Augmenting the one-shot framework by additional constraints

Torsten Bosse

To cite this article: Torsten Bosse (2016): Augmenting the one-shot framework by additional constraints, Optimization Methods and Software, DOI: 10.1080/10556788.2016.1180692

To link to this article: http://dx.doi.org/10.1080/10556788.2016.1180692

Published online: 12 May 2016.
Augmenting the one-shot framework by additional constraints

Torsten Bosse∗

Argonne National Laboratory, 9700 S. Cass Ave., Argonne, IL, USA

(Received 29 June 2015; accepted 16 April 2016)

The (multistep) one-shot method for design optimization problems has been successfully implemented for various applications. To this end, a slowly convergent primal fixed-point iteration of the state equation is augmented by an adjoint iteration and a corresponding preconditioned design update. In this paper we present a modification of the method that allows for additional equality constraints besides the usual state equation. A retardation analysis and the local convergence of the method in terms of necessary and sufficient conditions are given, which depend on key characteristics of the underlying problem and the quality of the utilized preconditioner.

Keywords: nonlinear optimization; automatic differentiation; piggyback; one-shot method; constraints; eigenvalue analysis

AMS Subject Classification: 49M05; 65F08; 65F15; 90M50; 90C30

1. Introduction

In the past decade, numerous applications and methods for the minimization of design optimization problems were considered. In these problems, one is interested in finding a control $u ∈ U$ that minimizes an objective function $f : U × Y → \mathbb{R}$ for some feasible state variable $y = y(u) ∈ Y$, which is implicitly defined by some state equation $c(u, y) = 0$. Such scenarios typically arise in PDE constraint optimization, where the state equation is some partial differential equation describing a physical process, and can be found in various applications [18–21,24,26,28,29,31]. For example, one can think of the shape optimization of an airfoil in order to minimize the drag, which is represented by the objective function $f$. Thus, the (parametrized) shape of the airfoil is given by the control $u$, and the feasible state $y$ describes its surrounding airflow that satisfies some version of the Navier–Stokes equation represented by the function $c$. The structure and the large number of unknowns make (the discretization of) these problems usually intractable for most standard nonlinear programming methods. In many of these examples, the state equation is given by an equivalent contractive fixed-point function $G : U × Y → Y$ such that the state variable $y$ satisfies the state equation $c(u, y) = 0$ if and only if it is a solution of the fixed-point equation $y = G(u, y)$ for any given control $u ∈ U$. The fixed-point function $G$ can be thought of as one iteration of a numerical procedure for the solution of the underlying (discretized) state equation, for example, a highly specialized simulation code for this particular physical application. It can be used to design so-called one-shot methods [1–3,8,11–13,16,
18.22], for the (local) solution of the resulting design optimization problem

$$\min_{(u,y)} f(u,y), \quad \text{s.t. } y = G(u,y)$$

(1)

if the functions $f$ and $G$ are sufficiently smooth and the fixed-point function satisfies the contraction condition

$$\|G_y(u,y)\| \leq \rho_G < 1, \quad (u,y) \in U \times Y,$$

(2)
in some appropriate operator norm on the vector space $Y$ for any fixed control $u \in U$. These methods are well suited for problems with a slow contraction rate of the fixed-point iteration, namely, for problems where the upper bound $\rho_G$ on the norm of the partial derivative of $G$ w.r.t. $y$ is close to one. They are based upon the Karush–Kuhn–Tucker (KKT) conditions for the first-order stationary points $(u^*, y^*)$ of the design optimization problem (1). In the finite-dimensional case, $Y = \mathbb{R}^n$ and $U = \mathbb{R}^m$, the KKT conditions ensure the existence of a unique adjoint variable $\tilde{y}^*$ in the corresponding dual space $\bar{Y} = \mathbb{R}^m$ such that

$$0 = L_u(u^*, y^*, \tilde{y}^*) = f_u(u^*, y^*) + G_u(u^*, y^*)^\top \tilde{y}^*$$
$$0 = L_y(u^*, y^*, \tilde{y}^*) = f_y(u^*, y^*) + (G_y(u^*, y^*) - I)^\top \tilde{y}^*$$
$$0 = L_{\tilde{y}}(u^*, y^*, \tilde{y}^*) = G(u^*, y^*) - y^*$$

holds under the stated assumptions. Here, $L : U \times Y \times \bar{Y} \to \mathbb{R}$ denotes the Lagrangian

$$L(u,y,\bar{y}) = f(u,y) + \bar{y}^\top (G(u,y) - y)$$

associated with (1). In combination with the contraction condition (2), the necessary optimality conditions yield an adjoint fixed-point iteration (see also [5,6,13])

$$\tilde{y}^+ = f_y(u,y) + G_y(u,y)^\top \bar{y}$$

for the adjoint $\tilde{y}$ that can be used in combination with a preconditioned design update

$$u^+ = u + \alpha B^{-1}(f_u(u,y) + G_u(u,y)^\top \bar{y})$$

for some suitable step-multiplier $\alpha \in \mathbb{R}^+$ and preconditioner matrix $B \in \mathbb{R}^{m \times m}$ to find such KKT points. The original fixed-point iteration, the adjoint fixed-point iteration, and the preconditioned design update motivate a number of different one-shot schemes to compute a sequence of iterates $(u_k, y_k, \tilde{y}_k)$ that converge to stationary points $(u_\star, y_\star, \tilde{y}_\star) = \lim_{k \to \infty} (u_k, y_k, \tilde{y}_k)$ for some initial guess $(u_0, y_0, \tilde{y}_0)$ sufficiently close to a solution. Three of these one-shot schemes can be briefly described by the propagation rules

$$\cdot \cdot \cdot \to (\text{design } u, \text{ state } y, \text{ adjoint } \tilde{y}_\cdot) \to \cdot \cdot \cdot,$$

(3)

$$\cdot \cdot \cdot \to \text{design } u \to \text{ state } y \to \text{ adjoint } \tilde{y} \to \cdot \cdot \cdot,$$

(4)

$$\cdot \cdot \cdot \to \text{design } u \to (\text{state } y)^\cdot \to (\text{adjoint } \tilde{y})^\cdot \to \cdot \cdot \cdot.$$

(5)

Here, the terminology $(t_1, t_2)$ denotes the parallel execution of the two updates $t_1$ and $t_2$, whereas $t_1 \to t_2$ indicates that one update $t_2$ is executed after one update $t_1$ is completed using the latest available information; consequently, $t_1^s$ abbreviates the $s$-times repetition $t_1 \to \cdots \to t_1$ of one
update $t_1$. In detail, the Jacobi–one-shot method (3), the Seidel–one-shot method (4), and the multistep Seidel–one-shot method (5) refer to the respective updates

\[
\begin{bmatrix}
    u^{k+1} \\
    y^{k+1} \\
    \tilde{y}^{k+1}
\end{bmatrix} = \begin{bmatrix}
    u^k - \alpha_k B_k^{-1}(f_u(u^k, y^k) + G_u(u^k, y^k)\tilde{y}^k) \\
    f_y(u^k, y^k) + G_y(u^k, y^k)\tilde{y}^k \\
    f_y(u^k, y^k) + G_y(u^k, y^k)\tilde{y}^k
\end{bmatrix},
\]

The derivatives in the definition of the updates can be computed by applying techniques from algorithmic differentiation [14,25]. For example, the adjoint product $G_y(u, y)^\top \tilde{y}$ can be efficiently evaluated by software packages such as ADOL-C [33] or Tapenade [17].

