

$$\|Q\| \leq q_1 \coloneqq \begin{cases} \left\|B^k\right\| s(B^k)^2 & \text{for general } B, \\ \dfrac{\|B\|^k}{1 - \|B\|^k} & \text{when } \|B\| < 1; \end{cases} \tag{51}$$

$$\|Q\| \leq q_2|\sin\theta|, \quad q_2 \coloneqq \begin{cases} \left\|B^k\right\| s(B^k)^2 & \text{for general } B, \\ \dfrac{\|B\|^k}{(1 - \|B\|^k)^2} & \text{when } \|B\| < 1. \end{cases} \tag{52}$$

Now we rewrite (48) as

$$\lambda^2(\lambda - 1)\|y\|^2 + \tau G(P^* + \mathrm{i}Q^*, P + \mathrm{i}Q) + \tau(\lambda - 1)L(P^* + \mathrm{i}Q^*, P + \mathrm{i}Q) = 0. \tag{53}$$

where

$$G(X, Y) = \langle M^* X T_k^* H^* H T_k Y M y, y\rangle, \quad L(X, Y) = \langle M^* X X_k Y M y, y\rangle$$

for $X, Y \in \mathbb{C}^{n_u \times n_u}$. G satisfies the following properties:

- $\forall X, Y_1, Y_2 \in \mathbb{C}^{n_u \times n_u}, \forall z_1, z_2 \in \mathbb{C}$: $\quad G(X, z_1Y_1 + z_2Y_2) = z_1G(X, Y_1) + z_2G(X, Y_2)$.
- $\forall X_1, X_2, Y \in \mathbb{C}^{n_u \times n_u}, \forall z_1, z_2 \in \mathbb{C}$: $\quad G(z_1X_1 + z_2X_2, Y) = z_1G(X_1, Y) + z_2G(X_2, Y)$.
- $\forall X \in \mathbb{C}^{n_u \times n_u}$: $\quad G(X^*, X) \in \mathbb{R}$.
- $\forall X, Y \in \mathbb{C}^{n_u \times n_u}$: $\quad G(X, Y) + G(Y^*, X^*) \in \mathbb{R}$, indeed

$$\begin{aligned} G(X, Y) &= \langle M^* X T_k^* H^* H T_k Y M y, y\rangle = \langle y, M^* Y^* T_k^* H^* H T_k X^* M y\rangle \\ &= \langle M^* Y^* T_k^* H^* H T_k X^* M y, y\rangle^* = G(Y^*, X^*)^*. \end{aligned}$$

Similarly, L has the same properties as G (note that $X_k^* = X_k$ by Lemma 4.1). With these properties of G and L, we expand (53) and take its real and imaginary parts, so we respectively obtain:

$$\Re(\lambda^3 - \lambda^2)\|y\|^2 + \tau G_1 + \tau[\Re(\lambda - 1)L_1 - \Im(\lambda - 1)L_2] = 0 \tag{54}$$

and

$$\Im(\lambda^3 - \lambda^2)\|y\|^2 + \tau G_2 + \tau[\Im(\lambda - 1)L_1 + \Re(\lambda - 1)L_2] = 0 \tag{55}$$

where

$$G_1 = G(P^*, P) - G(Q^*, Q), \quad G_2 = G(P^*, Q) + G(Q^*, P),$$

$$L_1 = L(P^*, P) - L(Q^*, Q), \quad L_2 = L(P^*, Q) + L(Q^*, P).$$

Step 2. Find a suitable combination of equations (54) and (55), choose τ so that we obtain a new equation with a left-hand side which is strictly positive/negative.

Let $\gamma = \gamma(\lambda) \in \mathbb{R}$, defined by cases as in Lemma A.4. Multiplying equation (55) with γ then summing it with equation (54), we obtain:

$$\begin{aligned}&[\Re(\lambda^3 - \lambda^2) + \gamma\Im(\lambda^3 - \lambda^2)] \|y\|^2 + \tau G(P^* + \gamma Q^*, P + \gamma Q) - (1+\gamma^2)\tau G(Q^*, Q) \\ &+\tau\left([\Re(\lambda - 1) + \gamma\Im(\lambda - 1)]L_1 + [\gamma\Re(\lambda - 1) - \Im(\lambda - 1)]L_2\right) \qquad = 0.\end{aligned} \tag{56}$$

Now we prepare some useful estimates.

- $\forall X \in \mathbb{C}^{n_u \times n_u}$: $\quad 0 \le G(X^*, X) = \|HT_k X M y\|^2 \le (\|H\| \|T_k\| \|M\| \|X\|)^2 \|y\|^2$.
  Since $\|Q\| \le q_1$ and $\|Q\| \le q_2 |\sin\theta|$, we have

  $$G(Q^*, Q) \le (\|H\| \|T_k\| \|M\| q_1)^2 \|y\|^2 \text{ and } G(Q^*, Q) \le (\|H\| \|T_k\| \|M\| q_2 \sin|\theta|)^2 \|y\|^2.$$

- By Cauchy-Schwarz inequality we have

  $$|\Re(\lambda - 1) + \gamma\Im(\lambda - 1)| \le \sqrt{1+\gamma^2}|\lambda - 1|; \quad |\gamma\Re(\lambda - 1) - \Im(\lambda - 1)| \le \sqrt{1+\gamma^2}|\lambda - 1|.$$

- $\forall X, Y \in \mathbb{C}^{n_u \times n_u}$: $\quad |L(X, Y)| = |\langle M^* X X_k Y M y, y\rangle| \le \|X_k\| \|M\|^2 \|X\| \|Y\| \|y\|^2$. Hence

  $$\begin{aligned}|L_1| &= |L(P^*, P) - L(Q^*, Q)| \le |L(P^*, P)| + |L(Q^*, Q)| \\ &\le \|X_k\| \|M\|^2 (\|P\|^2 + \|Q\|^2) \|y\|^2 \le \|X_k\| \|M\|^2 (p^2 + q_1^2) \|y\|^2,\end{aligned}$$

  $$\begin{aligned}|L_2| &= |L(P^*, Q) + L(Q^*, P)| \le |L(P^*, Q)| + |L(Q^*, P)| \\ &\le 2\|X_k\| \|M\|^2 \|P\| \|Q\| \|y\|^2 \le 2\|X_k\| \|M\|^2 p q_1 \|y\|^2,\end{aligned}$$

  and then

  $$\begin{aligned}&|[\Re(\lambda - 1) + \gamma\Im(\lambda - 1)]L_1 + [\gamma\Re(\lambda - 1) - \Im(\lambda - 1)]L_2| \\ \le\ &|\Re(\lambda - 1) + \gamma\Im(\lambda - 1)||L_1| + |\gamma\Re(\lambda - 1) - \Im(\lambda - 1)||L_2| \\ \le\ &\sqrt{1+\gamma^2}|\lambda - 1| \|X_k\| \|M\|^2 (p^2 + q_1^2 + 2pq_1) \|y\|^2 \\ =\ &\sqrt{1+\gamma^2}|\lambda - 1| \|X_k\| \|M\|^2 (p + q_1)^2 \|y\|^2.\end{aligned}$$

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

- *Case 1.* $\Re(\lambda^3 - \lambda^2) \ge 0$;
- *Case 2.* $\Re(\lambda^3 - \lambda^2) < 0$ and $\theta \in [\theta_0, \pi - \theta_0] \cup [-\pi + \theta_0, -\theta_0]$ for fixed $0 < \theta_0 < \frac{\pi}{6}$;
- *Case 3.* $\Re(\lambda^3 - \lambda^2) < 0$ and $\theta \in (-\theta_0, \theta_0)$ for fixed $0 < \theta_0 < \frac{\pi}{6}$;
- *Case 4.* $\Re(\lambda^3 - \lambda^2) < 0$ and $\theta \in (\pi - \theta_0, \pi) \cup (-\pi, -\pi + \theta_0)$ for fixed $0 < \theta_0 < \frac{\pi}{6}$.

The four cases will be treated in the following four lemmas (Lemmas 4.10–4.13), which together give the statement of this proposition. □

Lemma 4.10 (Case 1)**.** *For $k \ge 2$, equation (48) admits no solutions λ in Case 1 if we take*

- $\tau < \dfrac{s(B^k)^{-4}}{4\|H\|^2 \|M\|^2 \|T_k\|^2 \|B^k\|^2 + \sqrt{2}\|M\|^2 \|X_k\| (1 + 2\|B^k\|)^2}$ *if the denominator of the right-hand side is not 0;*

- *any* $\tau > 0$ *otherwise.*

Moreover, if $\|B\| < 1$*, we can take*

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

Proof. Writing (56) for $\gamma = \gamma_1$ as in Lemma A.4 (i) (in particular $\gamma_1^2 = 1$), we have

$$\begin{aligned}&[\Re(\lambda^3-\lambda^2)+\gamma_1\Im(\lambda^3-\lambda^2)]\|y\|^2 + \tau G(P^*+\gamma_1 Q^*, P+\gamma_1 Q) - 2\tau G(Q^*,Q)\\ &+\tau\left([\Re(\lambda-1)+\gamma_1\Im(\lambda-1)]L_1 + [\gamma_1\Re(\lambda-1)-\Im(\lambda-1)]L_2\right) = 0.\end{aligned} \tag{57}$$

Since $G(P^*+\gamma_1 Q^*, P+\gamma_1 Q) \geq 0$, and by estimating

$$G(Q^*,Q) \leq (\|H\|\,\|T_k\|\,\|M\|\,q_2\sin|\theta|)^2\|y\|^2,$$

$$[\Re(\lambda-1)+\gamma_1\Im(\lambda-1)]L_1 + [\gamma_1\Re(\lambda-1)-\Im(\lambda-1)]L_2 \geq -\sqrt{2}|\lambda-1|\,\|X_k\|\,\|M\|^2(p+q_1)^2\|y\|^2,$$

by Lemma A.4 (i) the left-hand side of (57) will be strictly positive if τ satisfies:

$$\left(2(\|H\|\,\|T_k\|\,\|M\|\,q_2)^2\frac{|\sin\theta|^2}{|\lambda-1|} + \sqrt{2}\,\|X_k\|\,\|M\|^2(p+q_1)^2\right)\tau < 1.$$

Since $\frac{|\sin\theta|^2}{|\lambda-1|} \leq \frac{|\sin\theta|^2}{2|\sin(\theta/2)|} = 2\left|\sin\frac{\theta}{2}\right|\cos^2\frac{\theta}{2} \leq 2$, we have the first part of the conclusion using definitions (50), (51), (52) of p, q_1, q_2. Finally, the conclusion in the case $\|B\| < 1$ can be obtained by Lemma 4.6. □

Lemma 4.11 (Case 2)**.** *For* $k \geq 2$*, equation* (48) *admits no solutions* λ *in Case 2 if we take*

- $\tau < \dfrac{s(B^k)^{-4}}{\left(\frac{1}{2\sin(\theta_0/2)}\|H\|^2\|M\|^2\|T_k\|^2 + \sqrt{2}\|M\|^2\|X_k\|\right)(1+2\|B^k\|)^2}$ *if the denominator of the right-hand side is not* 0*;*
- *any* τ *otherwise.*

Moreover, if $\|B\| < 1$*, we can take*

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

Proof. Writing (56) for $\gamma = \gamma_2$ as in Lemma A.4 (ii) (in particular $\gamma_2^2 = 1$), we have

$$\begin{aligned}&[\Re(\lambda^3-\lambda^2)+\gamma_2\Im(\lambda^3-\lambda^2)]\|y\|^2 + \tau G(P^*+\gamma_2 Q^*, P+\gamma_2 Q) - 2\tau G(Q^*,Q)\\ &+\tau\left([\Re(\lambda-1)+\gamma_2\Im(\lambda-1)]L_1 + [\gamma_2\Re(\lambda-1)-\Im(\lambda-1)]L_2\right) = 0.\end{aligned} \tag{58}$$

Since $G(Q^*,Q) \geq 0$, and by estimating $\|P+\gamma_2 Q\| \leq \|P\| + |\gamma_2|\,\|Q\| = \|P\| + \|Q\| \leq p+q_1$, so that

$$G(P^*+\gamma_2 Q^*, P+\gamma_2 Q) \leq [\|H\|\,\|T_k\|\,\|M\|\,(p+q_1)]^2\|y\|^2,$$

and

$$[\Re(\lambda-1)+\gamma_2\Im(\lambda-1)]L_1+[\gamma_2\Re(\lambda-1)-\Im(\lambda-1)]L_2 \\ \leq \sqrt{2}|\lambda-1|\,\|X_k\|\,\|M\|^2\,(p+q_1)^2\,\|y\|^2\,,$$

by Lemma A.4 (ii), the left-hand side of (58) will be strictly negative if τ satisfies:

$$\left([\|H\|\,\|T_k\|\,\|M\|\,(p+q_1)]^2\,\frac{1}{|\lambda-1|}+\sqrt{2}\,\|X_k\|\,\|M\|^2\,(p+q_1)^2\right)\tau<1.$$

Since $\frac{1}{|\lambda-1|}\leq\frac{1}{2\sin(\theta_0/2)}$, we have the first part of the conclusion using definitions (50), (51) of p,q_1. Finally, the conclusion in the case $\|B\|<1$ can be obtained by Lemma 4.6. □

Lemma 4.12 (Case 3). *Let $\delta_0>0$ be fixed and $c\coloneqq\frac{1+2\delta_0\sin\frac{5\theta_0}{2}+\delta_0^2}{\cos^2\frac{5\theta_0}{2}}$. For $k\geq 2$, equation (48) admits no solutions λ in Case 3 if we take*

- $\tau<s(B^k)^{-4}\Big/\left[\frac{2c\sin\frac{\theta_0}{2}}{\delta_0}\,\|H\|^2\,\|M\|^2\,\|T_k\|^2\left\|B^k\right\|^2+\frac{\sqrt{c}}{\delta_0}\,\|M\|^2\,\|X_k\|\,(1+2\left\|B^k\right\|+2\left\|B^k\right\|^2)\right.$
  $\left.+2\max\left(\frac{\sqrt{c}}{\delta_0},\frac{\sqrt{c}}{\cos 3\theta_0}\right)\|M\|^2\,\|X_k\|\,(\left\|B^k\right\|+\left\|B^k\right\|^2)\right]$ *if the denominator of the right-hand side is not 0;*
- *any $\tau>0$ otherwise.*

Moreover, if $\|B\|<1$, we can take

$$\tau<\frac{(1-\|B\|)^2}{\|H\|^2\|M\|^2}(1-\|B\|^k)^2\left[\frac{2c\sin\frac{\theta_0}{2}}{\delta_0}\left\|B^k\right\|^{2k}+\frac{\sqrt{c}}{\delta_0}(1-k\,\|B\|^{k-1}+(k-1)\,\|B\|^k)(1+\|B\|^{2k})\right.$$
$$\left.+2\max\left(\frac{\sqrt{c}}{\delta_0},\frac{\sqrt{c}}{\cos 3\theta_0}\right)(1-k\,\|B\|^{k-1}+(k-1)\,\|B\|^k)\,\|B\|^k\right]^{-1}.$$

