

# A proof that Anderson acceleration improves the convergence rate in linearly converging fixed point methods (but not in those converging quadratically)

Claire Evans<sup>\*</sup>    Sara Pollock<sup>†</sup>    Leo G. Rebholz<sup>‡</sup>    Mengying Xiao<sup>§</sup>

## Abstract

This paper provides the first proof that Anderson acceleration (AA) improves the convergence rate of general fixed point iterations. AA has been used for decades to speed up nonlinear solvers in many applications, however a rigorous mathematical justification of the improved convergence rate has remained lacking. The key ideas of the analysis presented here are relating the difference of consecutive iterates to residuals based on performing the inner-optimization in a Hilbert space setting, and explicitly defining the gain in the optimization stage to be the ratio of improvement over a step of the unaccelerated fixed point iteration. The main result we prove is that AA improves the convergence rate of a fixed point iteration to first order by a factor of the gain at each step. In addition to improving the convergence rate, our results indicate that AA increases the radius of convergence. Lastly, our estimate shows that while the linear convergence rate is improved, additional quadratic terms arise in the estimate, which shows why AA does not typically improve convergence in quadratically converging fixed point iterations. Results of several numerical tests are given which illustrate the theory.

## 1 Introduction

We study an acceleration technique for fixed point problems called Anderson acceleration, in which a history of search-directions is used to improve the rate of convergence of fixed-point iterations. The method was originally introduced by D.G. Anderson in 1965 in the context of integral equations [2]. It has recently been used in many applications, including multiseant methods for fixed-point iterations in electronic structure computations [5], geometry optimization problems [12], various types of flow problems [11, 13], radiation diffusion and nuclear physics [1, 16], molecular interaction [14], machine learning [6], improving the alternating projections method for computing nearest correlation matrices [7], and on a wide range of nonlinear problems in the context of generalized minimal residual (GMRES) methods in [17]. We further refer readers to [8, 10, 11, 17] and references

---

<sup>\*</sup>School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634 (ce-evans4@g.clemson.edu)

<sup>†</sup>Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (s.pollock@ufl.edu), partially supported by NSF grant DMS1719849.

<sup>‡</sup>School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634 (rebholz@clemson.edu), partially supported by NSF grant DMS1522191.

<sup>§</sup>Department of Mathematics, College of William & Mary, Williamsburg VA 23187 (mxiao01@wm.edu).

therein for detailed discussions on both practical implementation and a history of the method and its applications.

Despite a long history of use and a strong recent interest, the first mathematical convergence results for Anderson acceleration (for both linear and nonlinear problems) appear in 2015 in [15], under the usual local assumptions for convergence of Newton iterations. However, this theory does not prove that Anderson acceleration improves the convergence of a fixed point iteration, or in other words accelerates convergence in the sense of [4]. Rather, it proves that Anderson accelerated fixed point iterations will converge in the neighborhood of a fixed point; and, an upper bound on the convergence rate is shown to approach from above the convergence rate of the underlying fixed point iteration. While an important stage in the developing theory, this does not explain the efficacy of the method, which has gained popularity as practitioners have continued to observe a dramatic speedup and increase in robustness from Anderson acceleration over a wide range of problems.

The purpose of this paper is to address this gap in the theory by proving a rigorous estimate for Anderson acceleration that shows a guaranteed improvement in the convergence rate for fixed point iterations (for general  $C^2$  functions) that converge linearly (with rate  $\kappa$ ). By explicitly defining the gain of the optimization stage at iteration  $k$  to be the ratio  $\theta_k$  of the optimized objective function compared to that of the usual fixed point method, we prove the new convergence rate is  $\theta_k((1 - \beta_{k-1}) + \beta_{k-1}\kappa)$  at step  $k$ , where  $0 < \beta_{k-1} \leq 1$  is a damping parameter and  $\beta_{k-1} = 1$  produces the undamped iteration. The key ideas to the proof are an expansion of the residual errors, developing expressions relating the difference of consecutive iterates and residuals, and explicitly factoring in the gain from the optimization stage. A somewhat similar approach is used by the authors to prove that Anderson acceleration speeds up Picard iteration convergence for finite element discretizations of the steady Navier-Stokes equations in [13] (without the  $C^2$  assumption on the fixed-point operator), and herein we extend these ideas to general fixed point iterations.

In addition to the improved linear convergence rate, our analysis also indicates that Anderson acceleration introduces quadratic error terms, which is consistent with known results that Anderson acceleration does not accelerate quadratically converging fixed point methods (see the numerical experiments section below), establishing a barrier which theoretically prevents establishing an improved convergence rate for general fixed-point iterations. A third important result we show is that both Anderson acceleration and the use of damping can extend the radius of convergence for the method, i.e. Anderson acceleration can allow the iteration to converge even when outside the domain where the fixed point function is contractive. An illustrative example of this is shown in §5.2.

This paper is arranged as follows. In §2, we review Anderson acceleration, describe the problem setting, and give some basic definitions and notation. §3 gives several important technical results to make the later analysis cleaner and simpler. §4 gives the main result of the paper, proving that the linear convergence rate is improved by Anderson acceleration, but additional quadratic error terms arise. §5 gives results from numerical tests, with the intent of illustrating the current contributions to the theory. Conclusions are given in the final section.

## 2 Anderson acceleration

In what follows, we will consider a fixed-point operator  $g : X \rightarrow X$  where  $X$  is a Hilbert space with norm  $\|\cdot\|$  and inner-product  $(\cdot, \cdot)$ . The Anderson acceleration algorithm with depth  $m$  applied to

the fixed-point problem  $g(x) = x$  reads as follows.

**Algorithm 2.1** (Anderson iteration). *The Anderson-acceleration with depth  $m \geq 0$  and damping factors  $0 < \beta_k \leq 1$  reads:*

*Step 0: Choose  $x_0 \in X$ .*

*Step 1: Find  $\tilde{x}_1 \in X$  such that  $\tilde{x}_1 = g(x_0)$ . Set  $x_1 = \tilde{x}_1$ .*

*Step k: For  $k = 1, 2, 3, \dots$  Set  $m_k = \min\{k, m\}$ .*

*[a.] Find  $\tilde{x}_{k+1} = g(x_k)$ .*

*[b.] Solve the minimization problem for  $\{\alpha_j^{k+1}\}_{j=k-m_k}^k$*

$$\min_{\sum_{j=k-m_k}^k \alpha_j^{k+1} = 1} \left\| \sum_{j=k-m_k}^k \alpha_j^{k+1} (\tilde{x}_{j+1} - x_j) \right\|. \quad (2.1)$$

*[c.] For damping factor  $0 < \beta_k \leq 1$ , set*

$$x_{k+1} = (1 - \beta_k) \sum_{j=k-m_k}^k \alpha_j^{k+1} x_j + \beta_k \sum_{j=k-m_k}^k \alpha_j^{k+1} \tilde{x}_{j+1}. \quad (2.2)$$

We will use throughout this work the stage- $k$  residual and error terms

$$e_k := x_k - x_{k-1}, \quad \tilde{e}_k := \tilde{x}_k - \tilde{x}_{k-1}, \quad w_k := \tilde{x}_k - x_{k-1}. \quad (2.3)$$

Define the following averages given by the solution  $\alpha^{k+1} = \{\alpha_j^{k+1}\}_{j=k-m_k}^k$  to the optimization problem (2.1) by

$$x_k^\alpha = \sum_{j=k-m_k}^k \alpha_j^{k+1} x_j, \quad \tilde{x}_{k+1}^\alpha = \sum_{j=k-m_k}^k \alpha_j^{k+1} \tilde{x}_{j+1}, \quad w_{k+1}^\alpha = \sum_{j=k-m_k}^k \alpha_j^{k+1} (g(x_j) - x_j). \quad (2.4)$$

Then the update (2.2) can be written in terms of the averages  $x^\alpha$  and  $\tilde{x}^\alpha$ ,

$$x_{k+1} = (1 - \beta_k) x_k^\alpha + \beta_k \tilde{x}_{k+1}^\alpha, \quad (2.5)$$

and the stage- $k$  gain  $\theta_k$  can be defined by

$$\|w_{k+1}^\alpha\| = \theta_k \|w_{k+1}\|. \quad (2.6)$$

The key to showing the acceleration of this technique defined by taking a linear combination of a history of steps corresponding to the coefficients of the optimization problem (2.1) is connecting the gain  $\theta_k$  given by (2.6) to the differences of consecutive iterates and residual terms in (2.4). As such, the success (or failure) of the algorithm to reduce the residual is coupled to the success of the optimization problem at each stage of the algorithm. As  $\alpha_k^{k+1} = 1, \alpha_j^{k+1} = 0, j \neq k$  is an admissible solution to (2.1), it follows immediately that  $0 \leq \theta_k \leq 1$ . As discussed in the remainder, the improvement in the contraction rate of the fixed-point iteration is characterized by  $\theta_k$ .

The two main components of the proof of residual convergence at an accelerated rate are the expansion of the residual  $w_{k+1}$  into  $w_k^\alpha$  and error terms  $e_{k-m_{k-1}}, \dots, e_k$ ; and, control of the  $e_j$ 's in terms of the corresponding  $w_j$ 's. In the next section, the first of these is established for general  $m$ , and the second for the particular cases of depth  $m = 1$  and  $m = 2$ , with the result then extrapolated for general  $m$ .

### 3 Technical preliminaries

There are two main technical results used in our theory. The first is an expansion of the residual, and the second is a set of estimates relating the difference of consecutive iterates to residuals. These are shown in §3.1 and §3.2, respectively. The main results which depend on these estimates are then presented in §4.