In particular, the multistep Seidel approach (5) was investigated in [3], where the method was shown to be locally convergent for an appropriate choice of preconditioner matrices $B_k$ and a sufficiently large number $s_k \in \mathbb{N}$ of multiple state updates

\[
G^{s_k}(u^k, y^k) = \frac{G(u^k, G(u^k, \ldots, G(u^k, y^k) \ldots))}{s_k-	ext{times}}
\]

and corresponding adjoint updates. The choice for the preconditioner $B_k$ and the number $s_k$ in every iteration $k$ was related to problem-dependent quantities that could be estimated during the runtime of the procedure by using the information from previous iterates $(u_l, y_l, \tilde{y}_l)$ for $l = 0, 1, \ldots, k - 1$. Moreover, the proposed stepping scheme was shown to have a retardation factor of 2 in the ideal case. Here, the retardation factor is the efficiency measure of an optimization method that is defined by the ratio

\[
\frac{\text{Cost(Optimization)}}{\text{Cost(Simulation)}} \sim O(|\text{prob, mesh, load, \ldots}|^b),
\]

representing the slowdown of going from a full simulation to compute a feasible state to a full optimization of the design optimization problem.

In this paper, we extend the previous results for the original design optimization problem (1) and consider the modified design optimization problem

\[
\min_{(u, y_1, y_2)} f(u, y_1, y_2), \quad \text{s.t. } y_2 = G(u, y_1, y_2) \quad \text{and} \quad g(u, y_1, y_2) = 0. \tag{6}
\]

It has a similar structure to problem (1), except that now an additional equality constraint $g : U \times Y_1 \times Y_2 \to Y_1$ is present. For consistency, we adapt the previous notation and denote by $u \in U$ the control variables and by $(y_1, y_2) \in Y_1 \times Y_2$ the state variables, where the finite-dimensional spaces $Y_1$, $Y_2$, and $U$ are now given by $U = \mathbb{R}^m$, $Y_1 = \mathbb{R}^{n_1}$, and $Y_2 = \mathbb{R}^{n_2}$ with $m, n = n_1 + n_2 \in \mathbb{N}$. Analogous to before, $G : U \times Y_1 \times Y_2 \to Y_2$ represents a contractive fixed-point mapping for fixed variables $(u, y_1)$ satisfying the contraction condition

\[
||G_{y_2}(u, y_1, y_2)|| \leq \rho_G < 1 \quad \text{for } (u, y_1, y_2) \in U \times Y_1 \times Y_2, \tag{7}
\]

in some appropriate operator norm but now for the vector space $Y_2$. As already implicitly done in definition (6), we assume that the state variable $y$ can be separated into two state variables parts.
$y_1 \in Y_1$ and $y_2 \in Y_2$ such that for any choice of $y_2$ and a control $u \in U$ there exists $y_1$, which solves the additional constraint $g(u, y_1, y_2) = 0$. Mathematically, we require that the Jacobian $g_{y_1}(u, y_1, y_2)$ not be singular for all $(u, y_1, y_2) \in U \times Y_1 \times Y_2$ in a sufficiently large neighbourhood of the solution, such that there exists an implicit function $\phi : U \times Y_2 \to Y_1$ that satisfies $g(u, \phi(u, y_2), y_2) = 0$.

As before, the fixed-point iteration $G$ can be interpreted as a simulation code for the computation of the flow $y_2$, and the control vector $u$ represents all parameters for the shape of the airfoil. The additional constraint given by the function $g : \mathbb{R}^{n_1+n_2} \to \mathbb{R}^n$ describes, for example, the requirement for constant lift ($n_1 = 1$) and the second state variable $y_1 \in \mathbb{R}^1$ the angle of attack of the airfoil, which can be adjusted to solve the additional constraint.

Until now, it was not clear how the design optimization problem with additional constraints could be solved by the one-shot approach. One possibility was to use a penalty approach, where a penalty term is added to the objective, to incorporate the violation of the additional constraint, and apply one of the previously described methods on the modified problem. For example, Walther et al. [32] just recently proposed an extension for the Jacobi-one-shot method [15], which uses a preconditioner that is based on a doubly augmented penalty function. However, this method requires some heuristic for the adaptation of the penalty parameters and, in case of a bad choice, might lead to a slow convergence of the overall method. In detail, a retardation analysis for this approach still has to be investigated even if the stated numerical results are promising.

Therefore, we pursue the more intuitive approach and develop an extended one-shot method that directly incorporates updates for the additional constraint into the stepping scheme. The proposed method extends the multi-step Seidel–one-shot method (5) for the original problem (1) and is based on the first-order optimality conditions for the extended problem (6), which will be given at the beginning of Section 2. In detail, we replace the previous state and adjoint update in (5) by the extended state and adjoint updates
\[
(y_2 \to y_1) \quad \text{and} \quad (y_1, y_2) \to (\bar{y}_1, \bar{y}_2)
\]
that now include the quantities $y_1$ and $\bar{y}_1$, respectively. Both updates are motivated on a small illustrative counterexample and will be defined in Section 2 (see Equations (9) and (10)). The update for the state $(y_1, y_2)$ and adjoint $(\bar{y}_1, \bar{y}_2)$ can be thought of an extended mapping $G$ with its corresponding adjoint operation depending on some preconditioner $C$. In Section 3, we show that the preconditioner can be chosen such that $G$ satisfies the contraction condition (2) and, thus, allows use of the previous convergence results for the original multistep Seidel–oneshot method (5) for the extended fixed-point mapping $G$. A requirement for the existence of such a preconditioner is that the original fixed-point iteration $G$ has a sufficiently small contraction rate $\rho_G$ as can be seen by an eigenvalue analysis. The latter can be achieved by considering a sequence of $s_G$ updates for the state $y_2$ before the update of $y_1$ and the corresponding adjoint update. This inspiring the nested multistep one-shot scheme proposed in Section 4:
\[
\ldots \to \text{DESIGN} \to ((\text{STATE} \ y_2)^{s_G} \to \text{STATE} \ y_1)^s \to (\text{ADJOINT} \ y_1, y_2)^s \to \ldots
\]
A sufficient lower bound on the number $s_G$ for the number of fixed-point iterations $G$ will be given, in order to guarantee local convergence of the overall method, relying on the results given in [3] for the choice of $s$. Both lower bounds on $s_G$ and $s$ depend on problem-specific quantities and the quality of the corresponding preconditioners $C$ and $B$, respectively. For a suitable choice of both quantities, it can be shown that the proposed oneshot-method has a retardation factor of 4. A numerical validation for some parts of the theoretical results is illustrated with a simple example given in Section 5. The conclusion are given in Section 6 with a brief summary and suggestions for future work.
2. Fixed-point iteration for the augmented problem