Proof. Writing (56) for $\gamma=\gamma_3$ as in Lemma A.4 (iii), we have

$$[\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)]\,\|y\|^2+\tau G(P^*+\gamma_3Q^*,P+\gamma_3Q)-(1+\gamma_3^2)\tau G(Q^*,Q) \\ +\tau\,([\Re(\lambda-1)+\gamma_3\Im(\lambda-1)]L_1+[\gamma_3\Re(\lambda-1)-\Im(\lambda-1)]L_2) \qquad =0. \tag{59}$$

Since $G(P^*+\gamma_3Q^*,P+\gamma_3Q)\geq 0$, the left-hand side of (59) will be strictly positive if τ satisfies:

$$\tau<\frac{1}{\|y\|^2}\left[(1+\gamma_3^2)\frac{G(Q^*,Q)}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)}\right. \\ \left.+|L_1|\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)}+|L_2|\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)}\right]^{-1}.$$

By estimating

- $G(Q^*,Q)\leq(\|H\|\,\|T_k\|\,\|M\|\,q_2|\sin\theta|)^2\,\|y\|^2$
- $|L_1|\leq\|X_k\|\,\|M\|^2\,(p^2+q_1^2)\,\|y\|^2$;
- $|L_2|\leq 2\,\|X_k\|\,\|M\|^2\,pq_1\,\|y\|^2$

and using Lemma A.4 (iii), it suffices to choose

$$\left[(1+\gamma_3^2)\left(\|H\|\,\|T_k\|\,\|M\|\,q_2\right)^2 \frac{2|\sin\frac{\theta}{2}|\cos^2\frac{\theta}{2}}{\delta_0} + \|X_k\|\,\|M\|^2\,(p^2+q_1^2)\frac{\sqrt{1+\gamma_3^2}}{\delta_0}\right.$$
$$\left.+2\,\|X_k\|\,\|M\|^2\,pq_1\max\left(\frac{\sqrt{1+\gamma_3^2}}{\delta_0},\frac{\sqrt{1+\gamma_3^2}}{\cos 3\theta_0}\right)\right]\tau \qquad < 1.$$

Noting that $c = 1+\gamma_3^2$, the final result is obtained by definitions (50), (51), (52) of p, q_1, q_2. Finally, the conclusion in the case $0 < \|B\| < 1$ can be obtained by Lemma 4.6. □

Lemma 4.13 (Case 4). *For $k \geq 2$, equation (48) admits no solutions λ in Case 4 if we take*

- $\tau < \dfrac{\left[\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0\right] s(B^k)^{-4}}{\|H\|^2\,\|M\|^2\,\|T_k\|^2\,(1+\|B^k\|)^2+2\,\|M\|^2\,\|X_k\|\,(1+2\,\|B^k\|)^2}$ *if the denominator of the right-hand side is not 0;*
- *any $\tau > 0$ otherwise.*

Moreover, if $\|B\| < 1$, we can take

$$\tau < \frac{(1-\|B\|)^2}{\|H\|^2\,\|M\|^2}\cdot\frac{\left[\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0\right](1-\|B\|^k)^2}{(1-\|B\|^k)^2+2(1-k\,\|B\|^{k-1}+(k-1)\,\|B\|^k)(1+\|B\|^k)^2}.$$

Proof. Here it is enough to consider (54). By the properties of G

$$G(Q^*,Q)\geq 0, \quad G(P^*,P)\leq\left(\|H\|\,\|T_k\|\,\|M\|\,p\right)^2\|y\|^2$$

and Lemma A.4 (iv), we see that the left-hand side of (54) will be strictly negative if τ satisfies:

$$\left[\left(\|H\|\,\|T_k\|\,\|M\|\,p\right)^2\frac{1}{\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0}+\|X_k\|\,\|M\|^2\,(p+q_1)^2\frac{2}{\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0}\right]\tau \quad < 1.$$

Definitions (50), (51) of p, q_1 lead to the final result. Finally, the conclusion in the case $0 < \|B\| < 1$ can be obtained by Lemma 4.6. □

Similarly, with the help of Lemma A.5, we prove for the k-step one-shot method the analogue of Proposition 4.9. In particular, note that here just three cases of λ need to be considered, because the analogue of the fourth one is excluded by Lemma A.5 (iv).

Proposition 4.14 (k-step one-shot method). *$\exists\tau > 0$ sufficiently small such that equation (49) admits no solution $\lambda \in \mathbb{C}\backslash\mathbb{R}$, $|\lambda| \geq 1$. In particular, if $\|B\| < 1$, given any $\delta_0 > 0$ and $0 < \theta_0 < \frac{\pi}{4}$, take*

$$\tau < \frac{\min\{\psi_1(k,b),\ \psi_2(k,b),\ \psi_3(k,b)\}}{\|H\|^2\,\|M\|^2}$$

where

$$\psi_1(k,b) = \frac{(1-b)^2(1-b^k)^2}{4b^{2k}+\sqrt{2}(1-kb^{k-1}+(k-1)b^k)(1+b^k)^2}$$

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

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

and $c = \frac{1+2\delta_0\sin\frac{3\theta_0}{2}+\delta_0^2}{\cos^2\frac{3\theta_0}{2}}$.

4.4 Final result ($k \geq 2$)

Considering Remark 4.5, and taking the minimum between the bound in Proposition 4.7 for real eigenvalues and the bound in Proposition 4.9 for complex eigenvalues, we finally obtain a sufficient condition on the descent step τ to ensure convergence of the shifted multi-step one-shot method.

Theorem 4.15 (Convergence of shifted k-step one-shot, $k \geq 2$)**.** *Under assumption (4), the shifted k-step one-shot method, $k \geq 2$, converges for sufficiently small τ. In particular, for $\|B\| < 1$, it is enough to take*

$$\tau < \frac{\chi(k, \|B\|)}{\|H\|^2 \|M\|^2},$$

where $\chi(k, \|B\|)$ is an explicit function of k and $\|B\|$.

Similarly, by combining Remark 4.5, Propositions 4.8 and 4.14, we obtain a sufficient condition on the descent step τ to ensure convergence of the multi-step one-shot method.

Theorem 4.16 (Convergence of k-step one-shot, $k \geq 2$)**.** *Under assumption (4), the k-step one-shot method, $k \geq 2$, converges for sufficiently small τ. In particular, for $\|B\| < 1$, it is enough to take*

$$\tau < \frac{\psi(k, \|B\|)}{\|H\|^2 \|M\|^2},$$

where $\psi(k, \|B\|)$ is an explicit function of k and $\|B\|$.

5 Inverse problem with complex forward problem and real parameter

In this section we show that a linear inverse problem with associated complex forward problem and real parameter can be transformed into a linear inverse problem which matches with the real model at the beginning of Section 2, so that the previous theory applies. More precisely, here we study the state equation

$$u = Bu + M\sigma + F$$

where $u \in \mathbb{C}^{n_u}$, $\sigma \in \mathbb{R}^{n_\sigma}$, $B \in \mathbb{C}^{n_u \times n_u}$, $M \in \mathbb{C}^{n_u \times n_\sigma}$. We measure $Hu(\sigma) = f$ where $H \in \mathbb{C}^{n_f \times n_u}$ and we want to recover σ from f. Using the method of least squares, we consider the cost functional

$$J(\sigma) \coloneqq \frac{1}{2} \|Hu(\sigma) - f\|^2,$$

then by the Lagrangian technique with

$$\mathcal{L}(u, v, \sigma) = \frac{1}{2} \|Hu - f\|^2 + \Re\langle Bu + m\sigma + F - u, v\rangle,$$

we can define the adjoint state $p = p(\sigma)$ such that

$$p = B^* p + H^*(Hu(\sigma) - f),$$

which allows us to compute

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

By separating the real and imaginary parts of all vectors and matrices $u = u_1 + \mathrm{i}u_2$, $p = p_1 + \mathrm{i}p_2$, $B = B_1 + \mathrm{i}B_2$, $M = M_1 + \mathrm{i}M_2$, $F = F_1 + \mathrm{i}F_2$, $H = H_1 + \mathrm{i}H_2$, $f = f_1 + \mathrm{i}f_2$, we can transform

this inverse problem with complex forward problem into the inverse problem with real forward problem introduced at the beginning of Section 2. Indeed, note that $B^* = B_1^* - \mathrm{i}B_2^*$, $M^* = M_1^* - \mathrm{i}M_2^*$, $H^* = H_1^* - \mathrm{i}H_2^*$, so we have

$$\begin{cases} u_1 + \mathrm{i}u_2 = (B_1 + \mathrm{i}B_2)(u_1 + \mathrm{i}u_2) + (M_1 + \mathrm{i}M_2)\sigma + (F_1 + \mathrm{i}F_2) \\ p_1 + \mathrm{i}p_2 = (B_1^* - \mathrm{i}B_2^*)(p_1 + \mathrm{i}p_2) + (H_1^* - \mathrm{i}H_2^*)[(H_1 + \mathrm{i}H_2)(u_1 + \mathrm{i}u_2) - (f_1 + \mathrm{i}f_2)] \\ \nabla J(\sigma) = \Re[(M_1^* - \mathrm{i}M_2^*)(p_1 + \mathrm{i}p_2)], \end{cases}$$

which implies

$$\begin{cases} u_1 = B_1 u_1 - B_2 u_2 + M_1\sigma + F_1 \\ u_2 = B_2 u_1 + B_1 u_2 + M_2\sigma + F_2 \\ p_1 = B_1^* p_1 + B_2^* p_2 + (H_1^* H_1 + H_2^* H_2)u_1 - (H_2^* H_1 - H_1^* H_2)u_2 - (H_1^* f_1 + H_2^* f_2) \\ p_2 = -B_2^* p_1 + B_1^* p_2 + (H_2^* H_1 - H_1^* H_2)u_1 + (H_1^* H_1 + H_2^* H_2)u_2 - (-H_2^* f_1 + H_1^* f_2) \\ \nabla J(\sigma) = M_1^* p_1 + M_2^* p_2. \end{cases}$$

By setting

$$\tilde{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \tilde{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \tilde{B} = \begin{bmatrix} B_1 & -B_2 \\ B_2 & B_1 \end{bmatrix}, \tilde{M} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \tilde{F} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}, \tilde{H} = \begin{bmatrix} H_1 & -H_2 \\ H_2 & H_1 \end{bmatrix}, \tilde{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

we have

$$\begin{cases} \tilde{u} = \tilde{B}\tilde{u} + \tilde{M}\sigma + \tilde{F} \\ \tilde{p} = \tilde{B}^*\tilde{p} + \tilde{H}^*(\tilde{H}\tilde{u} - \tilde{f}) \\ \nabla J(\sigma) = \tilde{M}^*\tilde{p}, \end{cases}$$

that has the same structure as the inverse problem at the beginning of Section 2.

Finally we finish this section by two lemmas that match the assumptions of the inverse problem with complex state variable with the assumptions of the transformed inverse problem with real state variable.

Lemma 5.1. $\mathrm{Spec}(\tilde{B}) = \mathrm{Spec}(B) \cup \overline{\mathrm{Spec(B)}}$.

Proof. By writing

$$\tilde{B} = \begin{bmatrix} B_1 & -B_2 \\ B_2 & B_1 \end{bmatrix} = \underbrace{\begin{bmatrix} I & I \\ \mathrm{i}I & -\mathrm{i}I \end{bmatrix}}_{C^{-1}} \begin{bmatrix} \overline{B} & 0 \\ 0 & B \end{bmatrix} \underbrace{\begin{bmatrix} \frac{1}{2}I & -\frac{\mathrm{i}}{2}I \\ \frac{1}{2}I & \frac{i}{2}I \end{bmatrix}}_{C}, \tag{60}$$

we find that $\det(\tilde{B} - \lambda I) = \det(\overline{B} - \lambda I)\det(B - \lambda I)$. The conclusion is then deduced thanks to the fact that $\mathrm{Spec}(\overline{B}) = \overline{\mathrm{Spec(B)}}$. □

Lemma 5.2. *Assume that $\rho(B) < 1$, and $H(I - B)^{-1}M$ is injective. Then $\rho(\tilde{B}) < 1$, and $\tilde{H}(\tilde{I} - \tilde{B})^{-1}\tilde{M}$ is injective where $\tilde{I} \in \mathbb{R}^{2n_u \times 2n_u}$ is the identity matrix.*

Proof. The previous lemma says that $\rho(\tilde{B}) = \rho(B) < 1$. Therefore $(\tilde{I} - \tilde{B})^{-1}$ is well-defined and thanks to (60),

$$\begin{aligned} (\tilde{I} - \tilde{B})^{-1} &= \underbrace{\begin{bmatrix} I & I \\ \mathrm{i}I & -\mathrm{i}I \end{bmatrix}}_{C^{-1}} \begin{bmatrix} (I - \overline{B})^{-1} & 0 \\ 0 & (I - B)^{-1} \end{bmatrix} \underbrace{\begin{bmatrix} \frac{1}{2}I & -\frac{\mathrm{i}}{2}I \\ \frac{1}{2}I & \frac{i}{2}I \end{bmatrix}}_{C} \\ &= \frac{1}{2}\begin{bmatrix} (I - \overline{B})^{-1} + (I - B)^{-1} & -\mathrm{i}(I - \overline{B})^{-1} + \mathrm{i}(I - B)^{-1} \\ \mathrm{i}(I - \overline{B})^{-1} - \mathrm{i}(I - B)^{-1} & (I - \overline{B})^{-1} + (I - B)^{-1} \end{bmatrix}. \end{aligned}$$

Now we have

$$\begin{aligned}\tilde{H}(\tilde{I}-\tilde{B})^{-1}\tilde{M} &= \tfrac{1}{2}\begin{bmatrix} H_1 & -H_2 \\ H_2 & H_1 \end{bmatrix}\begin{bmatrix} (I-\overline{B})^{-1}+(I-B)^{-1} & -\mathrm{i}(I-\overline{B})^{-1}+\mathrm{i}(I-B)^{-1} \\ \mathrm{i}(I-\overline{B})^{-1}-\mathrm{i}(I-B)^{-1} & (I-\overline{B})^{-1}+(I-B)^{-1} \end{bmatrix}\begin{bmatrix} M_1 \\ M_2 \end{bmatrix} \\ &= \tfrac{1}{2}\begin{bmatrix} \overline{H}(I-\overline{B})^{-1}+H(I-B)^{-1} & -\mathrm{i}\overline{H}(I-\overline{B})^{-1}+\mathrm{i}H(I-B)^{-1} \\ \mathrm{i}\overline{H}(I-\overline{B})^{-1}-\mathrm{i}H(I-B)^{-1} & \overline{H}(I-\overline{B})^{-1}+H(I-B)^{-1} \end{bmatrix}\begin{bmatrix} M_1 \\ M_2 \end{bmatrix} \\ &= \tfrac{1}{2}\begin{bmatrix} \overline{H}(I-\overline{B})^{-1}\overline{M}+H(I-B)^{-1}M \\ \mathrm{i}\overline{H}(I-\overline{B})^{-1}\overline{M}-\mathrm{i}H(I-B)^{-1}M \end{bmatrix}.\end{aligned}$$