For the bounds in §3.2 relating the difference of consecutive iterates to residuals, the operator  $g : X \rightarrow X$  is assumed Lipschitz continuous and contractive, as in [13]; see Assumption 3.2, below. The results of §3.1 do not require the contractive property, but require the assumption that  $g$  is twice continuously differentiable to allow for Taylor expansions of the error terms. We denote the derivatives of  $g$  by  $g'(\cdot; \cdot)$  and  $g''(\cdot; \cdot, \cdot)$ , and employ the standard notation that forms  $g'(\cdot; \cdot)$  and  $g''(\cdot; \cdot, \cdot)$  are linear with respect to the arguments to the right of the semicolon.

**Assumption 3.1.** *Let  $X$  be a Hilbert space and  $g : X \rightarrow X$ . Assume  $g$  has a fixed point  $x^* \in X$ , and there are positive constants  $\kappa$  and  $\hat{\kappa}$  with*

1.  $g \in C^2(X)$ .
2.  $\|g'(y; u)\| \leq \kappa \|u\|$  for each  $y$  and all  $u \in X$ .
3.  $\|g''(y; u, v)\| \leq \hat{\kappa} \|u\| \|v\|$  for each  $y$  and all  $u, v \in X$ .

**Assumption 3.2.** *Let  $X$  be a Hilbert space and  $g : X \rightarrow X$ . Assume  $\|g(y) - g(x)\| \leq \kappa \|x - y\|$  for every  $x, y \in X$ , with  $\kappa < 1$ .*

By standard fixed-point theory, Assumption 3.2 implies the existence of a unique fixed-point  $x^*$  of  $g$  in  $X$ . In a slight abuse of notation, the difference of consecutive iterates,  $e_k = x_k - x_{k-1}$  is loosely referred to in this manuscript as an error term. As shown carefully in [13], the true error  $x_k - x^*$  is controlled in norm by  $e_j$ ,  $j = k - m_k, \dots, k$ , for the depth  $m$  algorithm so long as the coefficients from the optimization remain bounded. In the results of §4, the residual  $w_k$  is shown to converge to zero under Assumption 3.2. This is sufficient to establish convergence of the error  $x_k - x^*$  to zero as

$$\|x_k - x^*\| \leq \|x_k - g(x_k)\| + \|g(x_k) - g(x^*)\| \leq \|w_{k+1}\| + \kappa \|x_k - x^*\|,$$

by which  $\|x_k - x^*\| \leq (1 - \kappa)^{-1} \|w_{k+1}\|$ .

#### 3.1 Expansion of the residual

Based on Assumption 3.1 the error term  $\tilde{e}_k$  of (2.4) has a Taylor expansion

$$\tilde{e}_{k+1} := g(x_k) - g(x_{k-1}) = \int_0^1 g'(z_k(t); e_k) \, dt, \quad (3.1)$$

where  $z_k(t) = x_{k-1} + te_k$ . For each  $t \in [0, 1]$  a second application of Taylor's Theorem provides

$$g'(z_{k-1}(t); \cdot) = g'(z_k(t); \cdot) + \int_0^1 g''(\hat{z}_{k,t}(s); z_{k-1}(t) - z_k(t), \cdot) \, ds, \quad (3.2)$$

where  $\hat{z}_{k,t}(s) = z_{k-1}(t) + s(z_k(t) - z_{k-1}(t))$ . Using (3.1)-(3.2) we next derive an expansion of the residual  $w_{k+1}$  in terms of the differences of consecutive iterates  $e_k, \dots, e_{k-m_{k-1}}$ . We start with the definition of the residual by (2.4) and the expansion of iterate  $x_k$  by the update (2.5).

$$w_{k+1} = g(x_k) - x_k = (1 - \beta_{k-1})(g(x_k) - x_{k-1}^\alpha) + \beta_{k-1}(g(x_k) - \tilde{x}_k^\alpha). \quad (3.3)$$

Expanding the first term on the right hand side of (3.3) yields

$$\begin{aligned} g(x_k) - x_{k-1}^\alpha &= \sum_{j=k-m_{k-1}-1}^{k-1} \alpha_j^k (g(x_k) - x_j) \\ &= \sum_{j=k-m_{k-1}-1}^{k-1} \alpha_j^k (g(x_j) - x_j) + \sum_{j=k-m_{k-1}}^k \left( \sum_{n=k-m_{k-1}-1}^{j-1} \alpha_n^k \right) (g(x_j) - g(x_{j-1})) \\ &= w_k^\alpha + \sum_{j=k-m_{k-1}}^k \gamma_j \tilde{e}_{j+1}, \end{aligned} \quad (3.4)$$

where

$$\gamma_j := \sum_{n=k-m_{k-1}-1}^{j-1} \alpha_n^k. \quad (3.5)$$

It is worth noting that  $\gamma_k = 1$ . Expanding the second term on the right hand side of (3.3), we get

$$g(x_k) - \tilde{x}_k^\alpha = \sum_{j=k-m_{k-1}-1}^{k-1} \alpha_j^k (g(x_k) - g(x_j)) = \sum_{j=k-m_{k-1}}^k \gamma_j \tilde{e}_{j+1}. \quad (3.6)$$

Reassembling (3.3) with (3.4) and (3.6) followed by (3.1), we have

$$w_{k+1} = (1 - \beta_{k-1})w_k^\alpha + \sum_{j=k-m_{k-1}}^k \gamma_j \tilde{e}_{j+1} = (1 - \beta_{k-1})w_k^\alpha + \sum_{j=k-m_{k-1}}^k \gamma_j \int_0^1 g'(z_j(t); e_j) \, dt. \quad (3.7)$$

We now take a closer look at the last term of (3.7). For each  $j = k - m_{k-1}, \dots, k - 1$ , adding and subtracting intermediate averages allows

$$\int_0^1 g'(z_j(t); e_j) \, dt = \int_0^1 g'(z_k(t); e_j) \, dt + \sum_{n=j}^{k-1} \int_0^1 g'(z_n(t); e_j) - g'(z_{n+1}(t); e_j) \, dt. \quad (3.8)$$

Applying now (3.2) to each summand of (3.8) then yields

$$\int_0^1 g'(z_j(t); e_j) \, dt = \int_0^1 g'(z_k(t); e_j) \, dt + \sum_{n=j}^{k-1} \int_0^1 \int_0^1 g''(\hat{z}_{n+1,t}(s); z_n(t) - z_{n+1}(t), e_j) \, ds \, dt. \quad (3.9)$$

Summing over the  $j$ 's after (3.9) is applied to each term, the sum on the right hand side of (3.7) may be expressed as

$$\begin{aligned} \sum_{j=k-m_{k-1}}^k \gamma_j \int_0^1 g'(z_j(t); e_j) \, dt &= \int_0^1 g'(z_k(t); \sum_{j=k-m_{k-1}}^k \gamma_j e_j) \, dt \\ &+ \sum_{j=k-m_{k-1}}^{k-1} \int_0^1 \int_0^1 \left( \sum_{n=j}^{k-1} g''(\hat{z}_{n+1,t}(s); z_n(t) - z_{n+1}(t), \gamma_j e_j) \right) \, ds \, dt. \end{aligned} \quad (3.10)$$

The next calculation shows that  $\sum_{j=k-m_{k-1}}^k \gamma_j e_j$  is equal to  $\beta_{k-1} w_k^\alpha$ . First observe that  $\gamma_j - \gamma_{j-1} = \alpha_{j-1}^k$  and  $\gamma_{k-m_{k-1}} = \alpha_{k-m_{k-1}-1}^k$ . Separating the first term of the sum and using  $\gamma_k = 1$ ,

$$\begin{aligned} \sum_{j=k-m_{k-1}}^k \gamma_j e_j &= x_k - x_{k-1} + \sum_{j=k-m_{k-1}}^{k-1} \gamma_j (x_j - x_{j-1}) \\ &= x_k - x_{k-1} + \gamma_{k-1} x_{k-1} - \sum_{j=k-m_{k-1}-1}^{k-2} \alpha_j^k x_j \\ &= x_k - \alpha_{k-1}^k x_{k-1} - \sum_{j=k-m_{k-1}-1}^{k-2} \alpha_j^k x_j = x_k - x_{k-1}^\alpha. \end{aligned} \quad (3.11)$$

From (3.11) and the decomposition of  $x_k$  in terms of update (2.2), we have that

$$\sum_{j=k-m_{k-1}}^k \gamma_j e_j = x_k - x_{k-1}^\alpha = (1 - \beta_{k-1}) x_{k-1}^\alpha + \beta_{k-1} \tilde{x}_k^\alpha - x_{k-1}^\alpha = \beta_{k-1} (\tilde{x}_k^\alpha - x_{k-1}^\alpha) = \beta_{k-1} w_k^\alpha. \quad (3.12)$$

Putting (3.12) together with (3.10) and (3.7) then yields

$$\begin{aligned} w_{k+1} &= \int_0^1 (1 - \beta_{k-1}) w_k^\alpha + \beta_{k-1} g'(z_k(t); w_k^\alpha) \, dt \\ &+ \sum_{j=k-m_{k-1}}^{k-1} \int_0^1 \int_0^1 \left( \sum_{n=j}^{k-1} g''(\hat{z}_{n+1,t}(s); z_n(t) - z_{n+1}(t), \gamma_j e_j) \right) \, ds \, dt. \end{aligned} \quad (3.13)$$

Based on the expansion of  $w_{k+1}$  by (3.13) we now proceed to bound the higher order terms in the particular cases  $m = 1$  and  $m = 2$  to establish convergence of Algorithm 2.1 at an accelerated rate.