According to standard nonlinear optimization theory [27] and the stated assumption, there exists a unique pair of adjoint variables \( \bar{y}_1^* \) and \( \bar{y}_2^* \) in the corresponding dual spaces \( \bar{Y}_1 = \mathbb{R}^{n_1} \) and \( \bar{Y}_2 = \mathbb{R}^{n_2} \), respectively, such that the KKT conditions

\[
0 = L_u(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = f_u(u, y_1, y_2) + \mathcal{G}_u(u, y_1, y_2)^\top \bar{y}_2 + g_u(u, y_1, y_2)^\top \bar{y}_1 \\
0 = L_{y_1}(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = f_{y_1}(u, y_1, y_2) + \mathcal{G}_{y_1}(u, y_1, y_2)^\top \bar{y}_2 + g_{y_1}(u, y_1, y_2)^\top \bar{y}_1 \\
0 = L_{y_2}(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = f_{y_2}(u, y_1, y_2) + (\mathcal{G}_{y_2}(u, y_1, y_2) - I)^\top \bar{y}_2 + g_{y_2}(u, y_1, y_2)^\top \bar{y}_1 \\
0 = L_{\bar{y}_1}(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = g(u, y_1, y_2) \\
0 = L_{\bar{y}_2}(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = \mathcal{G}(u, y_1, y_2) - y_2
\]

are satisfied for any first-order stationary point \((u^*, y_1^*, y_2^*)\) of the extended problem (6), where linear independence constraint qualifications hold. Here, \( L : U \times Y_1 \times Y_2 \times \bar{Y}_1 \times \bar{Y}_2 \to \mathbb{R} \) denotes the Lagrangian function

\[
L(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = f(u, y_1, y_2) + (\mathcal{G}(u, y_1, y_2) - y_2)^\top \bar{y}_2 + g(u, y_1, y_2)^\top \bar{y}_1.
\]

In this paper, we provide a modification of the multistep Seidel–one-shot method (5) to find a stepping scheme that computes such stationary points \((u^*, y_1^*, y_2^*, \bar{y}_1^*, \bar{y}_2^*)\) of the problem (6). Although several stepping schemes are possible,

\[
\cdots \to (\text{DESIGN } u, \text{STATE } y_2, \text{STATE } y_1, \text{ADJOINT } \bar{y}_2, \text{ADJOINT } \bar{y}_1, \ldots)
\]

\[
\cdots \to \text{DESIGN } u \to (\text{STATE } y_2, \text{STATE } y_1)^s \to (\text{ADJOINT } \bar{y}_2, \text{ADJOINT } \bar{y}_1)^s \to \ldots,
\]

\[
\cdots \to \text{DESIGN } u \to (\text{STATE } y_2)^s \to (\text{STATE } y_1)^s \to (\text{ADJOINT } \bar{y}_2)^s \to \ldots, \ldots
\]

which correspond to the original Jacobian method (first scheme), a mixed Seidel–Jacobian approach (second scheme), and the pure Seidel approach (third scheme), respectively, we focus first on the specific stepping scheme

\[
\cdots \to \text{DESIGN } u \to (\text{STATE } y_2 \to \text{STATE } y_1)^s \to (\text{ADJOINT } \bar{y}_2 \to \text{ADJOINT } \bar{y}_1)^s \to \ldots
\]

(8)

that extends the previously presented multistep Seidel–one-shot approach (5) in a natural manner.

Example 1 (Motivating Counterexample) If we assume for the moment that \( s = 1 \), then we can formulate the state update for the primal variable \( y_2 \) at a given current iterate \((u, y_1, y_2)\) by the fixed-point iteration step

\[
y_2^+ = \mathcal{G}(u, y_1, y_2)
\]

motivated by the stationarity condition \( 0 = L_{\bar{y}_2}(u, y_1, y_2, \bar{y}_1, \bar{y}_2) \) and the assumption \( \|\mathcal{G}_{y_2}(u, y_1, y_2)\| \leq \rho_G < 1 \). Also, we can at least theoretically define the new state \( y_2^+ \) as the root of \( g(u, \ldots, y_2^+) = 0 \) such that \( 0 = L_{\bar{y}_2}(u, y_1^+, y_2^+, \bar{y}_1, \bar{y}_2) \) holds after one primal state cycle (state \( y_2 \to \text{STATE } y_1 \)). Analogously, we can compute the adjoint update \( \bar{y}_2 \) by the fixed-point iteration

\[
\bar{y}_2^+ = L_{\bar{y}_2}(u, y_1^+, y_2^+, \bar{y}_1, \bar{y}_2)^\top + \bar{y}_2
\]

according to the third stationarity condition and set the adjoint \( \bar{y}_1^+ \) to be the unique solution of \( 0 = L_{y_1}(u, y_1^+, y_2^+, \bar{y}_1^+, \bar{y}_2^+) \) since \( \|\mathcal{G}_{y_1}(u, y_1, y_2)\| \leq \rho_G < 1 \) and \( g_{y_1}(u, y_1, y_2) \) was assumed to be
invertible. Hence, we see that after one evaluation of the update sequence

$$\cdots \rightarrow (\text{STATE } y_2 \rightarrow \text{STATE } y_1)^{\dagger} \rightarrow (\text{ADJOINT } \tilde{y}_2 \rightarrow \text{ADJOINT } \tilde{y}_1)^{\dagger} \rightarrow \ldots,$$

at least the two stationarity conditions

$$0 = L_{y_1}(u, y_1, y_2, \tilde{y}_1, \tilde{y}_2) \quad \text{and} \quad 0 = L_{\tilde{y}_1}(u, y_1, y_2, \tilde{y}_1, \tilde{y}_2)$$

are exactly satisfied. However, this situation does not need to hold true for the other two stationarity conditions $0 = L_{y_2}(u, y_1, y_2, \tilde{y}_1, \tilde{y}_2)$ and $0 = L_{\tilde{y}_2}(u, y_1, y_2, \tilde{y}_1, \tilde{y}_2)$ since they are in general affected by the subsequent changes in the variables $y_1$ and $\tilde{y}_1$, respectively. In fact, this may lead to divergence of the state and adjoint cycles

$$(\text{STATE } y_2 \rightarrow \text{STATE } y_1)^{\ddagger} \quad \text{and} \quad (\text{ADJOINT } \tilde{y}_2 \rightarrow \text{ADJOINT } \tilde{y}_1)^{\ddagger}$$

even in the case when the updates are exact, as indicated in Figure 1 (left).

The basic idea is now to reduce the influence of the changes in the variables $y_1$ and $\tilde{y}_1$ by rescaling the corresponding updates $y_2$ and $\tilde{y}_2$ as discussed in the previous example and depicted in Figure 1 (middle, right) with some corresponding step multiplier and preconditioner. Therefore, we consider the extended mapping $G : U \times Y_1 \times Y_2 \rightarrow Y_1 \times Y_2$ given by

$$(y_1^+, y_2^+) = G(u, y_1, y_2) = (y_1 - \alpha_G C^{-1} g(u, y_1, G(u, y_1, y_2))), \ G(u, y_1, y_2), \quad (9)$$

which represents an update of the state variable $y_2$ that is used for a scaled update of $y_1$, as indicated in (8). To guarantee that the extended state update $G$ is contractive, we need to find a suitable preconditioner matrix $C \in \mathbb{R}^{n_1 \times n_1}$ and step multiplier $\alpha_G \in \mathbb{R}_+$ such that $G$ satisfies the contraction assumption (2) for $y = (y_1, y_2)$ and control $u$. The corresponding ADJOINT fixed-point iteration $G : U \times \bar{Y}_1 \times \bar{Y}_2 \rightarrow \bar{Y}_1 \times \bar{Y}_2$ can be derived by differentiating the Lagrangian