Now assume that there exists $x \in \mathbb{C}^{n_\sigma}$ such that $H(\tilde{I}-\tilde{B})^{-1}\tilde{M}x = 0$, then

$$\begin{cases} [\overline{H}(I-\overline{B})^{-1}\overline{M}+H(I-B)^{-1}M]x = 0 \\ [\mathrm{i}\overline{H}(I-\overline{B})^{-1}\overline{M}-\mathrm{i}H(I-B)^{-1}M]x = 0 \end{cases}$$

or, equivalently,

$$\begin{cases} [H(I-\overline{B})^{-1}\overline{M}+H(I-B)^{-1}M]x = 0 \\ [-\overline{H}(I-\overline{B})^{-1}\overline{M}+H(I-B)^{-1}M]x = 0. \end{cases}$$

By summing up these two equations we deduce that $H(I-B)^{-1}Mx = 0$, then $x = 0$ thanks to the injectivity of $H(I-B)^{-1}M$. □

6 Numerical experiments

Let us introduce a toy model to illustrate numerically the performance of the different methods. Given $\Omega \subset \mathbb{R}^n$ an open bounded Lipschitz domain, we consider the direct problem for the linearized scattered field $u \in \mathbb{H}^2(\Omega)$ given by the Helmholtz equation

$$\begin{cases} \mathbf{div}(\tilde{\sigma}_0\nabla u)+\tilde{k}^2u = \mathbf{div}(\sigma\nabla u_0), & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{61}$$

where the incident field $u_0 : \Omega \to \mathbb{R}$ satisfies

$$\begin{cases} \mathbf{div}(\tilde{\sigma}_0\nabla u_0)+\tilde{k}^2u = 0, & \text{in } \Omega, \\ u_0 = f, & \text{on } \partial\Omega \end{cases} \tag{62}$$

with the datum $f : \partial\Omega \to \mathbb{R}$. Here $\sigma : \overline{\Omega} \to \mathbb{R}$ such that $\sigma\big|_{\partial\Omega} = 0$; $\tilde{\sigma}_0 = \sigma_0 + \delta\sigma_\mathrm{r}$ is a given function with $\delta \geq 0$ and random σ_r. More precisely, given $\tilde{\sigma}_0$ and f, we solve for $u_0 = u_0(f)$ in (62), then insert u_0 into (61) to solve for $u = u(\sigma)$. The variational formulations for u and u_0 are respectively

$$\int_\Omega \tilde{\sigma}_0\nabla u\cdot\nabla v - \int_\Omega \tilde{k}^2uv = \int_\Omega \sigma\nabla u_0\cdot\nabla v, \quad \forall v \in \mathbb{H}^1_0(\Omega) \quad \text{and } u = 0 \text{ on } \partial\Omega, \tag{63}$$

$$\int_\Omega \tilde{\sigma}_0\nabla u_0\cdot\nabla v - \int_\Omega \tilde{k}^2uv = 0, \quad \forall v \in \mathbb{H}^1_0(\Omega) \quad \text{and } u_0 = f \text{ on } \partial\Omega. \tag{64}$$

We are interested in the inverse problem of finding σ from the measurement $Hu(\sigma)$ where $Hu := \tilde{\sigma}_0\frac{\partial u}{\partial\nu}\big|_{\partial\Omega}$. To solve this inverse problem we use the method of least squares. Denoting by σ^{ex} the exact σ and $g = \tilde{\sigma}_0\frac{\partial u(\sigma^{\mathrm{ex}})}{\partial\nu}\big|_{\partial\Omega}$ the corresponding measurement, we consider the cost functional

$J(\sigma) = \frac{1}{2}\|Hu(\sigma) - g\|^2_{\mathcal{L}^2(\partial\Omega)} = \frac{1}{2}\int_{\partial\Omega}(\tilde{\sigma}_0 \frac{\partial u(\sigma)}{\partial \nu} - g)^2$. The Lagrangian technique allows us to compute the gradient $\nabla_\sigma J(\sigma) = -\nabla u_0 \cdot \nabla p(\sigma)$, where the adjoint state $p = p(\sigma)$ satisfies

$$\int_\Omega \tilde{\sigma}_0 \nabla p \cdot \nabla v - \int_\Omega \tilde{k}^2 pv = 0, \quad \forall v \in \mathbb{H}^1(\Omega) \quad \text{and } p = \left(\tilde{\sigma}_0 \frac{\partial u(\sigma)}{\partial \nu}\bigg|_{\partial\Omega} - g\right) \text{ on } \partial\Omega. \tag{65}$$

By discretizing u by $\mathbb{P}^1$ finite elements on a mesh $\mathcal{T}_h^u$ of Ω, and σ by $\mathbb{P}^0$ finite elements on a coarser mesh $\mathcal{T}_h^\sigma$ of Ω, the discretization of (63) can be written as the linear system $A_1\vec{u} = A_2\vec{\sigma}$, where $\vec{u} \in \mathbb{R}^{n_u}$, $\vec{\sigma} \in \mathbb{R}^{n_\sigma}$. More precisely, A_1 and A_2 are respectively issued from the discretization of $\int_\Omega \tilde{\sigma}_0 \nabla u \cdot \nabla v - \int_\Omega \tilde{k}^2 uv$ and $\int_\Omega \sigma \nabla u_0 \cdot \nabla v$, where the Dirichlet boundary conditions are imposed by the penalty method. To rewrite the system in the form (1), we consider the naive splitting $A_1 = A_{11} + \delta A_{12}$, where A_{11} and A_{12} are respectively issued from the discretization of $\int_\Omega \sigma_0 \nabla u \cdot \nabla v - \int_\Omega \tilde{k}^2 uv$ and $\int_\Omega \sigma_{\mathrm{r}} \nabla u \cdot \nabla v$. Then we get

$$\vec{u} = A_{11}^{-1}(-\delta A_{12}\vec{u} + A_2\vec{\sigma}) \quad \text{and } \vec{u} = 0 \text{ on } \partial\Omega$$

and

$$\vec{p} = A_{11}^{-1}(-\delta A_{12}\vec{p}) \quad \text{and } \vec{p} = H\vec{u} - \vec{g} \text{ on } \partial\Omega$$

where $H \in \mathbb{R}^{n_f \times n_u}$ is the discretization of the above operator H by abuse of notation. Choosing δ such that $\delta \left\|A_{11}^{-1}A_{12}\right\|_2 < 1$, we consider (3) with $B = -\delta A_{11}^{-1}A_{12}$, $M = A_{11}^{-1}A_2$, $F = 0$. The application of A_{11}^{-1}, which has the same size as matrix A_1, is done by a direct solver; more practical fixed point iterations will be investigated in the future.

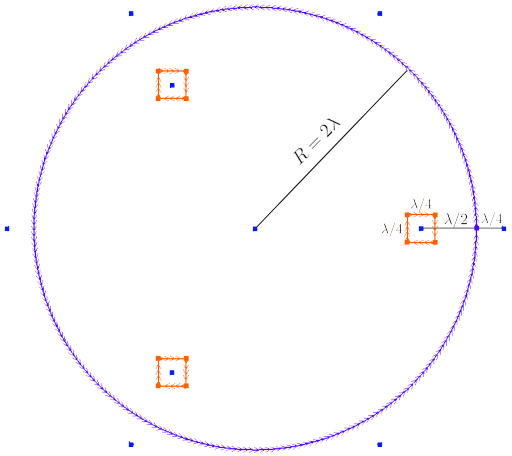


Figure 1: Domain with six source points for the numerical experiments. The unknown σ is supported on the three squares.

We then perform some numerical experiments in FreeFEM [12] with the following setting:

- Wavenumber $\tilde{k} = 2\pi$, $\sigma_0 = 1$, $\delta = 0.01$, σ_{r} is a random real function with range in the interval $[1, 2]$.
- Wavelength $\lambda = \frac{2\pi}{\tilde{k}}\sqrt{\sigma_0} = 1$, mesh size $h = \frac{\lambda}{20} = 0.05$. The domain Ω is the disk shown in Figure 1, where the squares are the support of function σ. Here $n_u = 5853$, $n_\sigma = 6$.

- We test with 6 data f given by zero-order Bessel function of the second kind centered at the points shown in Figure 1, and the cost functional is the normalized sum of the contributions corresponding to different data.
- We take $\sigma^{\text{ex}} = 10$ in every square and 0 otherwise. The initial guess for the inverse problem is 12 in every square and 0 otherwise.
- For the first iteration, we perform a line search to adapt the descent step τ, using a direct solver for the forward and adjoint problems.
- The stopping rule for the outer iteration is based on the relative value of the cost functional and on the relative norm of the gradient with a tolerance of 10^{-5}.

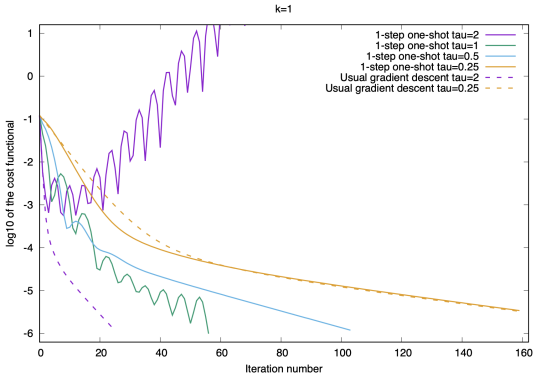
Recall that k is the number of inner iterations on the direct and adjoint problems. We are interested in two experiments.

In the first experiment, we study the dependence on the descent step τ. In Figure 2a and 2b we respectively fix $k = 1$ and $k = 2$ and compare k-step one-shot methods with the usual gradient descent method. On the horizontal axis we indicate the (outer) iteration number n in (5) and (9). We can verify that for sufficiently small τ, both one-shot methods converge. In particular, for $\tau = 2$, while 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 vanish when τ gets smaller. For sufficiently small τ, the convergence curves of both one-shot methods are comparable to the one of gradient descent.

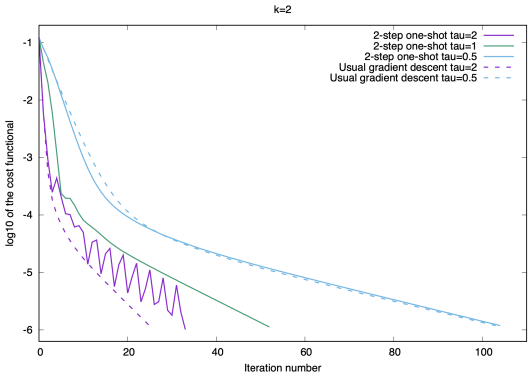
In the second experiment, we study the dependence on the number of inner iterations k, for fixed τ. First (Figures 2c–2d), we investigate for which k the convergence curve of k-step one-shot is comparable with the one of usual gradient descent. As in the previous pictures, on the horizontal axis we indicate the (outer) iteration number n in (5) and (9). For $\tau = 2$ (see Figure 2c), we observe that for $k = 3, 4$ the convergence curves of k-step one-shot are close to the one of usual gradient descent. Note that with 3 inner iterations the $\mathcal{L}^2$ error between u^n and the exact solution to the forward problem ranges between $4.3 \cdot 10^{-6}$ and 0.0136 for different n in (9); in fact this error is rather significant at the beginning then it tends to reduce when we are closer to 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 very particular case $\tau = 2.5$ (see Figure 2d), we observe an interesting phenomenon: when $k = 3, 5, 10$, with k-step one-shot the cost functional decreases even faster than with usual GD. For bigger k, for example $k = 14$, the convergence curve of one-shot is close to the one of usual gradient descent as expected. Next (Figures 2e–2f), since the overall cost of the k-step one-shot method increases with k, we indicate on the horizontal axis the accumulated inner iteration number, which sums up k from an outer iteration to the next. More precisely, because at the first outer iteration we perform a step search by a direct solver, we set to 1 the first accumulated inner iteration number; for the following outer iterations $n \geq 2$, the accumulated inner iteration number is set to $1 + (n-1)k$. In Figures 2e–2f we replot the results for the converging k-step one-shot methods of Figures 2c–2d with respect to the accumulated inner iteration number. For $\tau = 2$ (see Figure 2e), while $k = 2$ presents some oscillations, quite interestingly it appears that $k = 3$ gives a faster decrease of the cost functional with respect to $k = 4$, at least after the first iterations. For $\tau = 2.5$ (see Figure 2f) we observe that $k = 3$ is enough for the decrease of the cost functional, but with some oscillations, and the considered higher k appears again to give slower decrease.

A similar behavior can be observed for the shifted methods in Figure 3.

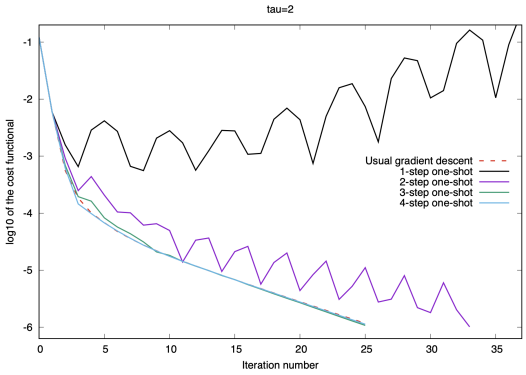
Finally we fix two particular values of τ and compare all considered methods in Figure 4. We note that shifted methods present more oscillations with respect to non-shifted ones, especially for larger τ.



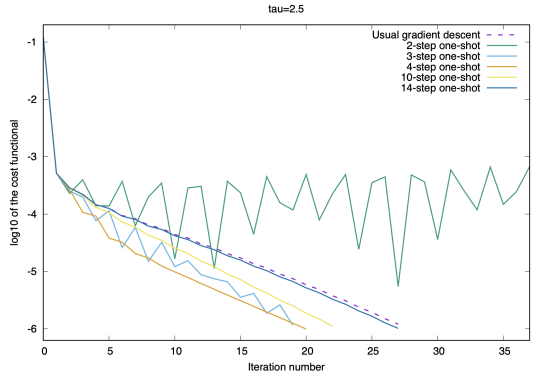
(a) Convergence curves of usual gradient descent and 1-step one-shot for different descent step τ.



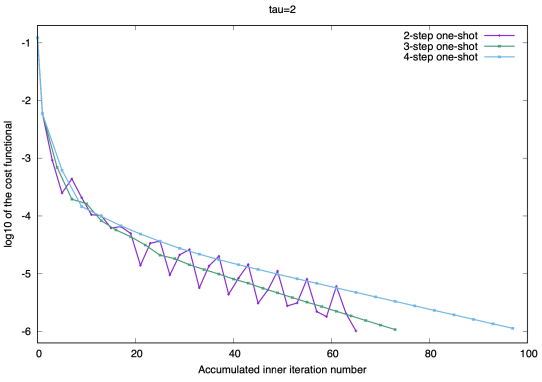
(b) Convergence curves of usual gradient descent and 2-step one-shot for different descent step τ.