### 3.2 Relating the difference of consecutive iterates to residuals

We now derive estimates to bound (in norm) the  $e_j$ 's from the right hand side of (3.13) by the corresponding  $w_j$ 's. The bounds in this subsection hold under Assumption 3.2, namely  $g$  is a contractive operator.

Under Assumption 3.2 we have the inequality

$$(1 - \kappa) \|e_n\| \leq \|e_n\| - \|\tilde{e}_{n+1}\| \leq \|\tilde{e}_{n+1} - e_n\| = \|w_{n+1} - w_n\|. \quad (3.14)$$

The next lemma establishes a bound for  $e_{j-1}$  in terms of  $w_j$  and  $w_{j-1}$  in the case of depth  $m = 1$ . The subsequent lemma generalizes the same idea for general  $m$ .

**Lemma 3.1.** *Under the conditions of Assumption 3.2, the following bounds hold true:*

$$|\alpha_{j-1}^j| \|e_{j-1}\| \leq \frac{1}{1 - \kappa} \|w_{j-1}\|, \quad (3.15)$$

$$|\alpha_{j-2}^j| \|e_{j-1}\| \leq \frac{1}{1 - \kappa} \|w_j\|. \quad (3.16)$$

*Proof.* Begin by rewriting the optimization problem (2.1) in the equivalent form

$$\eta = \operatorname{argmin} \|w_{j-1} + \eta(w_j - w_{j-1})\|^2,$$

where  $\alpha_{j-1}^j = \eta$  and  $\alpha_{j-2}^j = 1 - \eta$ . The critical point  $\eta$  then satisfies  $\eta \|w_j - w_{j-1}\|^2 = (w_{j-1}, w_j - w_{j-1})$ . Applying Cauchy-Schwarz and triangular inequalities yields  $|\eta| \|w_j - w_{j-1}\| \leq \|w_{j-1}\|$ . Applying (3.14) with  $n = j - 1$  yields the result (3.15).

Next, rewrite the optimization problem (2.1) in another equivalent form,

$$\gamma = \operatorname{argmin} \|w_j - \gamma(w_j - w_{j-1})\|^2, \quad (3.17)$$

where the equivalence follows with  $\alpha_{j-2}^j = \gamma$  and  $\alpha_{j-1}^j = 1 - \gamma$ . Following the same procedure as above yields  $|\gamma| \|w_j - w_{j-1}\| \leq \|w_j\|$ . Applying (3.14) at level  $n - 1$  then yields the second result (3.16).  $\square$

The use of  $\gamma$  as the second parameter of in the proof above is not purely coincidental, as this  $\gamma$  agrees with the  $\gamma_{j-1}^j$  used in §3.1. The same essential technique yields the necessary bounds for  $m \geq 2$ . The estimate for general  $m$  is given in the lemma below, with the particular estimate for  $m = 2$  given as a proposition.

As in the  $m = 1$  case above, two forms of the optimization problem are used. The  $\gamma$ -formulation is used to bound the terms  $\gamma_j \|e_j\|$  that appear from the expansion (3.13); whereas, the  $\eta$ -formulation is used to bound the terms  $\|e_j\|$  that appear in the numerator without leading optimization coefficients. It is then of particular importance that estimates of the form  $c \|e_j\| \leq \Sigma k_n \|w_n\|$  have the property that  $c$  is bounded away from zero. This is a reasonable assumption on the leading coefficient  $c = \alpha_{k-1}^k$  for each  $k$ , as some nonvanishing component in the latest search direction is necessary for progress. It is also a reasonable assumption on  $c = 1 - \alpha_{k-m_{k-1}-1}^k$ , meaning the coefficient of the earliest search direction considered is bounded away from unity. Presumably,  $|\alpha_{k-m_{k-1}-1}^k| < 1$  is a reasonable assumption to make, although this is not explicitly required (*cf.*, [13]).

**Lemma 3.2.** *Under the conditions of Assumption 3.2, the following bounds hold true:*

$$|\alpha_{j-1}^j| \|e_{j-1}\| \leq \frac{1}{1-\kappa} \left( |\eta_{j-1}| \|w_{j-1}\| + \sum_{n=j-m_{j-1}-1}^{j-2} |\alpha_{n-1}^j| \|w_n\| \right) \quad (3.18)$$

$$|1 - \alpha_{j-m_{j-1}-1}^j| \|e_{j-m_{j-1}}\| \leq \frac{1}{1-\kappa} \left( \sum_{n=j-m+2}^j |\alpha_{n-1}^j| \|w_n\| + |\eta_{j-m+2}| \|w_{j-m+1}\| + \|w_{j-m}\| \right) \quad (3.19)$$

$$\begin{aligned} |\gamma_{p-1}| \|e_{p-1}\| &\leq \frac{1}{1-\kappa} \left( \sum_{n=j-m_{j-1}}^{p-2} |\alpha_{n-1}^j| \|w_n\| + |\gamma_{p-2}| \|w_{p-1}\| + |\gamma_p| \|w_p\| \right. \\ &\quad \left. + \sum_{n=p+1}^j |\alpha_{n-1}^j| \|w_n\| \right). \end{aligned} \quad (3.20)$$

with  $\eta_{j-1} = \alpha_{j-1}^j + \alpha_{j-2}^j$  as in (3.21), and  $\gamma_p, \gamma_{p-1}, \gamma_{p-2}$ , given below by (3.22).

*Proof.* The optimization problem (2.1) at level  $j$  is to minimize

$$\left\| \sum_{n=j-m_{j-1}-1}^{j-1} \alpha_n^j w_{n+1} \right\| \quad \text{subject to} \quad \sum_{n=j-m_{j-1}-1}^{j-1} \alpha_n^j = 1.$$

Differencing from the left and right respectively, this can be posed as the following unconstrained optimization problems:

$$\text{minimize} \quad \left\| w_{j-m_{j-1}} + \sum_{n=j-m_{j-1}+1}^j \eta_n (w_n - w_{n-1}) \right\|^2, \quad \eta_n = \sum_{i=n-1}^{j-1} \alpha_i^j. \quad (3.21)$$

$$\text{minimize} \quad \left\| w_j - \sum_{n=j-m_{j-1}+1}^j \gamma_{n-1} (w_n - w_{n-1}) \right\|^2, \quad \gamma_n = \sum_{i=j-m_{j-1}-1}^{n-1} \alpha_i^j. \quad (3.22)$$

Note that (3.22) coincides with (3.5) which agrees with the unconstrained form of the optimization problem in for instance [5]. To help reduce notation, denote  $m = m_{j-1}$  for the remainder of the proof.

Starting with estimate (3.18) we are concerned with bounding in norm the leading term difference term  $w_j - w_{j-1}$ . Expanding the norm squared (3.21) as an inner-product and seeking the critical point for  $\eta_j$  yields

$$\eta_j \|w_j - w_{j-1}\|^2 + (w_j - w_{j-1}, w_{j-m}) + \sum_{n=j-m+1}^{j-1} \eta_n (w_j - w_{j-1}, w_n - w_{n-1}) = 0.$$

Recombining the terms inside the sum, noting  $\eta_{n-1} - \eta_n = \alpha_{n-2}^j$ , and  $\eta_j = \alpha_{j-1}^j$  obtain

$$\alpha_{j-1}^j \|w_j - w_{j-1}\|^2 = -(\alpha_{j-1}^j + \alpha_{j-2}^j)(w_j - w_{j-1}, w_{j-1}) - \sum_{n=j-m}^{j-2} \alpha_{n-1}^j (w_j - w_{j-1}, w_n).$$

Applying Cauchy-Schwarz and triangle inequalities then yields

$$|\alpha_{j-1}^j| \|w_j - w_{j-1}\| \leq |\alpha_{j-1}^j + \alpha_{j-2}^j| \|w_{j-1}\| + \sum_{n=j-m}^{j-2} \alpha_{n-1}^j \|w_n\|.$$

Applying (3.14), the result (3.18) follows.

Following the same idea for estimate (3.19), we are now concerned with bounding in norm the final difference term  $w_{j-m+1} - w_{j-m}$ . Again expanding (3.21) as an inner-product and seeking the critical point this time for  $\eta_{j-m+1}$  yields

$$\eta_{j-m+1} \|w_{j-m+1} - w_{j-m}\|^2 + (w_{j-m+1} - w_{j-m}, w_{j-m}) + \sum_{n=j-m+2}^j \eta_n (w_{j-m+1} - w_{j-m}, w_n - w_{n-1}) = 0.$$

Recombining terms noting  $\eta_{j-m+1} = 1 - \alpha_{j-m-1}^j$

$$\begin{aligned} (1 - \alpha_{j-m-1}^j) \|w_{j-m+1} - w_{j-m}\|^2 &= \sum_{n=j-m+2}^j \alpha_{n-1}^j (w_{j-m+1} - w_{j-m}, w_n) \\ &\quad - (w_{j-m+1} - w_{j-m}, \eta_{j-m+2} w_{j-m+1} + w_{j-m}). \end{aligned}$$

Applying Cauchy-Schwarz and triangle inequalities then yields

$$|1 - \alpha_{j-m-1}^j| \|w_{j-m+1} - w_{j-m}\| \leq \left( \sum_{n=j-m+2}^j |\alpha_{n-1}^j| \|w_n\| \right) + |\eta_{j-m+2}| \|w_{j-m+1}\| + \|w_{j-m}\|.$$

The result (3.19) follows by (3.14).