$$L(u, y_1, y_2, \bar{y}_1, \bar{y}_2) = f(u, y_1, y_2) + (G(u, y_1, y_2) - y_2)^\top (\bar{y}_1; \bar{y}_2)$$

$$= f(u, y_1, y_2) + (-\alpha_G C^{-1} g(u, y_1, G(u, y_1, y_2))), \ G(u, y_1, y_2) - y_2)^\top (\bar{y}_1; \bar{y}_2)$$

of the design optimization problem (1) for $G$ defined in (9) with respect to $(y_1, y_2)$ and incrementing

$$(\bar{y}_1^+, \bar{y}_2^+) = (\bar{y}_1 + L_{y_1}(u, y_1^+, y_2^+, \bar{y}_1, \bar{y}_2)^\top, \bar{y}_2 + L_{y_2}(u, y_1^+, y_2^+, \bar{y}_1, \bar{y}_2)^\top). \quad (10)$$
Thus, we do not have the stepping scheme (8) as proposed in the first place but
\[ \cdots \rightarrow \text{DESIGN } u \rightarrow (\text{STATE } y_2 \rightarrow \text{STATE } y_1)^s \rightarrow (\text{ADJOINT } (\tilde{y}_1, \tilde{y}_2))^s \rightarrow \cdots. \] (11)

Obviously, the primal preconditioning could and should depend on the current iterate, namely, \( \alpha_G = \alpha_G(u, y_1, y_2) \) and \( C = C(u, y_1, y_2) \), to prevent a too-conservative update strategy and, thus, slow convergence of the overall method to stationary points \((u^*, y_1^*, y_2^*, \tilde{y}_1^*, \tilde{y}_2^*)\) of the problem (6). For simplicity, we restrict ourselves on a local analysis to the extended multi-step–oneshot method defined by the stepping sequence of the updates (5) for \( u, y = (y_1, y_2) \) and \( \tilde{y} = (\tilde{y}_1, \tilde{y}_2) \) close to \((u^*, y_1^*, y_2^*, \tilde{y}_1^*, \tilde{y}_2^*)\), where the primal fixed-point iteration \( G \) and adjoint mapping \( \hat{G} \) are defined by (9) and (10), respectively. Based on an eigenvalue analysis, we will provide a suitable choice for the preconditioner matrix \( C \) and the stepsize \( \alpha_G \) to ensure the contractivity of the extended mapping \( G \). For such stepsizes \( \alpha_G \) and preconditioner matrices \( C \), we can then apply the results from [3] for the multistep Seidel–oneshot method \((s \geq 1)\) on \( G \) using the previously defined adjoint update. Therefore, we will assume for the moment that the contraction rate \( \rho_G \) is sufficiently small. One choice for the matrix \( C \) in the extended fixed-point iteration (8) is the projected Newton preconditioner
\[ C = g_{y_1} + g_{y_2} (I - \mathcal{G}_{y_2})^{-1} \mathcal{G}_{y_1} \] (12)
or a low-rank approximation of it [4,7,9,30]. The resulting algorithm and its (local) convergence behaviour heavily depend on the quality of the preconditioner, besides other problem-dependent quantities as we will see in the next section.

3. Eigenvalue analysis for the extended mapping

The eigenvalue analysis is based on an argumentation line similar to the one used in [3]. In the first step of the analysis, we show that all eigenvalues \( \lambda \in \mathbb{C} \) of the Jacobian matrix
\[ \mathcal{G}_{(y_1, y_2)} = \frac{\partial G}{\partial (y_1, y_2)}(u^*, y_1^*, y_2^*) \]
of the extended mapping \( G \) either are in the spectrum of the Jacobian matrix \( \mathcal{G}_{y_2} \) of the fixed-point iteration \( \mathcal{G} \) or are roots of a complex polynomial \( P(\cdot) : \mathbb{C} \rightarrow \mathbb{C} \). For all eigenvalues that are not in the spectrum of \( \mathcal{G}_{y_2} \), we can derive a necessary condition to be a root of this polynomial in terms of an inequality that includes problem specific-parameters. As we shall see in Section 4, some of these parameters can be adjusted such that the inequality is satisfied only for eigenvalues \( \lambda \) with \(|\lambda| < 1\). This then implies contractivity of the extended fixed-point mapping \( G \) at \((u^*, y_1^*, y_2^*)\) and, thus, also at points in a vicinity of the solution by a continuity argument. Therefore, let us consider the Jacobian matrix \( \mathcal{G}_{(y_1, y_2)}^* \) of the extended mapping \( G \)
\[ \mathcal{G}_{(y_1, y_2)}^* = \begin{bmatrix} I - \alpha_G C^{-1} g_{y_1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ \mathcal{G}_{y_1} & \mathcal{G}_{y_2} \end{bmatrix} = \begin{bmatrix} I - \alpha_G C^{-1}(g_{y_1} + g_{y_2} \mathcal{G}_{y_1}) & 0 \\ 0 & \mathcal{G}_{y_2} \end{bmatrix}, \]
where all occurring derivatives are evaluated at a solution \((u^*, y_1^*, y_2^*)\) of problem (6). For the eigenvalues of this matrix, we can show the following.
PROPOSITION 1 Any complex eigenvalue $\lambda \in \mathbb{C}$ of the Jacobian $G_{(y_1,y_2)}^*$ satisfies

$$\lambda \in \text{spec}(G_{y_2}) \quad \text{or} \quad \det(M(\lambda)) = 0,$$

where $M(\lambda) = (1 - \lambda)I - \alpha_G C^{-1}(g_{y_1} + g_{y_2} G_{y_1}) - \alpha_G C^{-1} g_{y_2} G_{y_2} (\lambda I - G_{y_2})^{-1} G_{y_1}$.

Proof The spectrum of the matrix $G_{(y_1,y_2)}^*$ is given by the complex roots of the polynomial

$$P(\lambda) = \det(G_{(y_1,y_2)}^* - \lambda I)$$

$$= \det \begin{bmatrix} (1 - \lambda)I - \alpha_G (C^{-1} g_{y_1} + C^{-1} g_{y_2} G_{y_1}) & -\alpha_G C^{-1} g_{y_2} G_{y_2} \\ G_{y_1} & \tilde{G}_{y_1} - \lambda I \end{bmatrix}.$$

According to the Laplacian expansion theorem (see [10]), we conclude that

$$P(\lambda) = \det(G_{y_2} - \lambda I) \det(M(\lambda)),$$

where the matrix $M(\lambda) \in \mathbb{C}^{n_1 \times n_1}$ is defined by the Schur complement

$$M(\lambda) = (1 - \lambda)I - \alpha_G C^{-1}(g_{y_1} + g_{y_2} G_{y_1}) - \alpha_G C^{-1} g_{y_2} G_{y_2} (\lambda I - G_{y_2})^{-1} G_{y_1}.$$

Thus, any eigenvalue is either in the spectrum of $G_{y_2}$ or a root of $\det(M(\lambda)) = 0$. ◼