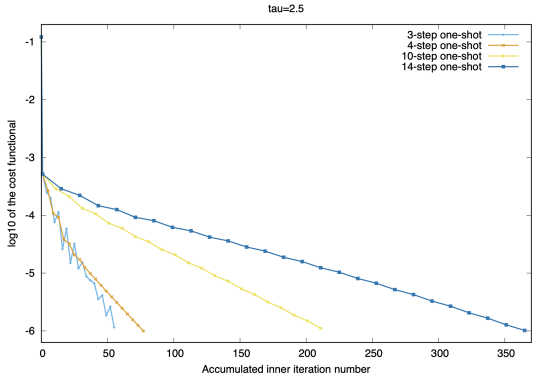
(c) Convergence curves of usual gradient descent and k-step one-shot for different k with $\tau = 2$.



(d) Convergence curves of usual gradient descent and k-step one-shot for different k with $\tau = 2.5$.

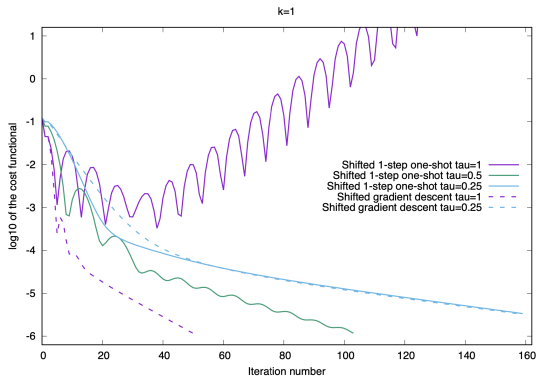


(e) Convergence curves of k-step one-shot for different k with $\tau = 2$.

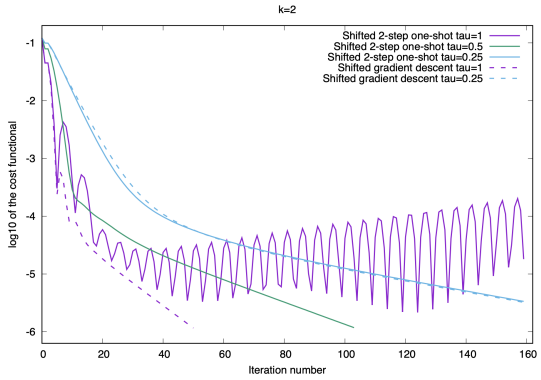


(f) Convergence curves of k-step one-shot for different k with $\tau = 2.5$.

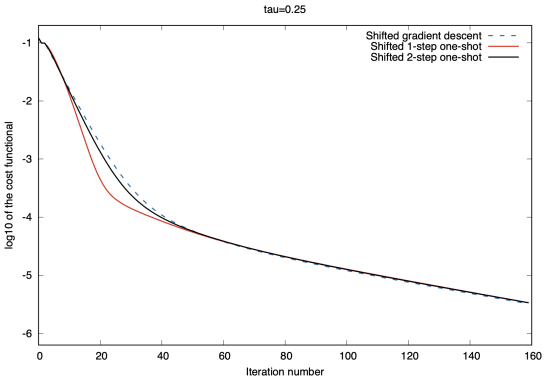
Figure 2: Convergence curves of usual gradient descent and k-step one-shot.



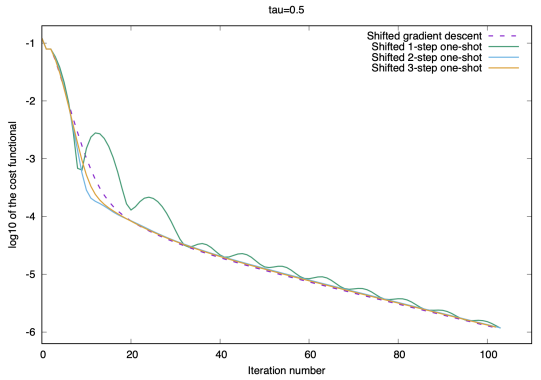
(a) Convergence curves of shifted gradient descent and shifted 1-step one-shot for different descent step τ.



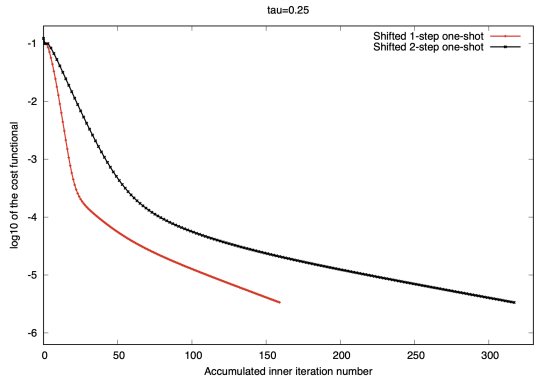
(b) Convergence curves of shifted gradient descent and shifted 2-step one-shot for different descent step τ.



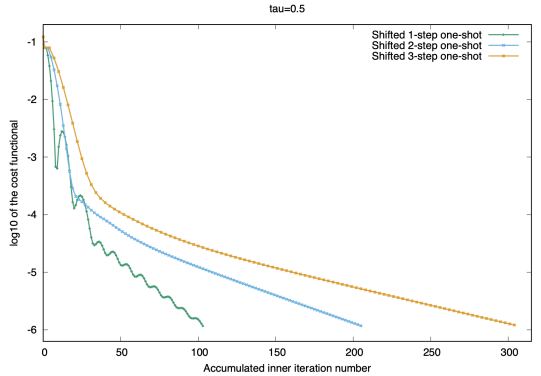
(c) Convergence curves of shifted gradient descent and shifted k-step one-shot for different k with $\tau = 0.25$.



(d) Convergence curves of shifted gradient descent and shifted k-step one-shot for different k with $\tau = 0.5$.

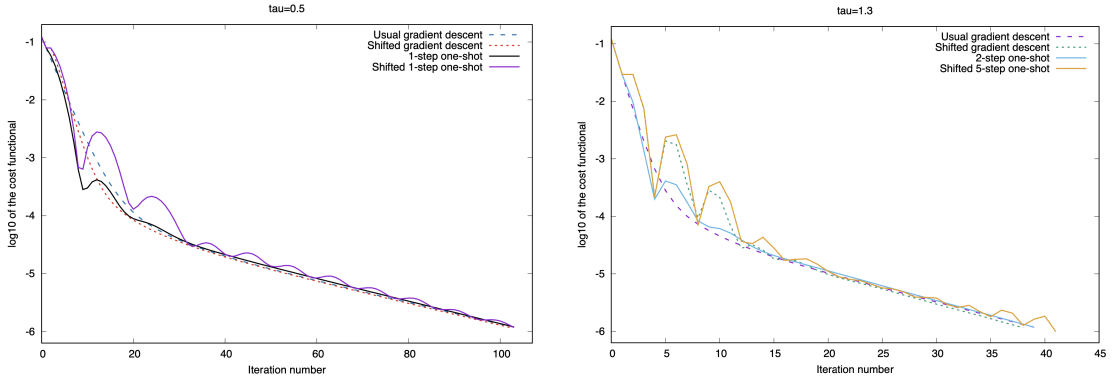


(e) Convergence curves of shifted k-step one-shot for different k with $\tau = 0.25$.



(f) Convergence curves of shifted k-step one-shot for different k with $\tau = 0.5$.

Figure 3: Convergence curves of shifted gradient descent and shifted k-step one-shot.



(a) Convergence curves with $\tau = 0.5$.

(b) Convergence curves with $\tau = 1.3$.

Figure 4: Comparison of usual gradient descent and k-step one-shot with shifted gradient descent and shifted k-step one-shot.

7 Conclusion

We have proved sufficient conditions on the descent step for the convergence of two variants of multi-step one-shot methods. 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 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 are enough to guarantee good convergence of the inversion algorithm.

These encouraging numerical results are preliminary in the sense that the considered fixed point iteration is not a practical one, since it involves a direct solve of a problem of the same size as the original forward problem. We will investigate in the future iterative solvers based on domain decomposition methods (see e.g. [3]), which are well adapted to large-scale problems. In addition, fixed point iterations could be replaced by more efficient Krylov subspace methods, such as conjugate gradient or GMRES.

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 non-linear and time-dependent inverse problems.

References

[1] S. Barnett. *Polynomials and linear control systems*, volume 77 of *Pure Appl. Math.* Marcel Dekker, Inc., New York, NY, 1983.

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

[3] 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.

[4] 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 *Con-*

strained Optimization and Optimal Control for Partial Differential Equations, International Series of Numerical Mathematics, pages 99–122. Springer Basel, 2012.

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

[6] 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.

[7] S. GÃ¼nther, 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.

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

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

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

[11] 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.

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

[13] E.I. Jury. On the roots of a real polynomial inside the unit circle and a stability criterion for linear discrete systems. *IFAC Proceedings Volumes*, 1(2):142–153, 1963. 2nd International IFAC Congress on Automatic and Remote Control: Theory, Basle, Switzerland, 1963.

[14] E.I. Jury. *Theory and Applications of the Z-Transform Method*. New York, 1964.

[15] 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.

[16] M. Marden. *The geometry of the zeros of a polynomial in a complex variable*, volume 3 of *Math. Surv.* American Mathematical Society (AMS), Providence, RI, 1949.

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

[18] 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.

[19] 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.

[20] I. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1917(147):205–232, 1917.

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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 (24), (25), (48), (49).

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

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

$$s(T) \coloneqq \sup_{z\in\mathbb{C}, |z|\geq 1} \left\|(I - T/z)^{-1}\right\| \tag{66}$$

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

Proof. The functional $z \mapsto \left\|(I-T/z)^{-1}\right\|$, with $z \in \mathbb{C}, |z| \geq 1$, is well-defined and continuous, and we use Lemma A.1. □

The following lemma says that, for $T \in \mathbb{C}^{n\times n}$ and $\lambda \in \mathbb{C}, |\lambda| \geq 1$, 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 $\lambda \in \mathbb{C}, |\lambda| \geq 1$. Write $\frac{1}{\lambda} = r(\cos\phi + \mathrm{i}\sin\phi)$ in polar form, where $0 < r \leq 1$ and $\phi \in [-\pi, \pi]$. Then*

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

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

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

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

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 3.3 and 4.3 we identify different cases of $\lambda \in \mathbb{C}$ and we need corresponding estimations, given in the two following lemmas. Lemma A.4 is used for the shifted k-step one-shot method and Lemma A.5 is used for the k-step one-shot method.

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 $\Re(\lambda^3 - \lambda^2) \geq 0$, let $\gamma_1 = \gamma_1(\lambda) = \begin{cases} 1, & \text{if } \Im(\lambda^3 - \lambda^2) \geq 0, \\ -1, & \text{if } \Im(\lambda^3 - \lambda^2) < 0 \end{cases}$ then

$$\Re(\lambda^3 - \lambda^2) + \gamma_1\Im(\lambda^3 - \lambda^2) \geq |\lambda - 1| \geq 2|\sin(\theta/2)|.$$

(ii) Let $0 < \theta_0 \leq \frac{\pi}{6}$. For λ satisfying $\Re(\lambda^3 - \lambda^2) < 0$ and $\theta \in [\theta_0, \pi - \theta_0] \cup [-\pi + \theta_0, -\theta_0]$, let $\gamma_2 = \begin{cases} -1, & \text{if } \Im(\lambda^3 - \lambda^2) \geq 0, \\ 1, & \text{if } \Im(\lambda^3 - \lambda^2) < 0 \end{cases}$ then

$$-\Re(\lambda^3 - \lambda^2) - \gamma_2\Im(\lambda^3 - \lambda^2) \geq |\lambda - 1| \geq 2\sin(\theta_0/2).$$

(iii) Let $0 < \theta_0 \leq \frac{\pi}{6}$ and $\delta_0 > 0$. For λ satisfying $\Re(\lambda^3 - \lambda^2) < 0$ and $\theta \in (-\theta_0, \theta_0)\backslash\{0\}$, let $\gamma_3 = \gamma_3(\mathrm{sign}(\theta)) = \begin{cases} \left(\delta_0 + \sin\frac{5\theta_0}{2}\right)/\cos\frac{5\theta_0}{2} & \text{if } \theta > 0, \\ -\left(\delta_0 + \sin\frac{5\theta_0}{2}\right)/\cos\frac{5\theta_0}{2} & \text{if } \theta < 0 \end{cases}$ then

$$\Re(\lambda^3 - \lambda^2) + \gamma_3\Im(\lambda^3 - \lambda^2) \geq 2\delta_0|\sin(\theta/2)|.$$

Moreover, if $0 < \theta_0 < \frac{\pi}{6}$, we have

$$\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)} \leq \frac{\sqrt{1+\gamma_3^2}}{\delta_0} \quad \textit{and} \quad \frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)} \leq \max\left(\frac{\sqrt{1+\gamma_3^2}}{\delta_0}, \frac{\sqrt{1+\gamma_3^2}}{\cos 3\theta_0}\right).$$

(iv) Let $0 < \theta_0 \leq \frac{\pi}{6}$. For λ satisfying $\Re(\lambda^3-\lambda^2) < 0$ and $\theta \in (\pi-\theta_0, \pi) \cup (-\pi, -\pi+\theta_0)$, we have

$$-\Re(\lambda^3-\lambda^2) \geq \sin\left(\frac{\pi}{2}-3\theta_0\right) + \cos 2\theta_0,$$

$$\frac{|\Re(\lambda-1)|}{-\Re(\lambda^3-\lambda^2)} \leq \frac{2}{\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0} \quad \textit{and} \quad \frac{|\Im(\lambda-1)|}{-\Re(\lambda^3-\lambda^2)} \leq \frac{2}{\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0}.$$

Proof. (i) From the definition of γ_1 we see that $\gamma_1^2 = 1$, $\gamma_1\Im(\lambda^3-\lambda^2) \geq 0$ and

$$\begin{aligned}\left[\Re(\lambda^3-\lambda^2)+\gamma_1\Im(\lambda^3-\lambda^2)\right]^2 &= \left[\Re(\lambda^3-\lambda^2)\right]^2 + \left[\Im(\lambda^3-\lambda^2)\right]^2 + 2\gamma_1\Re(\lambda^3-\lambda^2)\Im(\lambda^3-\lambda^2) \\ &\geq \left[\Re(\lambda^3-\lambda^2)\right]^2 + \left[\Im(\lambda^3-\lambda^2)\right]^2 = |\lambda^3-\lambda^2|^2,\end{aligned}$$

which yields $\Re(\lambda^3-\lambda^2)+\gamma_1\Im(\lambda^3-\lambda^2) \geq R^2|\lambda-1|$. Finally,

$$|\lambda-1| = |R\cos\theta - 1 + \mathrm{i}R\sin\theta| = \sqrt{R^2+1-2R\cos\theta} \geq \sqrt{2-2\cos\theta} = 2|\sin(\theta/2)|$$

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 $\left|\sin\frac{\theta}{2}\right| \geq \sin\frac{\theta_0}{2}$. From the definition of γ_2 we see that $\gamma_2^2 = 1$ and $\gamma_2\Im(\lambda^3-\lambda^2) \leq 0$. Similar to (i), we have $-\Re(\lambda^2-\lambda)-\gamma_2\Im(\lambda^2-\lambda) \geq |\lambda-1| \geq 2|\sin(\theta/2)|$, that implies the conclusion.
(iii) Note that $\cos 3\theta > 0, -\frac{\pi}{2} < 3\theta < \frac{\pi}{2}$, and $\sin 3\theta$ has the same sign as θ and γ_3, so we have

$$\begin{aligned}\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2) &= R^2(R\cos 3\theta - \cos 2\theta + \gamma_3 R\sin 3\theta - \gamma_3\sin 2\theta) \\ &\geq \cos 3\theta - \cos 2\theta + \gamma_3\sin 3\theta - \gamma_3\sin 2\theta \\ &= -2\sin\tfrac{5\theta}{2}\sin\tfrac{\theta}{2} + 2\gamma_3\cos\tfrac{5\theta}{2}\sin\tfrac{\theta}{2} \\ &= 2\sin\tfrac{\theta}{2}\left(\gamma_3\cos\tfrac{5\theta}{2} - \sin\tfrac{5\theta}{2}\right).\end{aligned}$$

Then we consider two cases: if $0 < \theta < \theta_0$ then $\gamma_3 > 0$, $\left|\sin\frac{\theta}{2}\right| = \sin\frac{\theta}{2} > 0$, $0 < \frac{5\theta}{2} < \frac{5\theta_0}{2} < \frac{\pi}{2}$ and $\gamma_3\cos\frac{5\theta}{2} - \sin\frac{5\theta}{2} > \gamma_3\cos\frac{5\theta_0}{2} - \sin\frac{5\theta_0}{2} = \delta_0$; if $-\theta_0 < \theta < 0$ then $-\gamma_3 > 0$, $\left|\sin\frac{\theta}{2}\right| = -\sin\frac{\theta}{2} > 0$, $-\frac{\pi}{2} < -\frac{5\theta_0}{2} < \frac{5\theta}{2} < 0$ and $-\gamma_3\cos\frac{5\theta}{2} + \sin\frac{5\theta}{2} > -\gamma_3\cos\frac{5\theta_0}{2} - \sin\frac{5\theta_0}{2} = \delta_0$.