Similarly for (3.20), expanding the norm of (3.22) as an inner product and seeking the critical point for each  $\gamma_p$  yields

$$\gamma_{p-1} \|w_p - w_{p-1}\|^2 = (w_p - w_{p-1}, w_j) - \sum_{n=j-m+1, n \neq p}^j \gamma_{n-1} (w_p - w_{p-1}, w_n - w_{n-1}).$$

Recombining the terms inside the sum using  $\gamma_n - \gamma_{n-1} = \alpha_{n-1}^j$ , and  $\gamma_{j-m} = \alpha_{j-m-1}^j$ , we obtain

$$\begin{aligned} \gamma_{p-1} \|w_p - w_{p-1}\|^2 &= \sum_{n=j-m}^{p-2} \alpha_{n-1}^j (w_p - w_{p-1}, w_n) - \gamma_{p-2} (w_p - w_{p-1}, w_{p-1}) + \gamma_p (w_p - w_{p-1}, w_p) \\ &\quad + \sum_{n=p+1}^j \alpha_{n-1}^j (w_p - w_{p-1}, w_n). \end{aligned}$$

Applying now Cauchy-Schwarz and triangle inequalities,

$$|\gamma_{p-1}| \|w_p - w_{p-1}\| \leq \sum_{n=j-m}^{p-2} |\alpha_{n-1}^j| \|w_n\| + |\gamma_{p-2}| \|w_{p-1}\| + |\gamma_p| \|w_p\| + \sum_{n=p+1}^j |\alpha_{n-1}^j| \|w_n\|.$$

Applying (3.14), the result (3.20) follows.  $\square$

For the convenience of subsequent calculations, the bounds (3.21) and (3.22) used to bound  $\|w_{k+1}\|$  for the case of depth  $m = 2$  are summarized in the following proposition.

**Proposition 3.3** (Depth  $m = 2$ ). *With depth  $m = 2$  the estimates (3.18) and (3.20) reduce to*

$$|\alpha_{j-1}^j| \|e_{j-1}\| \leq \frac{1}{1-\kappa} \left( |\alpha_{j-1}^j + \alpha_{j-2}^j| \|w_{j-1}\| + |\alpha_{j-3}^j| \|w_{j-2}\| \right) \quad (3.23)$$

$$|1 - \alpha_{j-2}^j| \|e_{j-2}\| \leq \frac{1}{1-\kappa} \left( |\alpha_{j-1}^j| \|w_j\| + |\alpha_{j-1}^j| \|w_{j-1}\| + \|w_{j-2}\| \right) \quad (3.24)$$

$$|\gamma_{j-1}| \|e_{j-1}\| \leq \frac{1}{1-\kappa} \left( \|w_j\| + |\alpha_{j-3}^j| \|w_{j-1}\| + |\alpha_{j-3}^j| \|w_{j-2}\| \right) \quad (3.25)$$

$$|\gamma_{j-2}| \|e_{j-2}\| \leq \frac{1}{1-\kappa} \left( |\alpha_{j-1}^j| \|w_j\| + |1 - \alpha_{j-1}^j| \|w_{j-1}\| \right). \quad (3.26)$$

The second two bounds (3.25) and (3.26) follow from (3.20) noting from (3.22), that for  $m = 2$  we have  $\gamma_{j-2} = \alpha_{j-3}^j$ ,  $\gamma_{j-1} = 1 - \alpha_{j-1}^j$  and  $\gamma_j = 1$ . The approach taken in [13] is to reduce the right hand side of (3.24) and (3.25) to two terms each by relating their expansion to that of (3.23) and (3.26), respectively. Here the terms are left as they are to emphasize the direct generality to greater depth  $m$ .

### 3.3 Explicit computation of the optimization gain

The stage- $k$  gain  $\theta_k$  has a simple description assuming the optimization is performed over a norm  $\|\cdot\|$  induced by an inner product  $(\cdot, \cdot)$ , in other words in a Hilbert space setting.

Consider the unconstrained  $\gamma$ -form of the optimization problem (3.22) at iteration  $k$  with depth  $m$ : Find  $\gamma_{k-m+1}, \dots, \gamma_k$  that minimize

$$\left\| w_{k+1} - \sum_{n=k-m+1}^k \gamma_n (w_{n+1} - w_n) \right\|^2 = \|w_{k+1} - F^k \gamma^k\|^2, \quad (3.27)$$

Where  $F$  is the matrix with columns  $w_{n+1} - w_n$ ,  $n = k-m+1, \dots, k$  and  $\gamma^k$  is the corresponding vector of coefficients  $\gamma_{k-m+1}, \dots, \gamma_k$ . Indeed, (3.27) (or equivalently reindexed) is the preferred way to state the optimization problem [17], particularly in the case where  $\|\cdot\|$  is the  $l_2$  norm and a fast  $QR$  algorithm can be used.

This is also the preferred statement of the problem to understand the gain  $\theta_k$  from (2.6), which satisfies  $\|w_{k+1}^\alpha\| = \theta_k \|w_{k+1}\|$ . Define the unique decomposition  $w_{k+1} = w_R + w_N$  with  $w_R \in \text{Range}(F^k)$  and  $w_N \in \text{Null}((F^k)^T)$ . Then  $w_N$  is the least-squares residual satisfying  $\|w_N\| = \|w_{k+1} - F^k \delta^k\| = \|w_{k+1}^\alpha\| = \theta_k \|w_{k+1}\|$  meaning

$$\theta_k = \sqrt{1 - \frac{\|w_R\|^2}{\|w_{k+1}\|^2}}, \quad (3.28)$$

and,  $\theta_k$  has the interpretation of the direction-sine between  $w_{k+1}$  and the subspace spanned by  $\{w_{n+1} - w_n\}_{n=k-m+1}^k$ . This is particularly clear in the case  $m = 1$  where by solving for the critical point  $\gamma$  of (3.17) yields

$$\gamma = \frac{(w_{k+1}, w_{k+1} - w_k)}{\|w_{k+1} - w_k\|^2}.$$

Expanding  $\theta_k^2 \|w_{k+1}\|^2 = \|w_{k+1} - \gamma(w_{k+1} - w_k)\|^2$  and using the particular value of  $\gamma$  above yields

$$1 - \theta_k^2 = \frac{(w_{k+1}, w_{k+1} - w_k)^2}{\|w_{k+1} - w_k\|^2 \|w_{k+1}\|^2},$$

with the clear interpretation that  $(1 - \theta_k^2)^{1/2}$  is the direction cosine between  $w_{k+1}$  and  $w_{k+1} - w_k$ , hence  $\theta_k$  is the direction-sine.

If indeed an (economy)  $QR$  algorithm  $F^k = Q_1 R_1$  is used to solve the optimization problem then  $\theta_k = \sqrt{1 - (\|Q_1^T w_{k+1}\| / \|w_{k+1}\|)^2}$ , which can be used to predict whether an accelerated step would be (sufficiently) beneficial. This explicit computation of  $\theta_k$  is used in §5.3 to propose an adaptive damping strategy based on the gain at each step. Finally, it is noted that the improvement in the gain  $\theta_k$  as  $m$  is increased depends on sufficient linear independence or small direction cosines between the columns of  $F^k$ , as information from earlier in the history is added. This is discussed in some greater depth in [17].

## 4 Convergence rates for depths $m = 1$ and $m = 2$

First we put the expansion (3.13) together with the bounds (3.15)-(3.16) for a convergence proof for the simplest case of  $m = 1$ .

**Theorem 4.1** (Convergence of the residual with depth  $m = 1$ ). *On satisfaction of Assumptions 3.1 and 3.2, if the coefficients  $\alpha_k^{k+1}, \alpha_{k-1}^k$  remain bounded and bounded away from zero, the following bound holds for the residual  $w_{k+1}$  from Algorithm 2.1 with depth  $m = 1$ :*

$$\|w_{k+1}\| \leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \mathcal{O}(\|w_k\|^2) + \mathcal{O}(\|w_{k-1}\|^2).$$

**Remark 4.1.** *The assumptions on the coefficients  $\alpha_j^k$  arising from the optimization problem are similar to those of [15]. These assumptions could be eliminated by solving instead a constrained optimization problem that enforces boundedness of the parameters, resulting in a modified gain  $\hat{\theta}_k$  which satisfies  $\theta_k \leq \hat{\theta}_k \leq 1$ .*

*Proof.* In this case the expansion found for  $w_{k+1}$  in (3.13) reduces to

$$\begin{aligned} w_{k+1} &= \int_0^1 (1 - \beta_{k-1}) w_k^\alpha + \beta_{k-1} g'(z_k(t); w_k^\alpha) dt \\ &\quad + \int_0^1 \int_0^1 g''(\hat{z}_{k,t}(s); z_{k-1}(t) - z_k(t), \gamma_{k-1} e_{k-1}) ds dt. \end{aligned} \quad (4.1)$$

Taking norms of both sides and applying Assumption 3.1, (2.6) and the triangle inequality,

$$\|w_{k+1}\| \leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \hat{\kappa}(\|e_k\| + \|e_{k-1}\|) \gamma_{k-1} \|e_{k-1}\|. \quad (4.2)$$

The preceding bound (4.2) holds regardless of whether  $g$  is globally contractive (Assumption 3.2), hence for error terms  $\|e_k\|$  and  $\|e_{k-1}\|$  small enough, contraction of the error may be observed depending on the search direction, particularly if a damping factor  $0 < \beta < 1$  is applied, and if the gain  $\theta_k$  is sufficiently less than one. This justifies the observation that Anderson acceleration can enlarge the effective domain of convergence of a fixed point iteration.