Since any eigenvalue of $G_{y_2}$ is already strictly smaller than, from the assumption of $G$ being a contractive fixed-point iteration, it is sufficient to guarantee that the condition

$$\det(M(\lambda)) = 0$$

is satisfied only for $\lambda \in \mathbb{C}$ with $|\lambda| < 1$ in order to prove the contractivity of the extended fixed-point iteration $G$. At least theoretically, we can assume that the variables were transformed by $y_1 = T^{-1} \tilde{y}_1$ such that the matrix $\tilde{H}_G(1)$ with

$$\tilde{H}_G(\lambda) \equiv \tilde{g}_{y_1} + \tilde{g}_{y_2} \tilde{G}_{y_1} + \tilde{g}_{y_2} \tilde{G}_{y_2} (\lambda I - \tilde{G}_{y_2})^{-1} \tilde{G}_{y_1}$$

is the unit for the transformed functions and variables that are annotated by a tilde; in other words, $\tilde{H}_G(1) = I$ for $\tilde{G}(u, \tilde{y}_1, y_2) = G(u, T\tilde{y}_1, y_2)$ and so on. For example, we can use the transformation

$$T^{-1} = H_G(1) \equiv g_{y_1} + g_{y_2} G_{y_1} + g_{y_2} G_{y_2} (I - G_{y_2})^{-1} G_{y_1} = g_{y_1} + g_{y_2} (I - G_{y_2})^{-1} G_{y_1},$$

if $H_G(1)^{-1}$ exists. Note that the second variable $y_2 = \tilde{y}_2$ is not affected by this transformation. As a result, we find the following necessary condition using the rational expression,

$$\mu(\eta, |\lambda|) \equiv \eta \left( \frac{|\lambda| + 1}{|\lambda| - \eta} \right),$$

and the transformed quantities such as the corresponding preconditioner matrix $\tilde{C}$, which should be equal to the unit in the ideal case.

PROPOSITION 2 All eigenvalues $\lambda \in \mathbb{C}$ of the Jacobian of the extended fixed-point iteration with the preconditioner matrix $\tilde{C}$ satisfy

$$|\lambda| \leq \rho_G \quad \text{or} \quad |\lambda| \leq \gamma_{\tilde{G}} + v_{\tilde{G}} c_{\tilde{G}} d_{\tilde{G}} \mu(\rho_G, |\lambda|),$$

where the constants are given by

$$\gamma_{\tilde{G}} = \|I - \alpha_G \tilde{C}^{-1}\|, \quad v_{\tilde{G}} = \alpha_G \|\tilde{C}^{-1}\|, \quad c_{\tilde{G}} = \|g_{y_2}\|, \quad \text{and} \quad d_{\tilde{G}} = \|(I - G_{y_2})^{-1} G_{y_1}\|\|T\|.$$
Proof. For the eigenvalues \( \lambda \in \mathbb{C} \) with \( |\lambda| \leq \rho_g \) there is nothing to show. Thus, we need to consider only eigenvalues with \( |\lambda| > \rho_g \). According to Proposition 2, it follows that for these values \( \det(M(\lambda)) = 0 \), which implies that there exists a kernel vector \( v \in \mathbb{C}^{n_1} \) of unit length such that

\[
\lambda v = [(I - \alpha \tilde{C}^{-1}) \tilde{H}_G(1)) - \alpha \tilde{C}^{-1} (\tilde{H}_G(\lambda) - \tilde{H}_G(1))] v
\]

and, therefore,

\[
|\lambda| \leq \|I - \alpha \tilde{C}^{-1}\| + \alpha \|\tilde{C}^{-1}\| \|\tilde{H}_G(\lambda) - \tilde{H}_G(1)\|.
\]

Here, the difference \( \tilde{H}_G(\lambda) - \tilde{H}_G(1) \) is given by

\[
\tilde{H}_G(\lambda) - \tilde{H}_G(1) = \tilde{g}_{\gamma_2} \tilde{g}_{\gamma_2} (\lambda I - \tilde{G}_{\gamma_2})^{-1} \tilde{g}_{\gamma_1} - \tilde{g}_{\gamma_2} \tilde{g}_{\gamma_2} (I - \tilde{G}_{\gamma_2})^{-1} \tilde{g}_{\gamma_1}
= [1 - \lambda] \tilde{g}_{\gamma_2} \tilde{g}_{\gamma_2} (\lambda I - \tilde{G}_{\gamma_2})^{-1} (I - \tilde{G}_{\gamma_2})^{-1} \tilde{g}_{\gamma_1}.
\]

Its norm can be bounded from above by using the submultiplicativity of the operator norm and the assumption \( 0 \leq \|\tilde{g}_{\gamma_2}\| \leq \rho_g \)

\[
\|\tilde{H}_G(\lambda) - \tilde{H}_G(1)\| \leq |1 - \lambda| c_{\tilde{g}} \frac{\rho_g}{|\lambda| - \rho_g} d_{\tilde{g}} \leq c_{\tilde{g}} d_{\tilde{g}} \rho_g |\lambda| + 1 |\lambda| - \rho_g = c_{\tilde{g}} d_{\tilde{g}} \mu(\rho_g, |\lambda|),
\]

where \( c_{\tilde{g}} = \|g_{\gamma_2}\| \) and \( d_{\tilde{g}} = \|(I - G_{\gamma_2})^{-1} G_{\gamma_2}\|\|T\|\). Thus, the asserted inequality \( |\lambda| \leq y_{\tilde{g}} + v_{\tilde{g}} c_{\tilde{g}} d_{\tilde{g}} \mu(\rho_g, |\lambda|) \) follows by defining \( y_{\tilde{g}} = \|I - \alpha \tilde{C}^{-1}\| \) and \( v_{\tilde{g}} = \alpha \|\tilde{C}^{-1}\| \).

As an immediate consequence for the Newton scenario, we find the following result.

Corollary 1. Assume that the Newton preconditioner (12) is invertible. Then the extended fixed-point mapping \( \tilde{G} \) is contractive for a suitable stepsize \( \alpha_G \) if \( \rho_G, c_{\tilde{g}}, \) and \( d_{\tilde{g}} \) are sufficiently small.

Proof. The proof is a direct consequence of Proposition 2 since the intersection points, where (13) (right) holds as equality, are given by the roots of a quadratic equation that is obtained by multiplication with \( |\lambda| - \rho \). Its solution can be arbitrarily close to zero for a sufficient choice of \( \alpha_G, \rho_G, c_{\tilde{g}}, \) and \( d_{\tilde{g}} \) using the given Newton preconditioner \( C \) with \( v_{\tilde{g}} = 1 \).