Next, if $0 < \theta_0 < \frac{\pi}{6}$, we will show that $\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)}$ and $\frac{|\gamma_3\Re(\lambda-1)-\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)}$ are both bounded. First,

$$\begin{aligned}\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)} &= \frac{|(\cos\theta+\gamma_3\sin\theta)R-1|}{R^2[(\cos 3\theta+\gamma_3\sin 3\theta)R-(\cos 2\theta+\gamma_3\sin 2\theta)]} \\ &\leq \frac{|(\cos\theta+\gamma_3\sin\theta)R-1|}{(\cos 3\theta+\gamma_3\sin 3\theta)R-(\cos 2\theta+\gamma_3\sin 2\theta)}.\end{aligned}$$

Since γ_3 does not depend on R, let us study $f_1(R) = \left(\frac{aR-1}{bR-c}\right)^2$ where $a = \cos\theta+\gamma_3\sin\theta$, $b = \cos 3\theta+\gamma_3\sin 3\theta$ and $c = \cos 2\theta+\gamma_3\sin 2\theta$. We observe that:

- $a, b, c > 0$. Indeed, $\cos\theta, \cos 2\theta, \cos 3\theta > 0$, and θ and γ_3 have the same sign.
- $bR - c > 0$ since $\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2) > 0$, thus $R > \frac{c}{b}$.
- $ac > b$ (equivalently $\frac{c}{b} > \frac{1}{a}$), since

$$ac = \cos\theta\cos 2\theta + \gamma_3^2\sin\theta\sin 2\theta + \gamma_3\sin 3\theta > \cos\theta\cos 2\theta - \sin\theta\sin 2\theta + \gamma_3\sin 3\theta = b.$$

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

$$\cos 2\theta + \gamma_3 \sin 2\theta < \cos 3\theta + \gamma_3 \sin 3\theta \Leftrightarrow |\gamma_3| > \frac{\left|\sin \frac{5\theta}{2}\right|}{\cos \frac{5\theta}{2}},$$

which is true since

$$|\gamma_3| = \frac{\delta_0 + \sin \frac{5\theta_0}{2}}{\cos \frac{5\theta_0}{2}} > \frac{\left|\sin \frac{5\theta}{2}\right|}{\cos \frac{5\theta}{2}} + \varepsilon_0 \quad \text{where} \quad \varepsilon_0 = \frac{\delta_0}{\cos \frac{5\theta_0}{2}}.$$

Then we study

$$f_1(1) = \left[\frac{\cos\theta - 1 + \gamma_3 \sin\theta}{\cos 3\theta - \cos 2\theta + \gamma_3(\sin 3\theta - \sin 2\theta)}\right]^2 = \left(\frac{-\sin\frac{\theta}{2} + \gamma_3 \cos\frac{\theta}{2}}{-\gamma_3 \sin\frac{5\theta}{2} + \gamma_3^2 \cos\frac{5\theta}{2}}\right)^2 \gamma_3^2.$$

We have:

- $(-\sin\frac{\theta}{2} + \gamma_3 \cos\frac{\theta}{2})^2 \leq 1 + \gamma_3^2$ by Cauchy-Schwarz inequality;
- $\gamma_3^2 = |\gamma_3|^2 > \frac{\gamma_3 \sin\frac{5\theta}{2}}{\cos\frac{5\theta}{2}} + \varepsilon_0|\gamma_3|$ that leads to $-\gamma_3 \sin\frac{5\theta}{2} + \gamma_3^2 \cos\frac{5\theta}{2} > \varepsilon_0 \cos\frac{5\theta_0}{2}|\gamma_3| = \delta_0|\gamma_3|$;

hence $f_1(1) \leq \frac{1+\gamma_3^2}{\delta_0^2}$ and finally $\frac{|\Re(\lambda-1)+\gamma_3\Im(\lambda-1)|}{\Re(\lambda^3-\lambda^2)+\gamma_3\Im(\lambda^3-\lambda^2)} \leq \frac{\sqrt{1+\gamma_3^2}}{\delta_0}$. Next, we have

$$\begin{aligned}\frac{|\gamma_3\Re(\lambda-1) - \Im(\lambda-1)|}{\Re(\lambda^2-\lambda) + \gamma_3\Im(\lambda^2-\lambda)} &= \frac{|(\gamma_3\cos\theta - \sin\theta)R - \gamma_3|}{R^2[(\cos 3\theta + \gamma_3 \sin 3\theta)R - (\cos 2\theta + \gamma_3 \sin 2\theta)]} \\ &\leq \frac{|(\gamma_3\cos\theta - \sin\theta)R - \gamma_3|}{(\cos 3\theta + \gamma_3 \sin 3\theta)R - (\cos 2\theta + \gamma_3 \sin 2\theta)}.\end{aligned}$$

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

- $\gamma_3 b - cd$ and θ have the same sign. Indeed, $\gamma_3 b - cd = (\gamma_3^2 + 1)\sin\theta\cos 2\theta$. Consequently, we always have $(\gamma_3 b - cd)\gamma_3 > 0$.
- We always have $\frac{\gamma_3}{d} > 1$. Indeed, if $\theta > 0$ then $d > 0$ since $\gamma_3 = \frac{\delta_0 + \sin\frac{5\theta_0}{2}}{\cos\frac{5\theta_0}{2}} > \frac{\sin\theta}{\cos\theta}$, also $\frac{\gamma_3}{d} = \frac{\gamma_3}{\gamma_3\cos\theta - \sin\theta} > 1$; if $\theta < 0$ then $d < 0$ since $-\gamma_3 = \frac{\delta_0 + \sin\frac{5\theta_0}{2}}{\cos\frac{5\theta_0}{2}} > -\frac{\sin\theta}{\cos\theta}$, also $\frac{\gamma_3}{d} = \frac{-\gamma_3}{-\gamma_3\cos\theta + \sin\theta} > 1$.

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

$$f_2(1) = \left(\frac{-\cos\frac{\theta}{2} - \gamma_3 \sin\frac{\theta}{2}}{-\gamma_3 \sin\frac{5\theta}{2} + \gamma_3^2 \cos\frac{5\theta}{2}}\right)^2 \gamma_3^2 \leq \frac{1+\gamma_3^2}{\delta_0^2},$$

and $\lim_{R\to+\infty} f_2(R) = \left(\frac{\gamma_3\cos\theta - \sin\theta}{\cos 3\theta + \gamma_3 \sin 3\theta}\right)^2 \leq \frac{1+\gamma_3^2}{\cos^2 3\theta_0}$. Therefore

$$\frac{|\gamma_3\Re(\lambda-1) - \Im(\lambda-1)|}{\Re(\lambda^2-\lambda) + \gamma_3\Im(\lambda^2-\lambda)} \leq \max\left(\frac{\sqrt{1+\gamma_3^2}}{\delta_0}, \frac{\sqrt{1+\gamma_3^2}}{\cos 3\theta_0}\right).$$

(iv) Since $\theta \in (\pi - \theta_0, \pi) \cup (-\pi, -\pi + \theta_0)$, we have

- $2\theta \in (2\pi - 2\theta_0, 2\pi) \cup (-2\pi, -2\pi + 2\theta_0) \subseteq \left(2\pi - \frac{\pi}{3}, 2\pi\right) \cup \left(-2\pi, -2\pi + \frac{\pi}{3}\right)$ thus $\cos 2\theta > \cos 2\theta_0 > 0$;
- $3\theta \in (3\pi - 3\theta_0, 3\pi) \cup (-3\pi, -3\pi + 3\theta_0) \subseteq \left(3\pi - \frac{\pi}{2}, 3\pi\right) \cup \left(-3\pi, -3\pi + \frac{\pi}{2}\right)$, thus $-\cos 3\theta > -\cos(3\pi - 3\theta_0) = \sin\left(\frac{\pi}{2} - 3\theta_0\right) \geq 0$;

So we have

$$-\Re(\lambda^3 - \lambda^2) = R^2(-R\cos 3\theta + \cos 2\theta) > \left[\sin\left(\frac{\pi}{2} - 3\theta_0\right) + \cos 2\theta_0\right] R^2 > 0.$$

Finally, $\frac{|\Re(\lambda-1)|}{-\Re(\lambda^3-\lambda^2)} \leq \frac{R+1}{\left[\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0\right]R^2} \leq \frac{2}{\sin\left(\frac{\pi}{2}-3\theta_0\right)+\cos 2\theta_0}$ and similarly for $\frac{|\Im(\lambda-1)|}{-\Re(\lambda^3-\lambda^2)}$. □

Lemma A.5. *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

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

(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. The proofs for (i) and (ii) are similar to those in Lemma A.4.
(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 $f_1(R) = \left(\frac{aR-1}{bR-a}\right)^2$ where $a = \cos\theta+\gamma_3\sin\theta$, $b = \cos 2\theta+\gamma_3\sin 2\theta$. 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-Schwarz 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 $f_2(R) = \left(\frac{cR-\gamma_3}{bR-a}\right)^2$ where $c = \gamma_3\cos\theta-\sin\theta$ 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 $1 \le R < \frac{\gamma_3}{c}$ and increases for $R > \frac{\gamma_3}{c}$. Moreover, like for $f_1(1)$, we can estimate

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

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

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

(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 $\Re(\lambda^2-\lambda) = R(R\cos 2\theta - \cos\theta) > 0$. □

B Descent step for usual and shifted gradient descent

Proposition B.1 (Descent step for the usual gradient descent)**.** *The usual gradient descent algorithm* (5) *converges if*

$$0 < \tau < \frac{2}{\|H(I-B)^{-1}M\|^2}.$$

Proof. The error system for (5) can be rewritten as

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} -\tau(I-B^*)^{-1}H^*H(I-B)^{-1}MM^* & 0 & (I-B^*)^{-1}H^*H(I-B)^{-1}M \\ -\tau(I-B)^{-1}MM^* & 0 & (I-B)^{-1}M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix} \tag{67}$$

Recall that a fixed point iteration converges if and only if the spectral radius of its iteration matrix is strictly less than 1. We can show that:

(i) If $\lambda \in \mathbb{C}\setminus\{0,1\}$ is an eigenvalue of the iteration matrix, then, proceeding as in Proposition 4.3, there exists $y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that

$$\lambda^2(\lambda-1) + \tau\frac{\left\|H(I-B)^{-1}My\right\|^2}{\|y\|^2}\lambda^2 = 0 \tag{68}$$

hence $\lambda = 1 - \tau\frac{\left\|H(I-B)^{-1}My\right\|^2}{\|y\|^2}$. If we take $\tau < \frac{2}{\|H(I-B)^{-1}M\|^2}$ then equation (68) admits no solution λ with $|\lambda| \ge 1$.

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix. To show this, we rewrite iteration (67) as

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

□

Proposition B.2 (Convergence of the shifted gradient descent)**.** *The shifted gradient descent algorithm* (6) *converges if*

$$0 < \tau < \frac{1}{\|H(I-B)^{-1}M\|^2}.$$

Proof. The error system for (6) can be rewritten as

$$\begin{bmatrix} p^{n+1} \\ u^{n+1} \\ \sigma^{n+1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & (I-B^*)^{-1}H^*H(I-B)^{-1}M \\ 0 & 0 & (I-B)^{-1}M \\ -\tau M^* & 0 & I \end{bmatrix} \begin{bmatrix} p^n \\ u^n \\ \sigma^n \end{bmatrix}. \tag{69}$$

Recall that a fixed point iteration converges if and only if the spectral radius of its iteration matrix is strictly less than 1. We can show that:

(i) If $\lambda \in \mathbb{C} \setminus \{0, 1\}$ is an eigenvalue of the iteration matrix, then, proceeding as in Proposition 4.2, there exists $y \in \mathbb{C}^{n_\sigma}, y \neq 0$ such that

$$\lambda^2(\lambda - 1) + \tau \frac{\left\|H(I-B)^{-1}My\right\|^2}{\|y\|^2}\lambda = 0. \tag{70}$$

By applying Lemma C.1 for

$$a_0 = 0, \quad a_1 = \tau \frac{\left\|H(I-B)^{-1}My\right\|^2}{\|y\|^2}, \quad a_2 = -1,$$

we see that equation (70) admits no solution λ with $|\lambda| \geq 1$ if we take $\tau < \frac{\|y\|^2}{\|H(I-B)^{-1}My\|^2}$. Then it is enough to take $\tau < \frac{1}{\|H(I-B)^{-1}M\|^2}$.

(ii) $\lambda = 1$ is not an eigenvalue of the iteration matrix. To show this, we rewrite iteration (69) as

$$\begin{bmatrix} \sigma^{n+1} \\ p^{n+1} \\ u^{n+1} \end{bmatrix} = \begin{bmatrix} I & -\tau M^* & 0 \\ (I-B^*)^{-1}H^*H(I-B)^{-1}M & 0 & 0 \\ (I-B)^{-1}M & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma^n \\ p^n \\ u^n \end{bmatrix}.$$

and proceed as in Proposition 4.2.