For the remainder of the calculation, we consider the case of a contractive operator, meaning Assumption 3.2 is satisfied. Applying (3.16) with  $j = k$  to the  $\gamma_{k-1} \|e_{k-1}\|$ , recalling by (3.5) we have  $\gamma_{k-1} = \alpha_{k-2}^k$ ; and, applying (3.15) with  $j = k + 1$  and  $j = k$  respectively to the remaining  $\|e_k\|$  and  $\|e_{k-1}\|$  allows

$$\begin{aligned} \|w_{k+1}\| &\leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \frac{\hat{\kappa}}{(1 - \kappa)^2} \left( \frac{\|w_k\|}{\alpha_k^{k+1}} + \frac{\|w_{k-1}\|}{\alpha_{k-1}^k} \right) \|w_k\| \\ &= \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \mathcal{O}(\|w_k\|^2) + \mathcal{O}(\|w_{k-1}\|^2). \end{aligned} \quad (4.3)$$

□

As discussed in §3.2,  $\alpha_k^{k+1}$  and  $\alpha_{k-1}^k$  are each the leading coefficients in their respective optimization problems, multiplying the most recent iterate. As such, these coefficients may be reasonably considered bounded away from zero.

The case of  $m = 2$  follows similarly, combining (3.13) with (3.23)-(3.26).

**Theorem 4.2** (Convergence of the residual with depth  $m = 2$ ). *On satisfaction of Assumptions 3.1 and 3.2, if the coefficients  $\alpha_{k-3}^k, \dots, \alpha_{k-1}^k$  remain bounded, and  $\alpha_{k-1}^k$  and  $1 - \alpha_{k-3}^k$  remain bounded away from zero, the following bound holds for the residual  $w_{k+1}$  from Algorithm 2.1 with depth  $m = 2$ .*

$$\|w_{k+1}\| \leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \mathcal{O}(\|w_k\|^2) + \mathcal{O}(\|w_{k-1}\|^2) + \mathcal{O}(\|w_{k-2}\|^2).$$

*Proof.* For depth  $m = 2$  the residual expansion (3.13) reduces to

$$\begin{aligned} w_{k+1} &= \int_0^1 (1 - \beta_{k-1}) w_k^\alpha + \beta_{k-1} g'(z_k(t); w_k^\alpha) \, dt \\ &\quad + \int_0^1 \int_0^1 g''(\hat{z}_{k,t}(s); z_{k-1}(t) - z_k(t), \gamma_{k-1} e_{k-1}) \, ds \, dt. \\ &\quad + \int_0^1 \int_0^1 g''(\hat{z}_{k-1,t}(s); z_{k-2}(t) - z_{k-1}(t), \gamma_{k-2} e_{k-2}) \, ds \, dt. \\ &\quad + \int_0^1 \int_0^1 g''(\hat{z}_{k,t}(s); z_{k-1}(t) - z_k(t), \gamma_{k-2} e_{k-2}) \, ds \, dt. \end{aligned}$$

Taking norms of both sides and applying (2.6) and the triangle inequality,

$$\begin{aligned} \|w_{k+1}\| &\leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \hat{\kappa}(\|e_k\| + \|e_{k-1}\|) |\gamma_{k-1}| \|e_{k-1}\| \\ &\quad + \hat{\kappa}(\|e_{k-2}\| + 2\|e_{k-1}\| + \|e_k\|) |\gamma_{k-2}| \|e_{k-2}\|. \end{aligned} \quad (4.4)$$

Applying (3.25) and (3.26) to (4.4) yields

$$\begin{aligned} \|w_{k+1}\| &\leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| \\ &\quad + \frac{\hat{\kappa}}{1 - \kappa} (\|e_k\| + \|e_{k-1}\|) \left( \|w_k\| + |\alpha_{k-3}^k| \|w_{k-1}\| + |\alpha_{k-3}^k| \|w_{k-2}\| \right) \\ &\quad + \frac{\hat{\kappa}}{1 - \kappa} (\|e_k\| + 2\|e_{k-1}\| + \|e_{k-2}\|) \left( |\alpha_{k-1}^k| \|w_k\| + |1 - \alpha_{k-1}^k| \|w_{k-1}\| \right). \end{aligned} \quad (4.5)$$

Applying (3.23) with  $j = k + 1$  and  $j = k$  together with (3.24) to (4.5) then yields

$$\begin{aligned}
\|w_{k+1}\| &\leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| \\
&+ \frac{\hat{\kappa}}{(1 - \kappa)^2} \left( \frac{1}{\alpha_k^{k+1}} \left( |\alpha_k^{k+1} + \alpha_{k-1}^{k+1}| \|w_k\| + |\alpha_{k-2}^{k+1}| \|w_{k-1}\| \right) \right. \\
&+ \frac{1}{\alpha_{k-1}^k} \left( |\alpha_{k-1}^k + \alpha_{k-2}^k| \|w_{k-1}\| + |\alpha_{k-3}^k| \|w_{k-2}\| \right) \Big) \\
&\times \left( \|w_k\| + |\alpha_{k-3}^k| \|w_{k-1}\| + |\alpha_{k-3}^k| \|w_{k-2}\| \right) \\
&+ \frac{\hat{\kappa}}{(1 - \kappa)^2} \left( \frac{1}{\alpha_k^{k+1}} \left( |\alpha_k^{k+1} + \alpha_{k-1}^{k+1}| \|w_k\| + |\alpha_{k-2}^{k+1}| \|w_{k-1}\| \right) \right. \\
&+ \frac{1}{\alpha_{k-1}^k} \left( |\alpha_{k-1}^k + \alpha_{k-2}^k| \|w_{k-1}\| + |\alpha_{k-3}^k| \|w_{k-2}\| \right) \\
&+ \frac{1}{1 - \alpha_{k-3}^k} \left( |\alpha_{k-1}^k| \|w_k\| + |\alpha_{k-1}^k| \|w_{k-1}\| + \|w_{k-2}\| \right) \Big) \\
&\times \left( |\alpha_{k-1}^k| \|w_k\| + |1 - \alpha_{k-1}^k| \|w_{k-1}\| \right). \tag{4.6}
\end{aligned}$$

And, (4.6) satisfies

$$\|w_{k+1}\| \leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \mathcal{O}(\|w_k\|^2) + \mathcal{O}(\|w_{k-1}\|^2) + \mathcal{O}(\|w_{k-2}\|^2), \tag{4.7}$$

where the higher order terms have bounded coefficients.  $\square$

**Remark 4.2.** To avoid the extra assumption that  $|1 - \alpha_{k-3}^k|$  remains bounded away from zero, the term  $\|e_{k-2}\|$  of (4.4) could be bounded instead by (3.20) with  $j = k - 1$ , by which (4.7) is replaced by

$$\begin{aligned}
\|w_{k+1}\| &\leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| \\
&+ \mathcal{O}(\|w_k\|^2) + \mathcal{O}(\|w_{k-1}\|^2) + \mathcal{O}(\|w_{k-2}\|^2) + \mathcal{O}(\|w_{k-3}\|^2).
\end{aligned}$$

Moreover this generalizes to higher order.

Finally, we state without proof the general result which can be extrapolated from (3.13) and (3.18)-(3.20) as was done explicitly for depth  $m = 2$ , above.

**Proposition 4.3.** On satisfaction of Assumptions 3.1 and 3.2, if the coefficients  $\alpha_{k-m-1}^k, \dots, \alpha_{k-1}^k$  remain bounded, and  $\alpha_{k-1}^k$  and  $1 - \alpha_{k-m-1}^k$  remain bounded away from zero, the following bound holds for the residual  $w_{k+1}$  from Algorithm 2.1 with depth  $m$ .

$$\|w_{k+1}\| \leq \theta_k((1 - \beta_{k-1}) + \kappa\beta_{k-1}) \|w_k\| + \sum_{j=0}^m \mathcal{O}(\|w_{k-j}\|^2).$$

As discussed in §3.3, even as the higher order terms accumulate, there is still an advantage to some extent to considering greater depth  $m$ , due to the improved gain from the optimization problem. However, in practice this must be weighed against the computational cost of raising  $m$  (which can become significant) and the accuracy of one's optimization solver. In our tests, little improvement is found past  $m = 3$ .

## 5 Numerical tests

We now give results of several numerical tests that illustrate the theory above. In particular, we illustrate that Anderson speeds up linear convergence, slows down quadratic convergence, and increases the radius of convergence in agreement with the presented theory. It is not our purpose in this section to show how well Anderson acceleration works on a wide variety of problems; for this, see the references in the introduction.

### 5.1 Simple illustrative tests for the scalar case

We start with results of some simple tests for scalar problems, which illustrate the theory above. For scalar fixed point iterations, it only makes sense to consider Anderson for  $m = 1$ , since one can solve explicitly for the optimization parameter that makes the objective function zero, hence  $\theta_k = 0$  at each step. We take  $\beta_k = 1$  in each of these 1D tests. We remark that for the 1D case with  $m = 1$ , Anderson acceleration of the fixed point problem with  $g(x)$  is equivalent to the secant method applied to  $f(x) = g(x) - x$  (this follows from [5] but could also be easily shown by writing out the methods), but still feel it is instructive to show these simple tests.