In other words, the extended fixed-point iteration \( \tilde{G} \) is contractive if the primal updates for \( y_2 \) and \( y_1 \) are Newton steps and there is only a slight coupling of the variables by the constraints. This situation can be seen by noting that (12) coincides with the total derivative \( dg(u, y_1, y_2)/dy_1 \). On the other hand, the situation depicted in Figure 1 (left) is reflected by the proposition; that is, even full Newton steps \( (\alpha_G = 1) \) for \( y_1 \) and arbitrary small contraction rates \( \rho_G \neq 0 \) for \( y_2 \) can lead to divergence. In this case the right inequality (13) implies only

\[
|\lambda| \leq 0 + v_{\tilde{g}} c_{\tilde{g}} d_{\tilde{g}} \mu(\rho_g, |\lambda|) = \|\tilde{C}^{-1}\| c_{\tilde{g}} d_{\tilde{g}} \rho_g \frac{|\lambda| + 1}{|\lambda| - \rho_g},
\]

which can be satisfied for any \( |\lambda| \in \mathbb{R}_+ \) for a sufficiently large choice of \( \|\tilde{C}^{-1}\| c_{\tilde{g}} d_{\tilde{g}} \). This situation might happen if small changes in \( y_2 \) have a large impact on the feasibility of the stationary condition \( g(u, y_1, y_2) = 0 \). The latter fact is represented by the quantity \( c_{\tilde{g}} = \|g_{\gamma_2}\| \) arising in formula (13), which measures the partial derivative \( \delta g/\delta y_2 \). The other quantity \( d_{\tilde{g}} \) can be understood as the influence of \( y_1 \) on \( y_2 \) since the projection matrix \( (I - G_{\gamma_2})^{-1} G_{\gamma_1} \) denotes the partial derivative \( \partial y_2/\partial y_1 \) according to the implicit function theorem. If one of the quantities is zero,
for example, the solution $y_1$ of $g(u, y_1, y_2) = 0$ is independent of the choice of $y_2$, then (13) simplifies to

$$|\lambda| \leq \|I - \alpha_G \tilde{C}^{-1}\|,$$

which suggests that any preconditioner $C$, or its transformed version $\tilde{C}$, and step-multiplier $\alpha_G$ with $\|I - \alpha_G \tilde{C}^{-1}\| \leq \rho_G$ preserve the contraction rate for the coupled iteration. In particular, we have no retardation at all for this choice, and using the exact preconditioner does not make sense since doing so would mean oversolving for $y_1$. If the exact Newton preconditioner is available, any stepsize $\alpha_G \in [1 - \rho_G, 1 + \rho_G]$ preserves $\rho_G = \rho_G$, as visualized in Figure 2 (left).

4. Enforcing contraction for the general case

In the preceding section, we showed that there exist stepsizes $\alpha_G$ and (projected Newton) preconditioners $C$ that guarantee that the extended mapping $G$ satisfies the contraction condition (2) and, thus, allow convergence of the overall method. A necessary condition for their existence was that there is only a slight coupling of the variables by the constraints, namely, $c_{\tilde{G}}$ or $d_{\tilde{G}}$ are sufficiently small. In this section, we discuss how their existence can be enforced in the case of a strong coupling, which will be achieved by choosing the primal contraction rate $\rho_G$ sufficiently small to compensate too large values $c_{\tilde{G}}d_{\tilde{G}} \gg 0$. The latter can be achieved by considering multiple updates $G^{s_G}$ instead of one update $G$ itself; that is, instead of just performing one fixed-point iteration for $y_2$, a sequence of $s_G$ updates is performed before updating $y_1$ in (9). This motivates the nested multistep one-shot method presented at the end of the introduction. For this method, we give a lower bound on the number of updates $s_G$ to ensure that the primal contraction rate $\rho_{G}^{s_G}$ of the multiple updates $G^{s_G}$ is sufficiently small. It is based on the following observations.

As mentioned earlier, we cannot prevent a contraction rate $\rho_G$ of the extended fixed-point iteration $G$ larger than one for the Newton update with (12)

$$y_1^+ = y_1 - \alpha_G C^{-1} g(u, y_1, G(u, y_1, y_2))$$

by choosing $\alpha_G$ sufficiently small or large. To see this, we depicted in Figure 2 (right) the right-hand side

$$\psi_{\alpha_G}(|\lambda|) = \|1 - \alpha_G \tilde{C}^{-1}\| + \alpha_G \|\tilde{C}^{-1}\|c_{\tilde{G}}d_{\tilde{G}} \mu(\rho_G, |\lambda|)$$

Figure 2. Feasible stepsizes for the exact Newton preconditioner in the decoupled case $c_{\tilde{G}}d_{\tilde{G}} = 0$ (left) and the situation for the general coupled case with $c_{\tilde{G}}d_{\tilde{G}} \gg 0$ (right).
of the second inequality (13) for two choices \(0 < \alpha_G^1 < \alpha_G^2 < 1\) and variable \(|\lambda|\). The lower bound of the function \(\psi_{\alpha_G}(|\lambda|)\) for \(|\lambda| > \rho_G\) is given by the limit
\[
\psi_{\alpha_G}^* = \lim_{|\lambda| \to \infty} \psi_{\alpha_G}(|\lambda|) = \lim_{|\lambda| \to \infty} \|1 - \alpha_G \bar{C}^{-1}\| + \alpha_G \|c_G d_G \rho_G\| \frac{|\lambda| + 1}{|\lambda| - \rho_G}
\]

since \(\mu(\rho_G, |\lambda|)\) is a monotonically decreasing function for these values. In particular, it might happen that no choice \(\alpha_G\) prevents (13) from being satisfied for \(|\lambda| \geq 1\), as depicted in Figure 2 for the Newton case with \(\bar{C}^{-1} = I\). Here, \(\alpha_G^2\) also allows for all \(\lambda \in [1, \lambda^*_G]\) and, thus, divergence of the extended fixed-point iteration \(G\). The basic reason is that the lower limit \(\psi_{\alpha_G}^*\) is strictly greater than one, which might be due to too large values for \(c_G\) and \(d_G\). Hence, \(|\lambda^*_G|\) can never be restricted below one if \(\|\bar{C}^{-1}\| c_G d_G \rho_G \geq 1\).

The remedy for this problem is simple. Note that we can always assume \(\rho_G\) being sufficiently small by considering a sequence of \(s_G\) updates for \(y_2\) before we perform an update on \(y_1\). In particular, we follow the multistep-Seidel idea and modify the extended stepping scheme (11) to be
\[
\ldots \to \text{design } u \to ((\text{state } y_2)^{s_G} \to \text{state } y_1)^{s_G} \to (\text{adjoint}, (\tilde{y}_1, \tilde{y}_2))^{s_G} \to \ldots \quad (14)
\]

with the corresponding multiple adjoint updates. Basically, we now write \(G^{s_G}\) instead of \(G\), which denotes the \(s_G\) times repeated application
\[
G^{s_G}(u, y_1, y_2) = G(u, y_1, G(u, y_1, \ldots G(u, y_1, y_2)))
\]

in all occurring equations such as (9) and (10). The derivatives \(G_{y_1}\) and \(G_{y_2}\) are replaced by
\[
G_{y_1}^{s_G} = (I + G_{y_2} + G_{y_2}^2 + \cdots + G_{y_2}^{s_G-1})G_{y_1} = (I - G_{y_2}^{s_G})(I - G_{y_2})^{-1}G_{y_1}
\]

and the product \(G_{y_2}^{s_G} = G_{y_2} \cdots G_{y_2} (s_G\text{-times})\), respectively. This modification does not alter the previous eigenvalue analysis; the quantities \(\bar{C}, \gamma_G, \nu_G, c_G, \text{ and } d_G\) of Proposition 2 are the same since the expressions \((I - G_{y_2}^{s_G})\) cancel out. The only difference is that the contraction rate \(\rho_G\) becomes \(\rho_G^{s_G}\) (i.e. the \(s_G\)th power of \(\rho_G\)). Hence, we can indeed assume that \(\rho_G\) is sufficiently small by choosing \(s_G\) sufficiently large.