□

C Convergence study for the scalar case

C.1 Notations and preliminary calculation

In the scalar case, that is when $n_u, n_\sigma, n_f = 1$, we change the notation from capital to lower case letters:

$$B \leftarrow b \in \mathbb{R}, b < 1, \quad M \leftarrow m \in \mathbb{R}, m \neq 0, \quad H \leftarrow h \in \mathbb{R}, h \neq 0,$$

$$T_k \leftarrow t_k = 1 + b + ... + b^{k-1} = \frac{1-b^k}{1-b}, \quad U_k \leftarrow u_k = kh^2 b^{k-1} \tag{71}$$

$$X_k \leftarrow x_k = \begin{cases} 0, & k = 1, \\ h^2[1 + 2b + 3b^2 + ... + (k-1)b^{k-2}], & k \geq 2. \end{cases}$$

The identity $1 + 2x + 3x^2 + ... + nx^{n-1} = \left(\frac{1-x^{n+1}}{1-x}\right)' = \frac{1-(n+1)x^n + nx^{n+1}}{(1-x)^2}$ says that

$$x_k = h^2 \frac{1 - kb^{k-1} + (k-1)b^k}{(1-b)^2}, \quad k \geq 1, \tag{72}$$

where we set $b^{k-1} = 1$ when $k = 1$ and $b = 0$. Now for each of algorithms (5), (6), (10), (9), we write the iterations for the errors in the scalar case and the corresponding iteration matrix $\mathcal{M}$ such that $[p^{n+1}, u^{n+1}, \sigma^{n+1}]^\intercal = \mathcal{M}[p^n, u^n, \sigma^n]^\intercal$.

- Usual gradient descent (usual GD):

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau m p^n \\ u^n = b u^n + m\sigma^n \\ p^n = b p^n + h^2 u^n \end{cases} \qquad \mathcal{M} = \begin{bmatrix} -h^2 m^2 (1-b)^{-2}\tau & 0 & h^2 m (1-b)^{-2} \\ -m^2 (1-b)^{-1}\tau & 0 & m(1-b)^{-1} \\ -m\tau & 0 & 1 \end{bmatrix} \tag{73}$$

- Shifted gradient descent (shifted GD):

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau m p^n \\ u^{n+1} = b u^{n+1} + m\sigma^n \\ p^{n+1} = b p^{n+1} + h^2 u^{n+1} \end{cases} \qquad \mathcal{M} = \begin{bmatrix} 0 & 0 & h^2 m (1-b)^{-2} \\ 0 & 0 & m(1-b)^{-1} \\ -m\tau & 0 & 1 \end{bmatrix} \tag{74}$$

- k-step one-shot:

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau m p^n \\ p^{n+1} = (b^k - \tau m^2 x_k) p^n + u_k u^n + m x_k \sigma^n \\ u^{n+1} = b^k u^n + m t_k \sigma^n - \tau m^2 t_k p^n \end{cases} \qquad \mathcal{M} = \begin{bmatrix} b^k - m^2 x_k \tau & u_k & m x_k \\ -m^2 t_k \tau & b^k & m t_k \\ -m\tau & 0 & 1 \end{bmatrix} \tag{75}$$

- Shifted k-step one-shot:

$$\begin{cases} \sigma^{n+1} = \sigma^n - \tau m p^n \\ p^{n+1} = b^k p^n + u_k u^n + m x_k \sigma^n \\ u^{n+1} = b^k u^n + m t_k \sigma^n \end{cases} \qquad \mathcal{M} = \begin{bmatrix} b^k & u_k & m x_k \\ 0 & b^k & m t_k \\ -m\tau & 0 & 1 \end{bmatrix}. \tag{76}$$

C.2 Necessary and sufficient conditions for convergence

In this simpler scalar case, we will be able to prove sufficient and also necessary conditions on the descent step τ for convergence. Our strategy to study the spectral radius $\rho(\mathcal{M})$ is as follows:

1. Compute $\det(\mathcal{M} - \lambda I)$ to write the eigenvalue equation $\mathcal{P}(\lambda) = 0$. For the considered methods, $\mathcal{P}$ turns out to be a polynomial of degree 3, $\mathcal{P}(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 + \lambda^3$, where $a_0, a_1, a_2 \in \mathbb{R}$ depend on h, m, b, τ. For the computations, the identity $u_k t_k - b^k x_k + x_k = h^2 t_k^2$, which is the scalar version of (41), can be helpful.

2. Apply to $\mathcal{P}$ Lemma C.1, which states a necessary and sufficient condition for a real coefficient polynomial of degree 3 to have all roots inside the unit circle of the complex plane. Then deduce conditions on τ.

Lemma C.1. *Let $a_0, a_1, a_2 \in \mathbb{R}$, then all roots of $\mathcal{P}(z) = a_0 + a_1 z + a_2 z^2 + z^3$ stay (strictly) inside the unit circle of the complex plane if and only if*

$$(a_0 - 1)(a_0 + 1) < 0, \tag{77}$$

$$(a_0^2 - a_2 a_0 + a_1 - 1)(a_0^2 + a_2 a_0 - a_1 - 1) > 0, \tag{78}$$

$$(a_0 + a_2 - a_1 - 1)(a_0 + a_2 + a_1 + 1) < 0. \tag{79}$$

The proof of Lemma C.1 is in Appendix D and is mainly based on Marden's works [17].

C.2.1 Descent step for the usual gradient descent

Here, the coefficients of $\mathcal{P}$ are

$$a_0 = 0, \quad a_1 = 0, \quad a_2 = h^2 m^2 (1 - b)^{-2}\tau - 1.$$

Conditions (77) and (78) of Lemma C.1 are automatically satisfied. Condition (79) gives

$$0 < \tau < \frac{2(1 - b)^2}{h^2 m^2},$$

that is (7) in the scalar case.

C.2.2 Descent step for the shifted gradient descent

Here, the coefficients of $\mathcal{P}$ are

$$a_0 = 0, \quad a_1 = h^2 m^2 (1 - b)^{-2}\tau, \quad a_2 = -1.$$

Condition (77) of Lemma C.1 is automatically satisfied, condition (79) is automatically satisfied for $\tau > 0$, and condition (78) gives us

$$\tau < \frac{(1 - b)^2}{h^2 m^2},$$

that is (8) in the scalar case.

C.2.3 Descent step for k-step one-shot

Here, the coefficients of $\mathcal{P}$ are

$$a_0 = -s^2, \quad a_1 = m^2(h^2 t_k^2 - x_k)\tau + (s^2 + 2s), \quad a_2 = m^2 x_k \tau - (2s+1)$$

where $s = b^k$. Condition (77) of Lemma C.1 is obviously satisfied since $|b| < 1$. Next we deal with condition (78). The computation shows that

$$a_0^2 - a_2 a_0 + a_1 - 1 = m^2(h^2 t_k^2 - x_k + x_k s^2)\tau + \underbrace{(s-1)^3(s+1)}_{<0}, \tag{80}$$

$$a_0^2 + a_2 a_0 - a_1 - 1 = -m^2(h^2 t_k^2 - x_k + x_k s^2)\tau + \underbrace{(s-1)(s+1)^3}_{<0} \tag{81}$$

and

$$h^2 t_k^2 - x_k + x_k s^2 = \frac{h^2 b^{k-1}(1-b^k)[k-(k+1)b + kb^k - (k-1)b^{k+1}]}{(1-b)^2}. \tag{82}$$

Lemma C.2. *$k-(k+1)b+kb^k-(k-1)b^{k+1} > 0$, $\forall |b| < 1$, $\forall k \geq 1$.*

Proof. We write $k-(k+1)b+kb^k-(k-1)b^{k+1} = (1-b)A$ where $A = k+1-\frac{1-b^k}{1-b}+(k-1)b^k$. It suffices to show $A > 0$. If $k = 1$ then $A = 1 > 0$. If either k is even, or $k \geq 3$ is odd and $0 \leq b < 1$, then $(k-1)b^k \geq 0$ and $\left|\frac{1-b^k}{1-b}\right| = |b^{k-1} + b^{k-2} + ... + b + 1| \leq |b^{k-1}| + |b^{k-2}| + ... + |b| + 1 < k$ give us the conclusion. If $k \geq 3$ is odd and $-1 < b < 0$ then $(k-1)(1+b^k+1) > 0$ and $\frac{1-b^k}{1-b} < 1$ therefore $A = 1 + \left(1 - \frac{1-b^k}{1-b}\right) + (k-1)(1+b^k) > 0$. □

Then, condition (78) imposes

- $\tau < \frac{(1-b)^2(1+b^k)(1-b^k)^2}{h^2 m^2 b^{k-1}[k-(k+1)b+kb^k-(k-1)b^{k+1}]}$ if $b^{k-1} > 0$;
- $\tau < \frac{(1-b)^2(1+b^k)^3}{h^2 m^2 b^{k-1}[-k+(k+1)b-kb^k+(k-1)b^{k+1}]}$ if $b^{k-1} < 0$;
- no condition on τ if $k \geq 2$ and $b = 0$.

Finally we check condition (79). We have $a_0 + a_2 + a_1 + 1 = h^2 m^2 t_k^2 \tau > 0$ and

$$a_0 + a_2 - a_1 - 1 = \frac{h^2 m^2 (1 - 2kb^{k-1} + 2kb^k - b^{2k})}{(1-b)^2}\tau - 2(1+s)^2,$$

therefore, condition (79) gives

- $\tau < \frac{2(1-b)^2(1+b^k)^2}{h^2 m^2 (1-2kb^{k-1}+2kb^k-b^{2k})}$ if $1 - 2kb^{k-1} + 2kb^k - b^{2k} > 0$;
- no condition on τ if $1 - 2kb^{k-1} + 2kb^k - b^{2k} \leq 0$.

In the following lemma we study the quantity $1 - 2kb^{k-1} + 2kb^k - b^{2k}$ that appears above.

Lemma C.3. *Let $f_k(b) = 1 - 2kb^{k-1} + 2kb^k - b^{2k}$ for $k \in \mathbb{N}^*$ and $-1 \leq b \leq 1$.*

(i) *$f_1(b) = -(1-b)^2 < 0, \forall -1 < b < 1$.*

(ii) *$f_2(b) = 1 - 4b + 4b^2 - b^4$ has a unique solution $b = -1 + \sqrt{2}$ in $(-1, 1)$; and $f_2(b) > 0$ if $-1 < b < -1 + \sqrt{2}$, $f_2(b) < 0$ if $-1 + \sqrt{2} < b < 1$.*

(iii) If $k \geq 3$ is odd then $f_k(b)$ has exactly two solutions $b_1(k) < b_2(k)$ in $(-1,1)$; if $k \geq 2$ is even then $f_k(b)$ has a unique solution $b_3(k)$ in $(-1,1)$. Moreover, for every odd $k \geq 3$:

- $-1 < b_1(k) < 0 < b_2(k) < 1$;
- $f_k(b) > 0 \Leftrightarrow b_1(k) < b < b_2(k)$;
- $f_k(b) < 0 \Leftrightarrow -1 < b < b_1(k) \vee b_2(k) < b < 1$.

and for every even $k \geq 2$:

- $0 < b_3(k) < 1$;
- $f_k(b) > 0 \Leftrightarrow -1 < b < b_3(k)$;
- $f_k(b) < 0 \Leftrightarrow b_3(k) < b < 1$.

(iv) $\lim\limits_{\substack{k \ odd \\ k\to\infty}} b_1(k) = -1$ *and* $\lim\limits_{\substack{k \ odd \\ k\to\infty}} b_2(k) = 1 = \lim\limits_{\substack{k \ even \\ k\to\infty}} b_3(k) = 1$.

Proof. (i) and (ii) are easy to verify. (iii) It remains to consider $k \geq 3$. We have

$$f_k'(b) = b^{k-2}\left[-2k(k-1) + 2k^2 b - 2kb^{k+1}\right], -1 < b < 1.$$

Set

$$g_k(b) = -2k(k-1) + 2k^2 b - 2kb^{k+1}, \quad -1 \leq b \leq 1, k \geq 3.$$

Case 1. [$k \geq 3$ is odd] By studying the sign of $g_k'(b)$, we find that

- g_k has a unique solution $v_1(k)$ in $(-1,1)$ and $0 < v_1(k) < \sqrt[k]{\frac{k}{k+1}} < 1$;
- $g_k(b) > 0 \Leftrightarrow v_1(k) < b < 1$;
- $g_k(b) < 0 \Leftrightarrow -1 < b < v_1(k)$.

Next, by studying the sign of $f_k'(b)$, we find that

- $f_k(b)$ has exactly two solutions $b_1(k) < b_2(k)$ in $(-1,1)$ and $-1 < b_1(k) < 0 < b_2(k) < 1$;
- $f_k(b) > 0 \Leftrightarrow b_1(k) < b < b_2(k)$;
- $f_k(b) < 0 \Leftrightarrow -1 < b < b_1(k) \vee b_2(k) < b < 1$.

Case 2. [$k \geq 4$ is even] By studying the sign of $g_k'(b)$, we find that

- g_k has a unique solution $v_2(k)$ in $(-1,1)$ and $0 < v_2(k) < \sqrt[k]{\frac{k}{k+1}} < 1$;
- $g_k(b) > 0 \Leftrightarrow v_2(k) < b < 1$;
- $g_k(b) < 0 \Leftrightarrow 0 < b < v_2(k)$.

Next, by studying the sign of $f_k'(b)$, we find that

- $f_k(b)$ has a unique solution $b_3(k)$ in $(-1,1)$ and $0 < b_3(k) < 1$;
- $f_k(b) > 0 \Leftrightarrow -1 < b < b_3(k)$;
- $f_k(b) < 0 \Leftrightarrow b_3(k) < b < 1$.

(iv) We have

$$f_k\left(\frac{1}{2}\right) = 1 - \frac{k}{2^{k-1}} - \frac{1}{2k}, \quad \forall k \geq 3 \quad \text{and} \quad f_k\left(-\frac{1}{2}\right) = 1 - \frac{3k}{2^{k-1}} - \frac{1}{2k}, \quad \forall \text{ odd } k \geq 3,$$

hence for sufficiently large k we have $f_k\left(\frac{1}{2}\right) > 0$ and for sufficiently large odd k we have $f_k\left(-\frac{1}{2}\right) > 0$. By the table of signs of f_k, we conclude that $b_1(k) < -\frac{1}{2}$ for large odd k, $b_2(k) > \frac{1}{2}$ for large odd k and $b_3(k) > \frac{1}{2}$ for large even k.
Case 1. [$k \geq 3$ is odd and sufficiently large] First we work with $b_1(k)$. We have

$$1 - 2kb_1(k)^{k-1} + 2kb_1(k)^k - b_1(k)^{2k} = 0$$

and $b_1(k) < -\frac{1}{2}$ so

$$-b_1(k)^{2k} + 2kb_1(k)^k + 1 = 2kb_1(k)^{k-1} = \underbrace{[-2kb_1(k)^k]}_{>0} \cdot \frac{1}{-b_1(k)} < [-2kb_1(k)^k].2 = -4kb_1(k)^k,$$

which leads to

$$b_1(k)^{2k} - 6kb_1(k)^k - 1 > 0 \Leftrightarrow [b_1(k)^k - 3k]^2 > 1 + 9k^2.$$

Since $-1 < b_1(k) < 0$ and k is odd, this tells us that

$$-1 < b_1(k) < -(-3k + \sqrt{1+9k^2})^{1/k} = \frac{-1}{(3k + \sqrt{1+9k^2})^{1/k}} < \frac{-1}{(7k)^{1/k}},$$

which yields $\lim\limits_{\substack{k \text{ odd} \\ k\to\infty}} b_1(k) = -1$. Next, we have

$$1 - 2kb_2(k)^{k-1} + 2kb_2(k)^k - b_2(k)^{2k} = 0$$

and $b_2(k) > \frac{1}{2}$ so

$$-b_2(k)^{2k} + 2kb_2(k)^k + 1 = 2kb_2(k)^{k-1} = 2kb_2(k)^k \cdot \frac{1}{b_2(k)} < 4kb_2(k)^k,$$

which leads to

$$b_2(k)^{2k} + 2kb_2(k)^k - 1 > 0 \Leftrightarrow [b_2(k)^k + k]^2 > 1 + k^2.$$

Since $0 < b_2(k) < 1$, this tells us that

$$1 > b_2(k) > (-k + \sqrt{1+k^2})^{1/k} = \frac{1}{(k + \sqrt{1+k^2})^{1/k}} > \frac{1}{(3k)^{1/k}},$$

which yields $\lim\limits_{\substack{k \text{ even} \\ k\to\infty}} b_2(k) = 1$.
Case 2. [$k \geq 4$ is even and sufficiently large] We repeat the same arguments as $b_2(k)$ for $b_3(k)$. □

In summary, we have the following proposition.

