CONVERGENCE ANALYSIS OF THE ALTERNATING ANDERSON-PICARD METHOD FOR NONLINEAR FIXED-POINT PROBLEMS*

XUE FENG†, M. PAUL LAIU‡, AND THOMAS STROHMER†

Abstract. Anderson Acceleration (AA) has been widely used to solve nonlinear fixed-point problems due to its rapid convergence. This work focuses on a variant of AA in which multiple Picard iterations are performed between each AA step, referred to as the Alternating Anderson-Picard (AAP) method. Despite introducing more 'slow' Picard iterations, this method has been shown to be efficient and even more robust in both linear and nonlinear cases. However, there is a lack of theoretical analysis for AAP in the nonlinear case, which this paper aims to address. We show the equivalence between AAP and a multisecant-GMRES method that uses GMRES to solve a multisecant linear system at each iteration. More interestingly, the incorporation of Picard iterations and AA establishes a deep connection between AAP and the Newton-GMRES method. This connection is evident in terms of the multisecant matrix, the approximate Jacobian inverse, search direction, and optimization gain—an essential factor in the convergence analysis of AA. We show that these terms converge to their corresponding terms in the Newton-GMRES method as the residual approaches zero. Consequently, we build the convergence analysis of AAP. To validate our theoretical findings, numerical examples are provided.

Key words. Anderson acceleration, alternating Anderson-Picard (AAP) method, Newton-GMRES, convergence

MSC codes. 68Q25, 68R10, 68U05

1. Introduction.

Fixed-point iterations are one of the cornerstones in scientific computing, formulated by

$$\boldsymbol{x} = g(\boldsymbol{x}), \tag{1.1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}^d$ is assumed to be a continuously differentiable operator in this work. The standard method for solving (1.1) is the fixed-point iteration, also known as the Picard iteration, noted for its simplicity but also for its potentially slow convergence. This paper concerns an acceleration method for nonlinear fixed-point iterations.

Anderson acceleration (AA), initially introduced in [2] to address partial differential equations, is one of the most popular acceleration schemes. AA is a scheme mixing history points with mixing coefficients computed by solving a least-squares (LS) problem. It can be regarded as a multisecant quasi-Newton method whose approximate Jacobian inverse satisfies a multisecant equation [13]. AA is known for

*Submitted to the editors DATE: July 14, 2024.

Funding: This work was supported, in part, by the Office of Advanced Scientific Computing Research and performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC for the US Department of Energy under Contract No. DE-AC05-00OR22725. T.S. and X.F. acknowledge support from NSF DMS-2208356, NIH R01HL16351. T.S., X.F. and P.L. acknowledge support from DE-AC05-00OR22725.

†University of California, Davis, Davis, CA (xffeng@ucdavis.edu, strohmer@math.ucdavis.edu).

‡Oak Ridge National Laboratory, Oak Ridge, TN (laiump@ornl.gov).

its simplicity and significantly improved convergence performance. It has wide applications in fields such as electronic structure calculations [3, 13], fluid dynamics [22], geometrical optimization [27], and more recently machine learning [15, 38].

The classical AA method takes an AA step in each iteration. There are different variants of AA such as restarted AA [26], type-I AA [13, 39], and EDIIS [7]. Among these alternatives, a simple method that applies AA at periodic intervals rather than every iteration draws our attention. Contrary to the belief that a reduced frequency of AA would degrade performance, this technique has been proven to enhance both the efficiency and robustness of AA. This method was first introduced to address large-scale linear systems within the classical Jacobi fixed-point iteration framework for electronic structure calculations, termed the Alternating Anderson-Jacobi (AAJ) method, in[31], where it has been shown to significantly outperform both the GMRES and Anderson-accelerated Jacobi methods. Subsequently, AAJ was generalized to include preconditioning, known as the Alternating Anderson-Richardson (AAR) method; it outperforms classical preconditioned Krylov solvers in terms of efficiency and scalability [35]. In nonlinear scenarios, [3] uses this strategy for accelerating self-consistent field iterations, and [18] applied a similar strategy on Riemannian optimization. On the theoretical side, [23] provides a convergence analysis of AAR, establishing its equivalence to GMRES. However, no theoretical convergence analysis exists yet for the application of this alternating approach to solve nonlinear fixed-point problems, which is the primary focus of this work.

We consider a specific variant of this periodically alternating method, referred to as AAP, which takes m Picard iterations between each AA step, and takes an AA step to mix all $m+1$ points. We refer to the method AAP(m) in this paper when the number of Picard iterations m needs to be specified. A detailed description of AAP is given in Algorithm 2.1. This method takes advantage of the efficiency of AA to improve the convergence, while maintaining the simplicity of the Picard iterations.