A necessary condition to ensure contraction for the extended fixed-point iteration with full stepsize \(\alpha_G = 1\) and Newton preconditioner \(\bar{C}\) is given by the lower bound
\[
s_G > \max \left(0, \left[ \frac{- \log(c_G d_G)}{\log(\rho_G)} \right] \right) \in \mathbb{N}.
\]

A sufficient choice for the number of inner iterations \(s_G\) is as follows.

**Proposition 3** Let \(\gamma_G = \|I - \alpha_G \bar{C}^{-1}\| < 1\). Then by adjusting \(s_G\) and, thus, \(\rho_G^{s_G}\), any rate \(\rho_G \in (\gamma_G, 1)\) can be attained as an upper bound on the spectrum of \(G_{y_1}^{s_G}(y_2)\). Sufficient is the following relation between \(s_G, \eta_G\), and \(\rho_G\) for given \(c_G, d_G, \gamma_G\), and \(\nu_G\) :
\[
\rho_G^{s_G} \leq \rho_G(\rho_G - \gamma_G) \leq \frac{\rho_G(\rho_G - \gamma_G)}{(\rho_G - \gamma_G) + (v_G c_G d_G)(1 + \rho_G)}.
\]
Proof From the inequality (13) it follows that any eigenvalue $\rho_G$ of $G$ needs to satisfy

$$\rho_G \leq \gamma_G + (v_G c_G d_G) \rho_G^2 \frac{\rho_G + 1}{\rho_G - \rho_G^2}.$$ 

Thus, inequality (15) must hold in order to exclude values greater than $\rho_G$, as can be seen by elementary operations. ■

Moreover, the lower bound $s_G$ for the number $s_G$ follows by setting $\rho_G = 1$ in (15), which implies that

$$s_G > s_G = \log_{(1/\rho_G)}[1 + 2(v_G c_G d_G)/(1 - \gamma_G)]$$

is sufficient to enforce contraction of the extended mapping in the general case with $\gamma_G = \|I - \alpha_G \tilde{C}^{-1}\| < 1$ (i.e. $\rho_G < 1$). Moreover, we can choose the values $\alpha_G, s_G, \tilde{C}^{-1}$, and $\rho_G$ such that the resulting algorithm is “optimal” in terms of the retardation factor.

Corollary 2 Let $\tilde{\gamma}_G \in (0, 1)$ be an upper bound on $\gamma_G = \|I - \alpha_G \tilde{C}^{-1}\|$, namely, $\tilde{\gamma}_G \geq \gamma_G$. Then there exists a preconditioner $\tilde{C}^{-1}$ such that $\rho_G$ can be chosen as the harmonic mean

$$\rho_G = \frac{2}{1 + \tilde{\gamma}_G} = 2 \left(1 - \frac{1}{1 + \tilde{\gamma}_G}\right) \Longleftrightarrow \tilde{\gamma}_G = \frac{\rho_G}{2 - \rho_G}$$

for the stepsize

$$\alpha_G = \frac{1 - \rho_G}{1 + \rho_G} \in (0, 1].$$

Furthermore, the minimal cycle length $s_G$ for the choice $\rho_G, \alpha_G, \text{and } \tilde{C}^{-1}$ is given by

$$s_G(\tilde{C}) = \left[2 \log_{\rho_G} \left(1 - \frac{1}{1 + \tilde{\gamma}_G}\right) - \log_{\rho_G} \left[\left(1 - \frac{1}{1 + \tilde{\gamma}_G}\right)/2 + \|\tilde{C}^{-1}\|c_G d_G/(2 + 2\tilde{\gamma}_G)\right]\right].$$

Proof Under the stated assumptions, we can bound the right-hand side of (15) from above and deduce by elementary arguments that the minimal cycle length $s_G$ must satisfy

$$s_G^2 = \frac{\rho_G(\rho_G - \tilde{\gamma}_G)}{(\rho_G - \tilde{\gamma}_G) + (v_G c_G d_G)(1 + \rho_G)} = \frac{\rho_G^2(1 - \rho_G)}{\rho_G(1 - \rho_G) + (v_G c_G d_G)(1 + \rho_G)(2 - \rho_G)}$$

$$= \frac{\rho_G^2}{\rho_G + \|\tilde{C}^{-1}\|c_G d_G/(2 - \rho_G)} = \left(1 - \frac{1}{1 + \tilde{\gamma}_G}\right)^2 / \left(1 - \frac{1}{1 + \tilde{\gamma}_G}\right)/2 + \|\tilde{C}^{-1}\|c_G d_G/(2 + 2\tilde{\gamma}_G).$$

As a direct consequence of Corollary 2, we see that the retardation factor of the extended fixed-point iteration $G$ w.r.t. to $G^{s_G}$ is 2 in the ideal case. In particular, we have

$$\lim_{\tilde{\gamma}_G \to 0} \log \rho_G^s / \log \rho_G = 2.$$
for the proposed choices of $\alpha_G$, $\rho_G$, $s_G(\tilde{C}^{-1})$ and a sufficiently accurate preconditioner $\tilde{C}^{-1}$, which satisfies $\gamma_G = \|I - \alpha \tilde{C}^{-1}\| \leq \tilde{y}_G$. A promising upper bound on $\tilde{y}_G$ is $\rho_G$ (or $\rho_G^s$) so that $\rho_G \to 0$ (or $s \to \infty$) implies $\gamma_G \to 0$ and $\alpha_G \to 1$; in other words, a very contractive fixed-point mapping $\tilde{G}$ requires a good approximation of the preconditioner for the extended fixed-point iteration and a stepsize $\alpha_G$ close to one. Thus, for the nested approach (14) the retardation factor w.r.t. $G$ is expected to be 4 in the ideal case with a sufficient choice for $s$ and $s_G$—independent of the meshsize!

Naturally, the quantities $\rho_G$, $c_G$, and $d_G$ needed for the choice of the number of inner cycles $s_G$ are usually unknown. Therefore, we propose to approximate them by corresponding estimates that can be derived by measurements during the optimization course analogous to [3], for example, by using differences of the gradients of the Lagrangian function. However, care must be taken for the estimates $\gamma_G$ and $\|\tilde{C}^{-1}\|$ since $\tilde{C}$ is in general non-symmetric and indefinite (but not singular because of the general assumptions). In particular, it is advisable to estimate now both quantities $\|\tilde{C}^{-1}\|$ and $\|I - \alpha \tilde{C}^{-1}\|$.

A simple example is given in the next section, where the required quantities such as the Newton preconditioner can be derived analytically.

5. Numerical results

Parts of the theoretical results are validated by using a discretized version of the Poisson equation over $\Omega = [0, 1] \subset \mathbb{R}$ with constant control $u_1$ and boundary conditions,

$$-y''(t) = u_1 \text{ for } t \in \Omega \quad \text{and} \quad y(0) = y(1) = u_2. \quad (17)$$

Besides the Dirichlet conditions we require that $y(t = \frac{1}{2}) = k$ for a given constant $k \in \mathbb{R}$. For an equidistant discretization of $\Omega$ with meshsize $h = 1/2n$, we can compute the $N = 2n - 1$ discrete state variables $y^{(i)}$ using the Jacobi method [23]

$$y^{(i)}_{\text{new}} = \frac{1}{2} [h^2 u_1 + y^{(i-1)} + y^{(i+1)}], \quad \text{for } i = 1, \ldots, N$$

and set $y^{(0)} = y^{(2n)} \equiv 0$ to solve the boundary problem (17), which represents a slowly convergent fixed-point solver $G : \mathbb{R}^N \to \mathbb{R}^N$ with contraction rate $\rho_G$ close to 1. The extra pointwise requirement translates to the scalar condition $g(u_1, u_2, y) = y^{(n)} - k = 0$ and provides the additional constraint for $y^{(n)} = y^{(n)}(u_1, u_2)$. Obviously, there always exists a unique solution $y = y(u_1, u_2)$ that is a fixed point of $G$ and satisfies $g$ if one of the quantities $u_1$ or $u_2$ is fixed.