Proposition C.4 (Convergence of k-step one-shot)**.** *Let* $\eta_1(k,b) \coloneqq +\infty$ *and*

$$\eta_{21}(k,b) \coloneqq \frac{(1-b)^2(1+b^k)(1-b^k)^2}{b^{k-1}[k-(k+1)b+kb^k-(k-1)b^{k+1}]};$$

$$\eta_{22}(k,b) \coloneqq \frac{-(1-b)^2(1+b^k)^3}{b^{k-1}[k-(k+1)b+kb^k-(k-1)b^{k+1}]};$$

$$\eta_3(k,b) \coloneqq \frac{2(1-b)^2(1+b^k)^2}{1-2kb^{k-1}+2kb^k-b^{2k}}$$

then the necessary and sufficient condition for the convergence of k-step one-shot in the scalar case is of the form $\tau < \frac{\eta(k,b)}{h^2m^2}$ where $\eta(k,b)$ is defined as follows:

(i) $\eta(1,b) = \eta_{21}(1,b) = (1-b)^3(1+b), -1 < b < 1$;

(ii) for odd $k \geq 3$,

$$\eta(k,b) = \begin{cases} \eta_{21}(k,b), & -1 < b \leq b_1(k) \vee b_2(k) \leq b < 1, \\ \min\left\{\eta_{21}(k,b), \eta_3(k,b)\right\}, & b_1(k) < b < b_2(k) \wedge b \neq 0, \\ 2, & b = 0 \end{cases}$$

where $-1 < b_1(k) < 0 < b_2(k) < 1$ *are the two solutions of*

$$1-2kb^{k-1}+2kb^k-b^{2k} = 0, \quad -1 < b < 1;$$

(iii) for even $k \geq 2$,

$$\eta(k,b) = \begin{cases} \eta_{21}(k,b), & b_3(k) \leq b < 1, \\ \min\left\{\eta_{21}(k,b), \eta_3(k,b)\right\}, & 0 < b < b_3(k), \\ 2, & b = 0, \\ \min\left\{\eta_{22}(k,b), \eta_3(k,b)\right\}, & -1 < b < 0 \end{cases}$$

where $0 < b_3(k) < 1$ *is the unique solution of*

$$1-2kb^{k-1}+2kb^k-b^{2k} = 0, \quad -1 < b < 1.$$

Note that $\lim\limits_{\substack{k \text{ odd} \\ k\to\infty}} b_1(k) = -1$ and $\lim\limits_{\substack{k \text{ odd} \\ k\to\infty}} b_2(k) = 1 = \lim\limits_{\substack{k \text{ even} \\ k\to\infty}} b_3(k)$, so the behavior of τ when $k \to \infty$ is consistent with the result $\tau < \frac{2(1-b)^2}{h^2m^2}, -1 < b < 1$ for the usual gradient descent. For illustrations of the function $\eta(k,b)$ for different k see section C.3.

C.2.4 Descent step for shifted k-step one-shot

Here, the coefficients of the polynomial $\mathcal{P}$ of the eigenvalue equation are

$$a_0 = h^2m^2v_k\tau - s^2, \quad a_1 = h^2m^2y_k\tau + s^2 + 2s, \quad a_2 = -2s-1, \tag{83}$$

where $s = b^k$, $y_k = \frac{x_k}{h^2} = \frac{1-kb^{k-1}+(k-1)b^k}{(1-b)^2}$ and $v_k = t_k^2 - y_k = \frac{b^{k-1}[k-(k+1)b+b^{k+1}]}{(1-b)^2}$. Note that v_k and b^{k-1} have the same sign, also $v_k = 0$ if and only if $k \geq 2$ and $b = 0$, since it is easy to show that $k-(k+1)b+b^{k+1} > 0, \forall |b| < 1, \forall k \geq 1$. Then, condition (77) of Lemma C.1 imposes

- $\tau < \frac{1+s^2}{h^2m^2v_k} = \frac{(1-b)^2(1+b^{2k})}{h^2m^2b^{k-1}[k-(k+1)b+b^{k+1}]}$ if $b^{k-1} > 0$;
- $\tau < \frac{-1+s^2}{h^2m^2v_k} = \frac{(1-b)^2(-1+b^{2k})}{h^2m^2b^{k-1}[k-(k+1)b+b^{k+1}]}$ if $b^{k-1} < 0$;
- no condition on τ if $k \geq 2$ and $b = 0$.

Next we study condition (78). We have

$$a_0^2 - a_2 a_0 + a_1 - 1 = v_k^2 (h^2 m^2 \tau)^2 + [(-2s^2 + 2s + 1)v_k + y_k] h^2 m^2 \tau + \underbrace{(s-1)^3 (s+1)}_{<0}$$

and

$$a_0^2 + a_2 a_0 - a_1 - 1 = v_k^2 (h^2 m^2 \tau)^2 - [(2s^2 + 2s + 1)v_k + y_k] h^2 m^2 \tau + \underbrace{(s-1)(s+1)^3}_{<0},$$

each of which, considered as a second order polynomial of $h^2 m^2 \tau$ if $v_k \neq 0$, has exactly two roots of opposite signs. Therefore if $v_k \neq 0$, condition (78) is equivalent to $(h^2 m^2 \tau - r_1)(h^2 m^2 \tau - r_2) > 0$ where

$$r_1 := \frac{(2s^2 - 2s - 1)v_k - y_k + \sqrt{(-4s+5)v_k^2 + y_k^2 + 2(-2s^2 + 2s + 1)v_k y_k}}{2v_k^2} > 0$$

and

$$r_2 := \frac{(2s^2 + 2s + 1)v_k + y_k + \sqrt{(8s^2 + 12s + 5)v_k^2 + y_k^2 + 2(2s^2 + 2s + 1)v_k y_k}}{2v_k^2} > 0.$$

Lemma C.5. *r_1 and r_2 cannot be both strictly less than $\frac{1+s^2}{v_k}$. r_1 and r_2 cannot be both strictly less than $\frac{-1+s^2}{v_k}$.*

Proof. Either $r_1 < \frac{1+s^2}{v_k}$ or $r_1 < \frac{-1+s^2}{v_k}$ implies $(s^2 + 4s + 1)v_k^2 + (s^2 + 1)v_k y_k > 0$. Either $r_2 < \frac{1+s^2}{v_k}$ or $r_2 < \frac{-1+s^2}{v_k}$ implies $(s^2 + 4s + 1)v_k^2 + (s^2 + 1)v_k y_k < 0$. □

Thanks to this lemma we see that condition (78), in combination with condition (77), gives

- $\tau < \frac{1}{h^2 m^2} \min\{r_1, r_2\}$ if $b^{k-1} \neq 0$;
- $\tau < \frac{1}{h^2 m^2}$ if $k \geq 2$ and $b = 0$.

Finally, we have $a_0 + a_2 + a_1 + 1 = h^2 m^2 t_k^2 \tau > 0$ and

$$a_0 + a_2 - a_1 - 1 = \frac{h^2 m^2}{(1-b)^2}[-1 + 2kb^{k-1} - 2kb^k + b^{2k}]\tau - 2(1 - b^k)^2,$$

thus condition (79) is equivalent to

- $\tau < \frac{2(1-b)^2(1-b^k)^2}{h^2 m^2 (-1 + 2kb^{k-1} - 2kb^k + b^{2k})}$ if $1 - 2kb^{k-1} + 2kb^k - b^{2k} < 0$;
- no condition on τ if $1 - 2kb^{k-1} + 2kb^k - b^{2k} \geq 0$.

One can look again at Lemma C.3 for the analysis of $1 - 2kb^{k-1} + 2kb^k - b^{2k}$. In summary, we have the following proposition.

Proposition C.6 (Convergence of shifted k-step one-shot)**.** *Let*

$$\kappa_{11}(k, b) := \frac{(1-b)^2(1+b^{2k})}{b^{k-1}[k - (k+1)b + b^{k+1}]};$$

$$\kappa_{12}(k,b) \coloneqq \frac{(1-b)^2(-1+b^{2k})}{b^{k-1}[k-(k+1)b+b^{k+1}]};$$

$$t_k \coloneqq \frac{1-b^k}{1-b}, \quad y_k \coloneqq \frac{1-kb^{k-1}+(k-1)b^k}{(1-b)^2}, \quad s \coloneqq b^k, \quad v_k \coloneqq t_k^2 - y_k,$$

$$\kappa_{21}(k,b) \coloneqq \frac{(2s^2-2s-1)v_k - y_k + \sqrt{(-4s+5)v_k^2 + y_k^2 + 2(-2s^2+2s+1)v_k y_k}}{2v_k^2},$$

$$\kappa_{22}(k,b) \coloneqq \frac{(2s^2+2s+1)v_k + y_k + \sqrt{(8s^2+12s+5)v_k^2 + y_k^2 + 2(2s^2+2s+1)v_k y_k}}{2v_k^2},$$

$$\kappa_2(k,b) \coloneqq \min\{\kappa_{21}(k,b), \kappa_{22}(k,b)\};$$

$$\kappa_3(k,b) \coloneqq \frac{2(1-b)^2(1-b^k)^2}{-1+2kb^{k-1}-2kb^k+b^{2k}}$$

then the necessary and sufficient condition for the convergence of shifted k-step one-shot in the scalar case is of the form $\tau < \frac{\kappa(k,b)}{h^2m^2}$ *where* $\kappa(k,b)$ *is defined as follows:*

(i) $\kappa(1,b) = \min\{\kappa_{11}(1,b), \kappa_2(1,b), \kappa_3(1,b)\}$*, also note that*

$$\kappa_{11}(1,b) = 1+b^2, \quad \kappa_{21}(1,b) = \frac{2b^2-2b-1+\sqrt{-4b+5}}{2},$$

$$\kappa_{22}(1,b) = \frac{2b^2+2b+1+\sqrt{8b^2+12b+5}}{2}, \quad \kappa_3(1,b) = 2(1-b)^2;$$

(ii) for odd $k \geq 3$*,*

$$\kappa(k,b) = \begin{cases} \min\{\kappa_{11}(k,b), \kappa_2(k,b), \kappa_3(k,b)\}, & -1 < b < b_1(k) \vee b_2(k) < b < 1, \\ \min\{\kappa_{11}(k,b), \kappa_2(k,b)\}, & b_1(k) \leq b \leq b_2(k) \wedge b \neq 0, \\ 1, & b = 0 \end{cases}$$

where $-1 < b_1(k) < 0 < b_2(k) < 1$ *are the two solutions of*

$$1 - 2kb^{k-1} + 2kb^k - b^{2k} = 0, \quad -1 < b < 1;$$

(iii) for even $k \geq 2$*,*

$$\kappa(k,b) = \begin{cases} \min\{\kappa_{11}(k,b), \kappa_2(k,b), \kappa_3(k,b)\}, & b_3(k) < b < 1, \\ \min\{\kappa_{11}(k,b), \kappa_2(k,b)\}, & 0 < b \leq b_3(k), \\ 1, & b = 0, \\ \min\{\kappa_{12}(k,b), \kappa_2(k,b)\}, & -1 < b < 0 \end{cases}$$

where $0 < b_3(k) < 1$ *is the unique solution of*

$$1 - 2kb^{k-1} + 2kb^k - b^{2k} = 0, \quad -1 < b < 1.$$

Remark C.7. In implementation, we rewrite $\kappa_{21}(k,b)$ as

$$\frac{b(1-b)^2(b^k-1)}{k-(k+1)b+b^{k+1}} + \frac{2\cdot\left[-b^k+1+\frac{b(1-b)^2(1-b^k)y_k}{k-(k+1)b+b^{k+1}}\right]}{y_k+v_k+\sqrt{(-4s+5)v_k^2+y_k^2+2(-2s^2+2s+1)v_k y_k}}$$

to avoid numerical errors. Also in this formula, we see that $\kappa_{21}(k,b) \overset{k\to\infty}{\longrightarrow} (1-b)^2$ (note that $y_k = \frac{1-kb^{k-1}+(k-1)b^k}{(1-b)^2} \overset{k\to\infty}{\longrightarrow} \frac{1}{(1-b)^2}$ and $v_k = t_k^2 - y_k \overset{k\to\infty}{\longrightarrow} 0$).

For illustrations of the function $\kappa(k,b)$ for different k see section C.3.

C.3 Comparison of the bounds for the descent step

In summary, in the scalar case, the necessary and sufficient convergence conditions on the descent step $\tau > 0$ are:

$$\tau < \frac{2(1-b)^2}{h^2m^2}, \qquad \tau < \frac{(1-b)^2}{h^2m^2}, \qquad \tau < \frac{\eta(k,b)}{h^2m^2}, \qquad \tau < \frac{\kappa(k,b)}{h^2m^2},$$

respectively for usual GD, shifted GD, k-step one-shot (with $\eta(k,b)$ given in Proposition C.4), shifted k-step one-shot (with $\kappa(k,b)$ given in Proposition C.6). By taking $m = h = 1$, in Figure 5 we plot for different k the functions: $b \mapsto 2(1-b)^2$ (usual GD), $b \mapsto (1-b)^2$ (shifted GD), $b \mapsto \eta(k,b)$ (k-step one-shot) and $b \mapsto \kappa(k,b)$ (shifted k-step one-shot).

From these plots we can draw two important conclusions. First, when k increases the visualized curves for k-step one-shot and shifted k-step one-shot tend to the corresponding curves for usual and shifted gradient descent, as expected. Second, even in this scalar case, it appears difficult to establish a simplified expression for $\eta(k,b)$ in Proposition C.4 and $\kappa(k,b)$ in Proposition C.6 to find a practical upper bound for the descent step τ.