The fixed point iterations we consider are:

$$\begin{aligned} FPP_1: \quad x_{k+1} &= g_1(x_k) = 1 + \frac{2}{x_k}, & x_0 &= 2.1, \\ FPP_2: \quad x_{k+1} &= g_2(x_k) = x_k - \frac{\cos(x_k) - \sin(x_k)}{-\sin(x_k) - \cos(x_k)}, & x_0 &= 1, \\ FPP_3: \quad x_{k+1} &= g_3(x_k) = x_k^2 - 2, & x_0 &= 4. \end{aligned}$$

Results from these iterations, with ( $m = 1$ ) and without ( $m = 0$ ) Anderson acceleration are shown in Figure 1. For  $FPP_1$  with  $m = 0$  we expect and observe linear convergence with a rate of  $|g'(2)| = 0.5$  to  $x^* = 2$ , but with  $m = 1$  the convergence becomes superlinear. Since  $\theta_k = 0$ , our theory shows that error then depends only on quadratic terms, which is consistent with these results.

$FPP_2$  is the Newton iteration for finding the zero of  $f(x) = \cos(x) - \sin(x)$ , and the fixed point the method converges to is  $x^* = \frac{\pi}{4}$ . Since here the  $m = 0$  test is Newton's method with a smooth  $g$  and good initial guess, the convergence is expected and observed to be quadratic. With  $m = 1$ , we see convergence is slightly worse, which agrees with the theory above: Anderson acceleration adds additional quadratic terms to the residual, which are significant in a quadratically converging iteration.

Lastly in 1D, we consider  $FPP_3$ , which for  $m = 0$  is not expected to converge to  $x^* = 2$  when  $x_0 > 2$  since  $g_3$  is not contractive near the fixed point ( $g'(2) = 4$ ). As expected, with  $m = 0$ , the iteration grows exponentially and by iteration 4 has reached a value of  $10^{10}$ . However, with  $m = 1$  the convergence radius is increased (from 0) to be large enough that the iteration converges even with  $x_0 = 4$ .

### 5.2 Numerical tests for steady incompressible Navier-Stokes equation

Here we present numerical experiments to show the improved convergence provided by Anderson acceleration for solving the steady incompressible Navier-Stokes equations (NSE), which are given

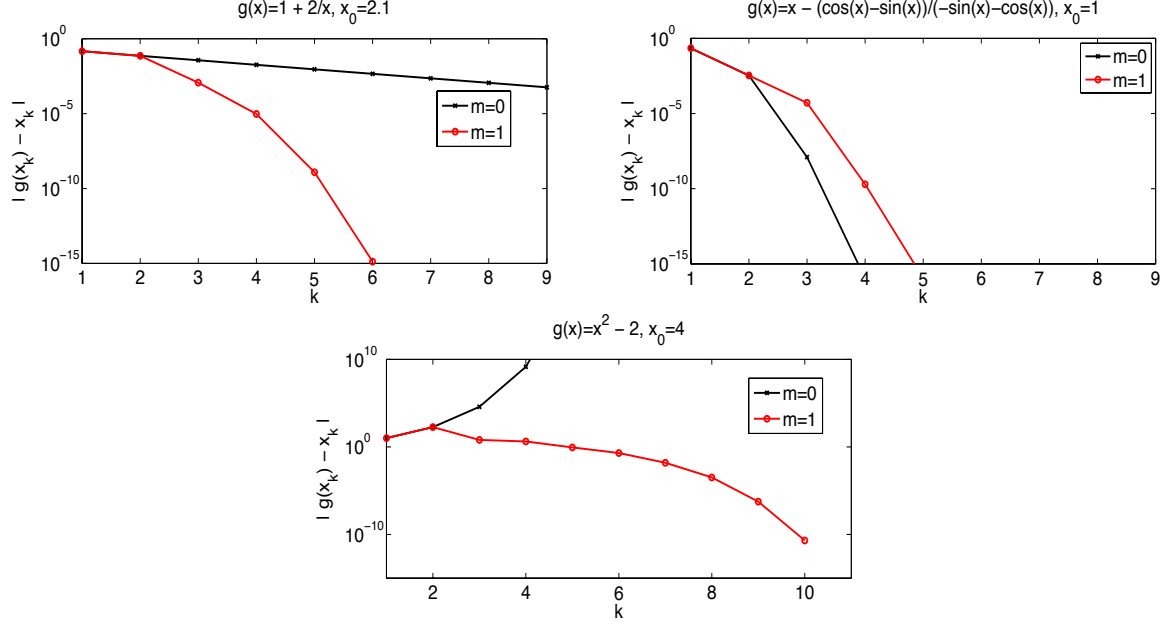


Figure 1: Shown above are the residuals for three scalar fixed point iterations, with and without Anderson acceleration.

in a domain  $\Omega$  by

$$u \cdot \nabla u + \nabla p - \nu \Delta u = f, \quad (5.1)$$

$$\nabla \cdot u = 0, \quad (5.2)$$

where  $\nu$  is the kinematic viscosity,  $f$  is a forcing,  $u$  and  $p$  represent velocity and pressure, and the system must be equipped with appropriate boundary conditions. The  $L^2(\Omega)$  norm and inner product will be denoted by  $\|\cdot\|$  and  $(\cdot, \cdot)$  in this subsection.

The tests we consider are for the 2D lid-driven cavity problem, which uses a domain  $\Omega = (0, 1)^2$ , no slip ( $u = 0$ ) boundary conditions on the sides and bottom, and a ‘moving lid’ on top which is implemented by the Dirichlet boundary condition  $u(x, 1) = \langle 1, 0 \rangle^T$ . There is no forcing ( $f = 0$ ), and the kinematic viscosity is set to be  $\nu := Re^{-1}$ , where  $Re$  is the Reynolds number, and in our tests we use  $Re$  varying between 1000 and 10,000. Plots of the velocity streamlines for the steady NSE at  $Re = 2500$  and 6000 are shown Figure 2.

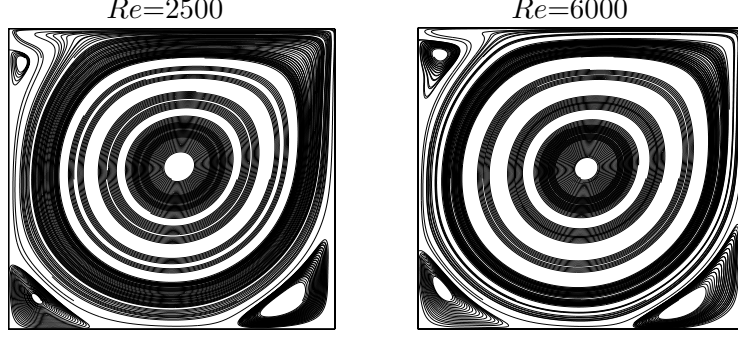


Figure 2: Streamline plots of the steady NSE driven cavity solutions with varying  $Re$ .

We discretize with  $(X_h, Q_h) = ((P_2)^2, P_1)$  Taylor-Hood finite elements on a  $\frac{1}{256}$  uniform triangular mesh that provides 592,387 total degrees of freedom, and for the initial guess we used  $u_h^0 = 0$  but satisfying the boundary conditions. Define the trilinear form  $b^*$  by

$$b^*(u, v, w) := (u \cdot \nabla v, w) + \frac{1}{2}((\nabla \cdot u)v, w).$$

The discrete steady incompressible NSE problem (with skew-symmetrized nonlinear term) reads as follows: Find  $(u, p) \in (X_h, Q_h)$  satisfying for all  $(v, q) \in (X_h, Q_h)$ ,

$$-(p, \nabla \cdot v) + \nu(\nabla u, \nabla v) + b^*(u, u, v) = (f, v), \quad (5.3)$$

$$(\nabla \cdot u, q) = 0. \quad (5.4)$$

Since this problem is nonlinear, we need a nonlinear solver. We consider two common nonlinear iterations, Picard and Newton, which are defined as follows.

**Algorithm 5.1** (Picard iteration for steady NSE).

*Step 1:* Choose  $u_0 \in X_h$ .

*Step k:* Find  $(u_k, p_k) \in (X_h, Q_h)$  satisfying for all  $(v, q) \in (X_h, Q_h)$ ,

$$b^*(u_{k-1}, u_k, v) - (p_k, \nabla \cdot v) + \nu(\nabla u_k, \nabla v) = (f, v), \quad (5.5)$$

$$(\nabla \cdot u_k, q) = 0. \quad (5.6)$$

**Algorithm 5.2** (Newton iteration for steady NSE).

*Step 1:* Choose  $u_0 \in X_h$ .

*Step k:* Find  $(u_k, p_k) \in (X_h, Q_h)$  satisfying for all  $(v, q) \in (X_h, Q_h)$ ,

$$b^*(u_{k-1}, u_k, v) + b^*(u_k, u_{k-1}, v) - b^*(u_{k-1}, u_{k-1}, v) - (p_k, \nabla \cdot v) + \nu(\nabla u_k, \nabla v) = (f, v), \quad (5.7)$$

$$(\nabla \cdot u_k, q) = 0. \quad (5.8)$$

For sufficiently small data, the steady NSE and these iterations are well-posed [9]. Hence we can consider both the Picard and Newton iterations as fixed point iterations  $u_{k+1} = g(u_k)$ , where  $g$  is a solution operator of (5.5)-(5.6) for Picard or (5.7)-(5.8) for Newton. In this way, we can apply Anderson acceleration to both methods. Below, we test both the Picard and Newton iterations

with Anderson acceleration (but note that we apply only the basic Picard and Newton methods, i.e. without relaxation or other variation that can aid in convergence). The linear systems are solved with a sparse direct solver.