Since we are interested primarily in the contraction of the extended fixed-point iteration (2), we identify $u = u_2 \in \mathbb{R}$, $y_1 = u_1 \in \mathbb{R}$, $y_2 = y \in \mathbb{R}^{2n+1}$ and assume that $u_2 = 0$ is constant; that is, we do not consider the overall one-shot optimization (14) but only state cycles (9) to find a state $(y_1^*, y_2^*)$ satisfying the Poisson equation with zero boundary conditions. Hence, we can formulate the extended fixed-point mapping $G : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R} \times \mathbb{R}^N$ with preconditioner $C \in \mathbb{R}^{1 \times 1}$ as follows:

$$G(0, y_1, y_2) = [y_1 - \alpha_G C^{-1} \left( \frac{1}{2} [h^2 y_1 + y_2^{(n-1)} + y_2^{(n+1)}] - k \right), G(0, y_1, y_2)],$$

where the boundaries are defined to be $y_2^{(0)} = y_2^{(2n)} \equiv 0$. Moreover, we find that the corresponding derivative matrices are given by

$$g_{y_1} = [0] \in \mathbb{R}^{1 \times 1}, \quad g_{y_2} = [0, \ldots, 0, 1, 0, \ldots, 0] \in \mathbb{R}^{1 \times N},$$

$$G_{y_1} = \left[ \frac{h^2}{2}, \ldots, \frac{h^2}{2} \right] \in \mathbb{R}^{N \times 1}, \quad G_{y_2} = 0.5 \ \text{tridiag}[1, 0, 1] \in \mathbb{R}^{N \times N}.$$
Figure 3. Snapshots (every 250 iterations) of the intermediate states (gray) and the solution (orange) of $y_2$ for the extended (left) and the original (right) fixed-point iteration, where the original iteration was computed at $y_1^\ast$.

Figure 4. Convergence history of the residuals for the extended fixed-point iteration $G$ (orange) and its two components (light/dark gray) compared with the original fixed-point iteration $G$ (purple) with fixed states $y_1^{\text{init}}$ (purple) and $y_1^\ast$ (yellow) using random initial values ($k = 1$, $n = 50$, $s_G = 1$).

Figure 3 depicts the snapshots of the intermediate states $y_2$ after every 250 iterations (gray) and the solution $y_2^\ast$ (orange) of the extended and the original fixed-point mapping $G$ and $G$ (at $y_1^\ast$), respectively. Here, we use the full-step projected Newton preconditioner matrix $C$ for the choice $k = 1$, $n = 50$, and random initial values as stated in Table 1. The convergence history of the residuals for the extended fixed-point iteration $\| (y_1^\ast, y_2^\ast) - G(0, y_1^{\text{init}}, y_2^\ast) \|_2$ (orange) and its two components $\| y_1^\ast - y_1 \|_2$ (light gray) and $\| y_2^\ast - G(0, y_1^\ast, y_2) \|_2$ (dark gray) can be found in Figure 4, where we also provide the residual $\| y_2^\ast - G(0, y_1^{\text{init}}, y_2) \|_2$ of the pure original fixed-point iteration $G$ for the initial fixed state $y_1^{\text{init}}$ (purple) and its solution $y_1^\ast$ (yellow). In particular, we can deduce from the graphics that the extended mapping is a contractivity fixed-point iteration that converges toward the solution $(y_1^\ast, y_2^\ast)$.
Table 1. Matlab code example for the extended fixed-point iteration without graphical output.

maxiter = 1e4; %Number of maximum iterations
tol=1e-8; %Stopping-tolerance
k = 1.0; %Constant for pointwise condition
Ndis = 50+1; %Number N of free states y_2
h2=1.0/(Ndis-1)^2; %Mesh-size^2 of discretization
I=speye(Ndis,Ndis); %Derivatives of g and \cal{G}
gy1=0.0; gy2=zeros(1,Ndis); gy2(1,ceil(Ndis/2))=1.0;
Gy1=h2/2.*ones(Ndis,1);
Gy2=0.5*(spdiags(ones(Ndis,1),-1,Ndis,Ndis)+...
spdiags(ones(Ndis,1),1,Ndis,Ndis));
C=gy1+(gy2/(I-Gy2))*Gy1; %Projected Newton-preconditioner
rho=normest(Gy2); %Primal contraction rate
alpha=(1-rho)/(1+rho); %Step-size
y1=randn(1,1); %Random initial values
y2=randn(Ndis,1);
u=0.0; %Boundary condition value
y2(1)=u; y2(Ndis)=u;
%Extended fixed-point equation
for i=1:maxiter
  y1new=y1-alpha*(C\(gy2*Gy1*y1+gy2*Gy2*y2-k));
  y2new=Gy1*y1+Gy2*y2; y2new(1)=u; y2new(end)=u;
  res1=norm(y1-y1new)^2; res2=norm(y2-y2new)^2;
  y1=y1new;
  y2=y2new;
  if(max(res1,res2)<tol)
    break;
  end
end

6. Conclusion

We considered an extension of the multistep one-shot method presented in [3] for design optimization problems with additional equality constraints (6). The convergence theory is based on an eigenvalue analysis that suggests using the nested approach (14). The resulting method is in the limit $s, s_G \to \infty$ similar to a fully hierarchical approach, where exact feasibility is established after each iteration. Local convergence of the method can be proven for a sufficient choice of preconditioners and cycle lengths $s_G$ and $s$. The lower bound on $s_G$, which depends on problem-specific quantities and the quality of the preconditioner, was given in Corollary 2. The latter quantities can be estimated during the optimization analogous to the approximations used for $s$ presented in [3]. The retardation factor is expected to be 2 for the constraint restoration part and 4 for the overall nested multistep one-shot method in the ideal case, namely, if the preconditioner
is exact and the step-size for the Newton steps is in the limit one. Some theoretical results and observations were validated on a simple discrete test problem.

Computations for real applications have not been conducted so far. Also, the question remains open of whether corresponding results can be formulated in a functional analytic setting and how additional inequality constraints can be embedded into the approach for more general design optimization problems.

Acknowledgements

Government license section: The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (‘Argonne’). Argonne, a US Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The US Government retains for itself, and others acting on its behalf, a paid-up non-exclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This material is based upon work partly supported by the U.S. Department of Energy, Office of Science [contract number DE-AC02-06CH11357].

Note

1. The derivatives will be denoted with subscripts, and the argument is skipped whenever it is unambiguous where the derivative is evaluated; for example, $G_y$, $g_y$, and $f_y$ denote the Jacobians of $G$, $g$, and the gradient of $f$ with respect to $y$, respectively.

References