Remark C.8. For $k \geq 2$, we observe that for some b the admissible range of τ of k-step one-shot is larger than the one of usual GD, that is not intuitive. This is indeed verified numerically using FreeFEM: when $b = 0.2$ and $\tau = 2.08$, 2-step one-shot converges while the usual GD does not.

D A proof of Lemma C.1 based on Marden's works

Definition D.1. We say that a complex coefficient polynomial has property $\mathscr{P}$ if all its zeros lie (strictly) inside the unit circle $|z| = 1$.

We recall some definitions from Marden's works [17].

Definition D.2. Let $P(z) = a_0 + a_1 z + ... + a_n z^n$ where $a_k \in \mathbb{R}, k = 0, ..., n$ (we do not require $a_n \neq 0$ here). We define

$$\tilde{P}(z) \coloneqq a_n + a_{n-1} z + ... + a_0 z^n$$

and call it the reverse polynomial of P. One can also see that $\tilde{P}(z) = z^n P(1/z)$.

Definition D.3. Let $P(z) = a_0 + a_1 z + ... + a_n z^n$ where $a_k \in \mathbb{R}, k = 0, ..., n$. We define a polynomial sequence $\{P_k\}_{0 \leq k \leq n}$ where

$$P_k(z) = a_0^{(k)} + a_1^{(k)} z + ... + a_{n-k}^{(k)} z^{n-k}$$

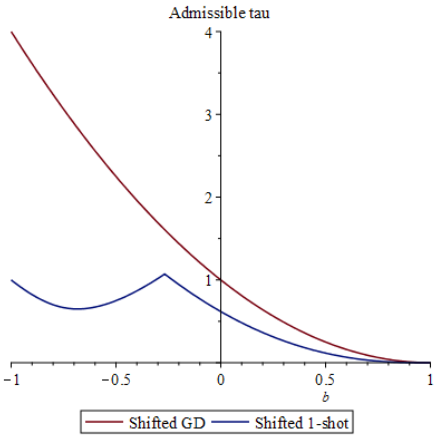
as follows:

- $P_0 = P$;
- $P_{k+1} = a_0^{(k)} P_k - a_{n-k}^{(k)} \tilde{P}_k$ for $0 \leq k \leq n-1$.

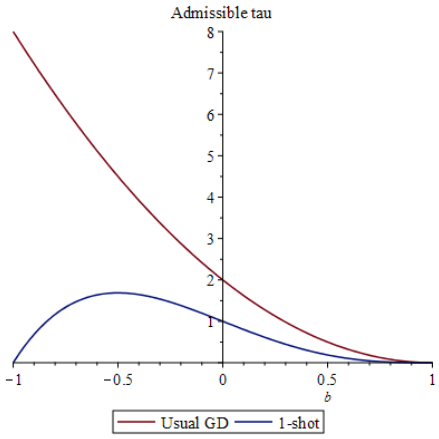
Then we define

$$m_k(P) = a_0^{(1)} a_0^{(2)} \cdots a_0^{(k)}, \quad 1 \leq k \leq n.$$

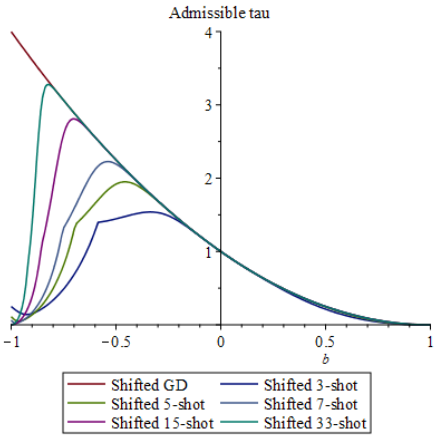
The coefficients of these polynomials can be gathered in the following table, that we call Marden's table:



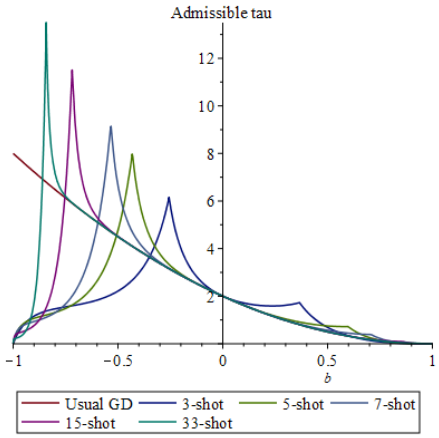
(a) Shifted 1-step one-shot



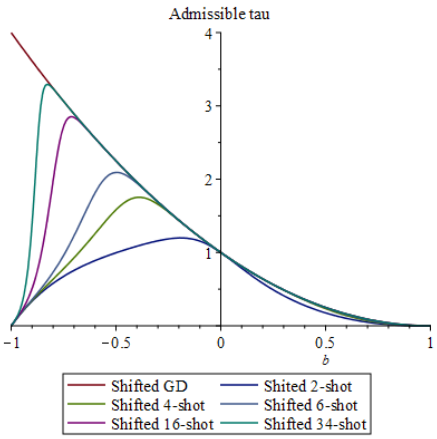
(b) 1-step one-shot



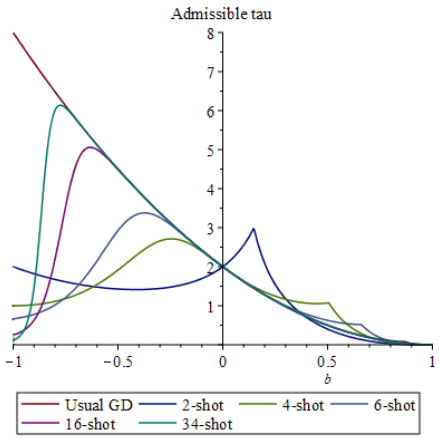
(c) Shifted k-step one-shot, odd $k \geq 3$



(d) k-step one-shot, odd $k \geq 3$



(e) Shifted k-step one-shot, even $k \geq 2$



(f) k-step one-shot, even $k \geq 2$

Figure 5: Admissible τ in the scalar case as a function of b.

	1	x	x^2	...	x^{n-1}	x^n
P_0	a_0	a_1	a_2	...	a_{n-1}	a_n
$\tilde{P}_0$	a_n	a_{n-1}	a_{n-2}	...	a_1	a_0
P_1	$a_0^{(1)}$	$a_1^{(1)}$	$a_2^{(1)}$	...	$a_{n-1}^{(1)}$	
$\tilde{P}_1$	$a_{n-1}^{(1)}$	$a_{n-2}^{(1)}$	$a_{n-3}^{(1)}$	...	$a_0^{(1)}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_{n-1}	$a_0^{(n-1)}$	$a_1^{(n-1)}$				
$\tilde{P}_{n-1}$	$a_1^{(n-1)}$	$a_0^{(n-1)}$				
P_n	$a_0^{(n)}$					

We have a nice and simple criterion mainly based on the works of Marden [16, 17] and Jury [13, 14], known as Jury-Marden Criterion:

Theorem D.4 (Jury-Marden Criterion)**.** *The polynomial P has property $\mathscr{P}$ if and only if*

$$a_0^{(1)} < 0; \quad a_0^{(k)} > 0, \forall\, 2 \le k \le n.$$

This necessary and sufficient condition is mentioned several times in the literature (see e.g. [1, Theorem 3.10]), but it is not easy to find an explicit proof, so we provide a proof for the reader's convenience. Before proving this result, we apply Jury-Marden Criterion to a polynomial of degree 3 and obtain precisely Lemma C.1, that is the following proposition.

Proposition D.5. *Let $P(z) = a_0 + a_1 z + a_2 z^2 + z^3, z \in \mathbb{C}$ where $a_0, a_1, a_2 \in \mathbb{R}$. Then P has property $\mathscr{P}$ if and only if*

$$\begin{cases} (a_0 - 1)(a_0 + 1) < 0 \\ (a_0^2 - a_2 a_0 + a_1 - 1)(a_0^2 + a_2 a_0 - a_1 - 1) > 0 \\ (a_0 + a_2 - a_1 - 1)(a_0 + a_2 + a_1 + 1) < 0. \end{cases}$$

Proof. By directly applying Jury-Marden Criterion to P, we obtain Marden's table as follows:

	1	x	x^2	x^3
$P_0 = P$	a_0	a_1	a_2	1
$\tilde{P}_0$	1	a_2	a_1	a_0
P_1	$a_0^2 - 1$	$a_1 a_0 - a_2$	$a_2 a_0 - a_1$	
$\tilde{P}_1$	$a_2 a_0 - a_1$	$a_1 a_0 - a_2$	$a_0^2 - 1$	
P_2	$(a_0^2 - 1)^2 - (a_2 a_0 - a_1)^2$	$(a_1 a_0 - a_2)(a_0^2 - a_2 a_0 + a_1 - 1)$		
$\tilde{P}_2$	$(a_1 a_0 - a_2)(a_0^2 - a_2 a_0 + a_1 - 1)$	$(a_0^2 - 1)^2 - (a_2 a_0 - a_1)^2$		

and

$$P_3(x) = \left[(a_0^2 - 1)^2 - (a_2 a_0 - a_1)^2\right]^2 - (a_1 a_0 - a_2)^2 (a_0^2 - a_2 a_0 + a_1 - 1)^2.$$

Hence

$$\begin{aligned} a_0^{(1)} &= a_0^2 - 1 = (a_0 - 1)(a_0 + 1), \\ a_0^{(2)} &= (a_0^2 - 1)^2 - (a_2 a_0 - a_1)^2 = (a_0^2 - a_2 a_0 + a_1 - 1)(a_0^2 + a_2 a_0 - a_1 - 1), \\ a_0^{(3)} &= \left[(a_0^2 - 1)^2 - (a_2 a_0 - a_1)^2\right]^2 - (a_1 a_0 - a_2)^2 (a_0^2 - a_2 a_0 + a_1 - 1)^2 \\ &= \left[(a_0^2 + a_2 a_0 - a_1 - 1)^2 - (a_1 a_0 - a_2)^2\right] (a_0^2 - a_2 a_0 + a_1 - 1)^2 \\ &= [a_0^2 + (a_2 - a_1) a_0 + a_2 - a_1 - 1][a_0^2 + (a_2 + a_1) a_0 - a_2 - a_1 - 1] \\ &\quad (a_0^2 - 1 - a_2 a_0 + a_1)^2 \\ &= (a_0 + 1)(a_0 + a_2 - a_1 - 1)(a_0 - 1)(a_0 + a_2 + a_1 + 1)(a_0^2 - 1 - a_2 a_0 + a_1)^2. \end{aligned}$$

Then the condition $a_0^{(1)} < 0, a_0^{(2)} > 0, a_0^{(3)} > 0$, after being simplified, is equivalent to three inequalities of the statement. □

Now, to prove Jury-Marden Criterion, we need the following two results.

Theorem D.6 (Marden, [17], Theorem 42.1)**.** *Let P be a real coefficient polynomial of n-th degree. If the sequence*

$$m_1(P), m_2(P), ..., m_n(P)$$

has exactly p negative elements and $n-p$ positive elements (hence no null elements), then P has p complex roots (including multiplicities) inside the unit circle $|z| = 1$, no roots on this circle and $n-p$ complex roots (including multiplicities) outside this circle.

Lemma D.7 (Schur, [20])**.** *Let $P(z) = a_0 + a_1 z + ... + a_n z^n$ where $a_k \in \mathbb{R}, \forall 1 \leq k \leq n$. Assume that $|a_0| < |a_n|$. Then $\deg \tilde{P}_1 = n-1$, and P has property $\mathscr{P}$ if and only if $\tilde{P}_1$ has property $\mathscr{P}$.*

Proof of Jury-Marden Criterion D.4. The sufficient condition for P having property $\mathscr{P}$ is a direct consequence of Marden's Theorem D.6. It remains to prove the necessary one.

For that, we will prove the following statement $M(n)$ by induction: "For every real-coefficient polynomial P of n-th degree having property $\mathscr{P}$, the sequence $a_0^{(1)}, ..., a_0^{(n)}$ obtained by Marden's algorithm must satisfy

$$a_0^{(1)} < 0, \quad a_0^{(k)} > 0, \forall\, 2 \leq k \leq n."$$

To check $M(1)$, let $P(z) = a_0 + a_1 z$ where $a_0, a_1 \in \mathbb{R}, a_1 \neq 0$. Then $P(z) = 0 \Leftrightarrow z = -a_0/a_1$ and $|-a_0/a_1| < 1 \Leftrightarrow |a_0| < |a_1| \Leftrightarrow a_0^{(1)} = a_0^2 - a_1^2 < 0$.

Now supposing that $M(n-1)$ is true for some $n \in \mathbb{N}, n \geq 2$, we show that $M(n)$ is true. Let $P(z) = a_0 + a_1 z + ... + a_n z^n$ where $a_k \in \mathbb{R}, k = 0, ..., n$ and $a_n \neq 0$. Assume that P has property $\mathscr{P}$. First, $a_0^{(1)} = a_0^2 - a_n^2 < 0$. Indeed, let $z_1, z_2, ..., z_n$ be the n zeros including multiplicities of P, then by Viète's formulas $z_1 z_2 \cdots z_n = (-1)^n (a_0/a_n)$. Taking the module of both sides of this identity and noting that P has property $\mathscr{P}$, we have $|a_0/a_n| < 1$, thus $a_0^{(1)} = a_0^2 - a_n^2 < 0$. Next, by Lemma D.7, $\tilde{P}_1$ is of $(n-1)$-th degree and it also has property $\mathscr{P}$. Marden's table for $\tilde{P}_1$ can be easily found:

	1	x	x^2	...	x^{n-3}	x^{n-2}	x^{n-1}
$\tilde{P}_1$	$a_{n-1}^{(1)}$	$a_{n-2}^{(1)}$	$a_{n-3}^{(1)}$	...	$a_2^{(1)}$	$a_1^{(1)}$	$a_0^{(1)}$
P_1	$a_0^{(1)}$	$a_1^{(1)}$	$a_2^{(1)}$	...	$a_{n-3}^{(1)}$	$a_{n-2}^{(1)}$	$a_{n-1}^{(1)}$
$-P_2$	$-a_0^{(2)}$	$-a_1^{(2)}$	$-a_2^{(2)}$	...	$-a_{n-3}^{(2)}$	$-a_{n-2}^{(2)}$	
$-\tilde{P}_2$	$-a_{n-2}^{(2)}$	$-a_{n-3}^{(2)}$	$a_{n-4}^{(1)}$	...	$-a_1^{(2)}$	$-a_0^{(2)}$	
P_3	$a_0^{(3)}$	$a_1^{(3)}$	$a_2^{(3)}$	...	$a_{n-3}^{(3)}$		
$\tilde{P}_3$	$a_{n-3}^{(3)}$	$a_{n-4}^{(3)}$	$a_{n-5}^{(3)}$	...	$a_0^{(3)}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_{n-1}	$a_0^{(n-1)}$	$a_1^{(n-1)}$					
$\tilde{P}_{n-1}$	$a_1^{(n-1)}$	$a_0^{(n-1)}$					
P_n	$a_0^{(n)}$						

By $M(n-1)$, we must then have $-a_0^{(2)} < 0,\ a_0^{(k)} > 0, \forall 3 \leq k \leq n$. □

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399