For Picard iterations, we observe in Figure 3 (left side) that Picard without acceleration is converging linearly, although slowly; after 40 iterations, the residual is still  $O(10^{-4})$ . Anderson acceleration makes a very significant improvement in the Picard convergence, with big improvement offered by  $m = 1$  and  $m = 2$ , and even more by  $m = 3$ . With  $m = 3$  the residual after 40 iterations is about  $O(10^{-9})$ , and it would take usual Picard about another 50 iterations to reach this level for its residual.

On the right side of Figure 3, we display the convergence behavior of the Newton iterations. We observe the usual Newton iteration diverges, but with Anderson acceleration it converges for each of  $m = 1, 2, 3$ . This is an example of Anderson acceleration increasing the radius of convergence of a fixed point iteration. The  $m = 1$  Anderson accelerated Newton iteration with  $m = 1$  achieves a residual of  $10^{-14}$  after just 13 iterations. It is important to note that such an improvement with small  $m$  is also observed in [11].

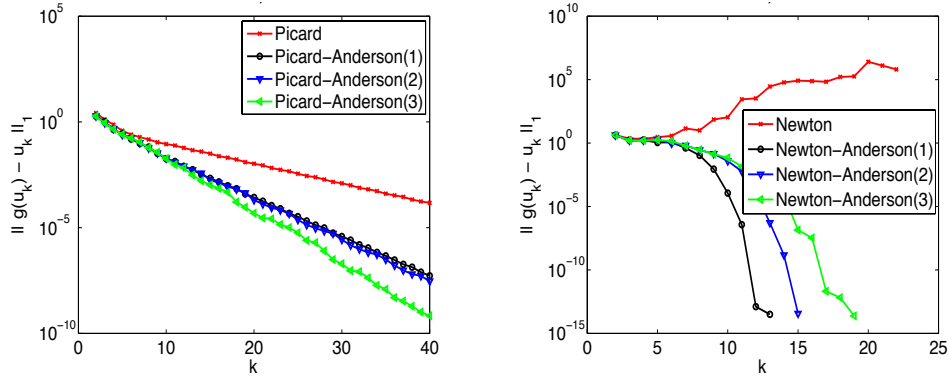


Figure 3: Convergence of the Anderson accelerated Picard and Newton iterations with  $Re = 2500$ .

Results for  $Re = 6000$  are shown in Figure 4. The usual Picard iteration fails here, as the residual over the last 20 iterations grows (although slightly), so  $\kappa > 1$  in this case. Anderson acceleration helps Picard significantly, and will allow for convergence. Usual Newton and  $m = 1, 2$  Anderson-accelerated Newton iterations all failed (diverged), and we do not show these results in the plot. The Anderson-accelerated Newton iteration with  $m = 3$  converged, and quite rapidly, reaching a residual of  $O(10^{-14})$  in just 23 iterations.

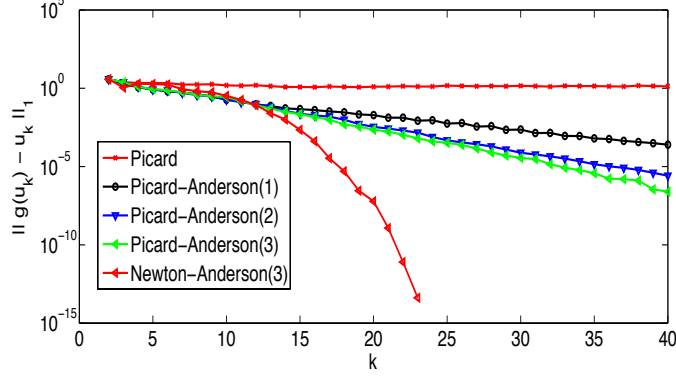


Figure 4: Convergence of the Anderson accelerated Picard and Newton iterations with  $Re = 6000$ .

Box-plots of the  $\theta_k$ 's from the Picard iterations are shown in figure 5. For  $Re = 2500$  (left side), there is a clear decreasing trend in distribution of  $\theta$ 's as  $m$  increases, while for  $Re = 6000$  the boxplots look rather similar but with  $m = 1$  seemingly a little lower overall compared to  $m = 2$ . However, the lower values and outliers in these plots are critical, since one multiplication of a small factor takes many multiplications of larger factors to achieve the same residual decrease.

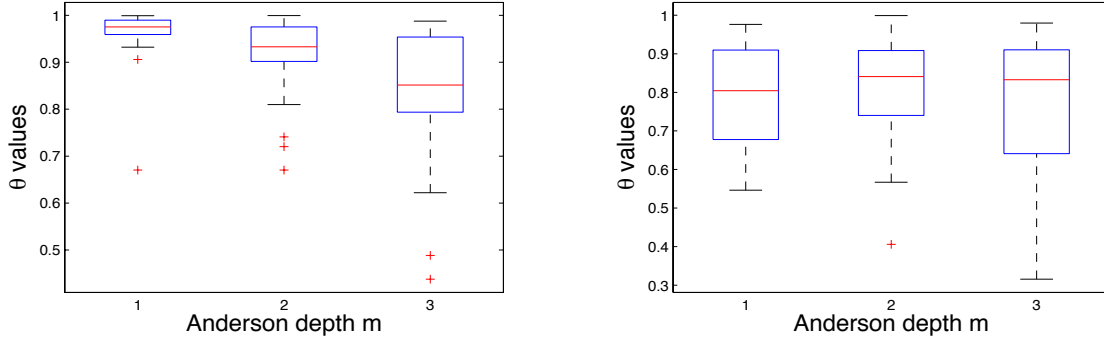


Figure 5: Box-plots of  $\theta$  values for the Picard iterations with  $Re = 2500$  (left) and  $Re = 6000$  (right).

As a final part of this test, we compare the number of iterations needed to converge the residual for the Picard iteration in the  $H^1$  norm to a tolerance of  $10^{-8}$ , for varying  $Re$  and varying  $m$ . Results are shown in Table 1, and again we observe a dramatic improvement from Anderson acceleration. Even  $m = 1$  is enough to provide convergence up to  $Re = 10000$ , although additional gain is made by increasing to  $m = 2$  and to  $m = 3$ . It is interesting that convergence of the steady NSE is achieved for  $Re = 9000$  and  $10000$  since the bifurcation point transition to transient flow is around  $8000$  [3], and thus the solutions found are seemingly unstable steady NSE solutions. F denotes

failure in the table, which we define as not converging within 500 iterations (but we note that inspection of the last few iterations of each of these that failed indicates the iterations are nowhere near, or even approaching, convergence).

Re / m	0	1	2	3
1000	36	32	29	26
2000	48	41	40	34
3000	86	49	45	37
4000	158	59	46	40
5000	363	55	48	44
6000	F	62	55	49
7000	F	65	61	53
8000	F	78	70	58
9000	F	94	83	68
10000	F	105	97	71

Table 1: Shown above are the number of Picard iterations needed to converge the nonlinear residual for the steady NSE up to  $10^{-8}$  in the  $H^1$  norm, for varying  $Re$  and  $m$ . F denotes a failure to reach convergence by 500 iterations.

### 5.3 Damping tests with a quasilinear equation

The damping parameter  $\beta$  of Algorithm 2.1 may become important for convergence in the case of a fixed-point operator  $g$  that is not contractive. A simple example of this type of problem is the quasilinear equation  $-\operatorname{div}(a(u)\nabla u) = f$  in a domain  $\Omega$  with homogeneous Dirichlet boundary conditions. In weak form

$$(a(u)\nabla u, \nabla v) = (f, v), \quad (5.9)$$

where  $(\cdot, \cdot)$  denotes the  $L^2$  inner product as in the above example. This can be thought of as a simple model of the effective nonlinearity in a steady Richards' type equation modeling the pressure  $u$  in partially saturated media, where  $a(u)$  is the hydraulic conductivity which depends nonlinearly on the pressure head via the saturation. In this example we take  $\Omega = (0, 1)$  and

$$a(u) = k + \tanh((u - u_0)/\varepsilon), \quad \text{with } u_0 = 0.5, \quad k = 1.01, \quad \text{and } \varepsilon = 0.1.$$

The function  $f$  is chosen so the exact solution is  $u^* = 10\sin(\pi x)$ . For the results below, the 1D problem is discretized with piecewise linear (P1) finite elements with a uniform meshsize of  $h = 1/16384$ . For this example  $m_k = 0$  for  $k < m$  and  $m$  otherwise. The optimization problem is solved with an economy  $QR$  decomposition and  $\theta_k$  is computed as described in §3.3. The fixed-point operator  $\tilde{u}^{k+1} = g(u^k)$  solves  $(a(u^k)\nabla g(u^k), \nabla v) = (f, v)$ , as in a basic Picard iteration.

As seen by the expansion (3.13), the results of Theorems 4.1 and 4.2 as well as Proposition 4.3, the damping factor  $\beta_{k-1}$  affects the first order term  $\theta_k(1 - \beta_{k-1} + \kappa\beta_{k-1})\|w_k\|$ , but not the higher order terms. If the operator  $g : X \rightarrow X$  is not assumed contractive, then Assumption 3.2 does not hold, and (3.13) then provides a blueprint for bounding  $\|w_{k+1}\|$  by  $\|w_k\|$  and higher-order terms

involving differences of consecutive iterates

$$\|w_{k+1}\| \leq \theta_k(1 - \beta_{k-1} + \kappa\beta_{k-1})\|w_k\| + \mathcal{O}(\|e_k\|^2) + \dots + \mathcal{O}(\|e_{k-m}\|^2),$$

as the bounds of §3.2 controlling the difference between consecutive iterates by the residuals do not hold in the noncontractive setting. It is remarked however that in the contractive setting of §3.2, the control of the error terms  $\|e_j\|$  in terms of residuals  $\|w_n\|$  is independent of the damping.

This first order effect of the damping agrees with that seen for the error in the fixed-point iteration alone. If the update step of the damped fixed-point iteration for operator  $g$  with fixed-point  $x^*$  is given by  $u_{k+1} = (1 - \beta)u_k + \beta g(u_k)$  then

$$u_{k+1} - u^* = (1 - \beta)(u_k - u^*) + \beta(g(u_k) - g(u^*)) = (1 - \beta)(u_k - u^*) + \beta g'(z_k^*(t); u_k - u^*),$$

with  $z_k^*(t) = u^* + t(u_k - u^*)$ . As this example is easily seen to satisfy Assumption 3.1, this immediately yields the norm-bound  $\|u_{k+1} - u^*\| \leq ((1 - \beta) + \kappa\beta)\|u_k - u^*\|$ .

If  $g$  is contractive then  $\beta = 1$  (no damping) gives the best convergence rate. In this example however, Assumption 3.2 does not hold globally. For instance near the boundary  $u$  approaches zero and  $k + (\tanh(u - u_0)/\varepsilon)$  is close to  $k - 1 = 10^{-2}$ , and the locally small ellipticity coefficient can cause failure of the method to converge. This is demonstrated in the first plot of Figure 6 where on the left the fixed-point iteration fails to converge to a tolerance of  $10^{-5}$  with  $\beta = \{1, 0.8, 0.6\}$ , although more accuracy is attained with the damped iterations. It is also clear from this first plot that in the regime where the operator  $g$  is contractive (the beginning of the calculation), the damping has the predicted linear effect on the convergence rate.

The second and third plots of Figure 6 show the effect of damping factors  $\beta = \{1.0, 0.8, 0.6\}$  as well as an adaptive strategy for the cases of Anderson depths  $m = 1$  and  $m = 2$ . The adaptive strategy is based on the convergence rates found in Theorems 4.1 - 4.2 and Proposition 4.3, meaning  $\beta$  plays an active role in decreasing the coefficient of the first order term particularly when  $\theta$  is not small enough. So  $\beta_{adapt} = 1 - \theta_k/2$  is chosen as a simple heuristic to set a factor between 0.5 and 1.0 that is close to unity when  $\theta_k$  is small and approaches 0.5 as  $\theta$  approaches one.

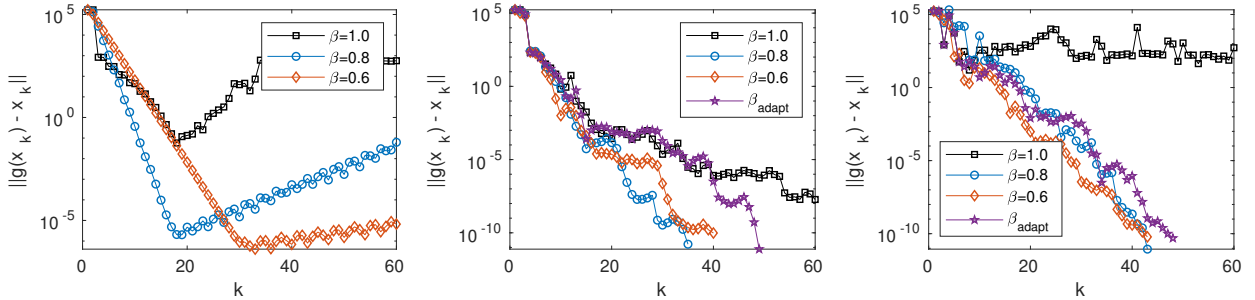


Figure 6: Left: Damped iterations for (5.9) with  $m = 0$ . Center: Damped iterations with  $m = 1$ . Right: Damped iteration with  $m = 2$ .

The three plots of Figure 7 illustrate the behavior of the same damping factors for greater Anderson depths,  $m = \{4, 6, 8\}$ . While the three examples in Figure 6 failed to converge without damping, the three examples for greater depth  $m$  in Figure 7 converged both with and without, but

generally better with some damping. The adaptive strategy, while not optimal, demonstrates proof of concept that with the gain  $\theta_k$  taken into account, damping can be designed without extensive experimentation or additional computation to stabilize the convergence for difficult problems.

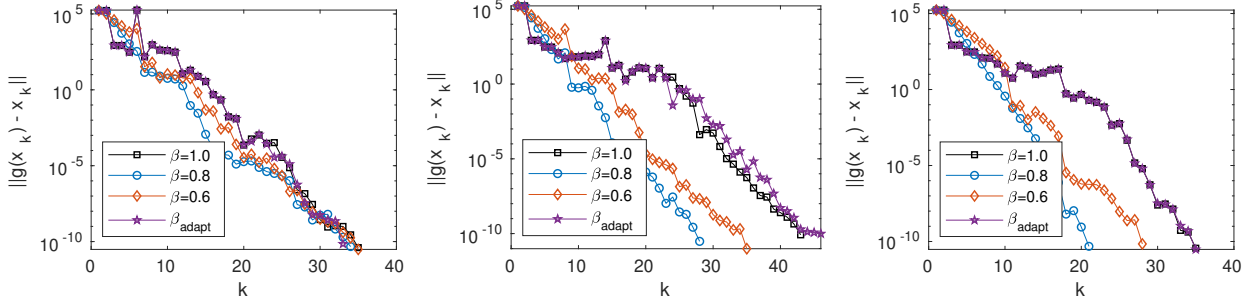


Figure 7: Left: Damped iterations for (5.9) with  $m = 4$ . Center: Damped iterations with  $m = 6$ . Right: Damped iteration with  $m = 8$ .

## 6 Conclusion

We have proven that Anderson acceleration improves the first-order convergence rate for fixed point iterations, in agreement with decades of experimental results. We show that the increase in the linear convergence rate at each step depends on the gain from the optimization step, but that additional quadratic error terms arise. Hence as long as the gain from the optimization stage dominates these quadratic error terms, the convergence rate will be increased. In particular for linearly convergent fixed point methods, an improved convergence rate from Anderson acceleration is expected; however, for methods converging quadratically, the convergence will typically be slightly slowed. Additionally, our results provide justification that both increasing the depth  $m$  and using damping increases the radius of convergence. Results of numerical tests have been provided to illustrate our theory.

## References

- [1] H. An, X. Jia, and H.F. Walker. Anderson acceleration and application to the three-temperature energy equations. *Journal of Computational Physics*, 347:1–19, 2017.
- [2] D. G. Anderson. Iterative procedures for nonlinear integral equations. *J. Assoc. Comput. Mach.*, 12(4):547–560, 1965.
- [3] F. Auteri, N. Parolini, and L. Quartapelle. Numerical investigation on the stability of singular driven cavity flow. *J. Comput. Phys.*, 183(1):1–25, 2002.
- [4] C. Brezinski. Convergence acceleration during the 20th century. *J. Comput. Appl. Math*, 122:1–21, 2000.
- [5] H. Fang and Y. Saad. Two classes of multisecant methods for nonlinear acceleration. *Numer. Linear Algebra Appl.*, 16(3):197–221, 2009.

- [6] M. Geist and B. Scherrer. Anderson acceleration for reinforcement learning. *Submitted*, 2018.
- [7] N. Higham and N. Strabic. Anderson acceleration of the alternating projections method for computing the nearest correlation matrix. *Numerical Algorithms*, 72:1021–1042, 2016.
- [8] C.T. Kelley. Numerical methods for nonlinear equations. *Acta Numerica*, 27:207–287, 2018.
- [9] W. Layton. *An Introduction to the Numerical Analysis of Viscous Incompressible Flows*. SIAM, Philadelphia, 2008.
- [10] J. Loffeld and C. Woodward. Considerations on the implementation and use of Anderson acceleration on distributed memory and GPU-based parallel computers. *Advances in the Mathematical Sciences*, pages 417–436, 2016.
- [11] P. A. Lott, H. F. Walker, C. S. Woodward, and U. M. Yang. An accelerated Picard method for nonlinear systems related to variably saturated flow. *Adv. Water Resour.*, 38:92–101, 2012.
- [12] Y. Peng, B. Deng, J. Zhang, F. Geng, W. Qin, and L. Liu. Anderson acceleration for geometry optimization and physics simulation. *Submitted*, 2018.
- [13] S. Pollock, L. Rebholz, and M. Xiao. Anderson-accelerated convergence of picard iterations for incompressible Navier-Stokes equations. *SIAM J. Numer. Anal.*, 2019. Accepted.
- [14] P. Stasiak and M.W. Matsen. Efficiency of pseudo-spectral algorithms with anderson mixing for the SCFT of periodic block-copolymer phases. *Eur. Phys. J. E*, 34:110:1–9, 2011.
- [15] A. Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM J. Numer. Anal.*, 53(2):805–819, 2015.
- [16] A. Toth, C.T. Kelley, S. Slattery, S. Hamilton, K. Clarno, and R. Pawlowski. Analysis of Anderson acceleration on a simplified neutronics/thermal hydraulics system. *Proceedings of the ANS MC2015 Joint International Conference on Mathematics and Computation (M&C), Supercomputing in Nuclear Applications (SNA) and the Monte Carlo (MC) Method*, ANS MC2015 CD:1–12, 2015.
- [17] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49(4):1715–1735, 2011.