Solving (1.1) is equivalent to finding the root of the residual function:

$$0 = f(\boldsymbol{x}) := g(\boldsymbol{x}) - \boldsymbol{x}. \tag{1.2}$$

Newton-type methods are fundamental in this context due to their characteristic fast convergence, achieved by utilizing Jacobian information [24]. One such method, Newton-GMRES [8, 1], iteratively updates the solution by solving the root of the linear approximation of f using GMRES [20]. In this work, we will study the properties of the AAP method. The core is to establish the equivalence between AAP and a multisecant-GMRES method, which iteratively uses GMRES to find the root of a multisecant linear approximation of f. In addition, compared to other variants of AA, the characteristic that all history points used in each AA step are generated by Picard iterations allows AAP capture local information more effectively. Formally, we will show that AAP has a deep connection to the Newton-GMRES method.

This paper is organized as follows. Section 2 details the AAP algorithm and related work. Section 3 demonstrates that, at each global iteration t, AAP(m) implicitly solves a multisecant linear system using the GMRES(m) method, followed by one additional linearized Picard iteration. Section 4 explores the connection between AAP and the Newton-GMRES method, focusing on the multisecant matrix, the approximate Jacobian inverse, search direction, and optimization gain – an essential factor in the convergence analysis of AA. We show that as the residual approaches zero, these terms converge to their corresponding terms in the Newton-GMRES method. Section 5 presents the convergence analysis of AAP, including a one-step residual bound

and the local convergence result. Numerical examples are provided in Section 6. Appendix A discusses an assumption for guaranteeing the convergence of AAP.

1.1. Notation. The vector norm is denoted by $\|\cdot\|$ and refers to the Euclidean norm (2-norm). The matrix norm is also denoted by $\|\cdot\|$ and refers to the induced 2-norm, which is the largest singular value. A sequence of matrices $\boldsymbol{A}_k \to \boldsymbol{A}$ means $\|\boldsymbol{A}_k - \boldsymbol{A}\| \to 0$. In this paper, a sequence of matrices $\boldsymbol{A}_k \to \widetilde{\boldsymbol{A}}_k$, where $\widetilde{\boldsymbol{A}}_k$ is another sequence of matrices, means $\|\boldsymbol{A}_k - \widetilde{\boldsymbol{A}}_k\| \to 0$; we refer to this as the convergence of $\boldsymbol{A}_k$ to $\widetilde{\boldsymbol{A}}_k$. The Frobenius norm of the matrix is denoted as $\|\cdot\|_F$. The identity matrix is denoted by $\boldsymbol{I}$. For better presentation in this paper, $\boldsymbol{A}^{\ell}$ represents a superscript on matrix $\boldsymbol{A}$; $\boldsymbol{A}^{(\ell)}$ refers to the matrix $\boldsymbol{A}$ raised to the power ℓ. $\boldsymbol{A}^{-1}$ denotes the inverse of matrix $\boldsymbol{A}$. $\boldsymbol{A}^{\dagger}$ denotes the left pseudoinverse of matrix $\boldsymbol{A}$. The ith largest singular value of $\boldsymbol{A}$ is denoted as $\sigma_i(\boldsymbol{A})$, and $\sigma_{\min}(\boldsymbol{A})$ is the smallest singular value of $\boldsymbol{A}$. The condition number of $\boldsymbol{A}$ is denoted as $\mathrm{cond}(\boldsymbol{A})$. The subspace spanned by the columns of matrix $\boldsymbol{A}$ is denoted as $\mathrm{span}(\boldsymbol{A})$.

For a given matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, the n-th Krylov subspace is defined as:

$$\mathcal{K}_n(\boldsymbol{A}, \boldsymbol{b}) = \mathrm{span}\{\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \boldsymbol{A}^2\boldsymbol{b}, \ldots, \boldsymbol{A}^{n-1}\boldsymbol{b}\}.$$

Krylov subspaces are shift invariant, i.e., for any $\nu \in \mathbb{R}$, $\mathcal{K}_n(\boldsymbol{A} - \nu\boldsymbol{I}, \boldsymbol{b}) = \mathcal{K}_n(\boldsymbol{A}, \boldsymbol{b})$. The Generalized Minimal Residual Method (GMRES) is one of the most popular Krylov subspace methods and solves the linear systems $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual within a Krylov subspace at iteration n:

$$\min_{\boldsymbol{x} \in \mathcal{K}_n(\boldsymbol{A}, \boldsymbol{b})} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|.$$

The following assumptions will be used in this paper:

ASSUMPTION 1.1. *The following properties are assumed for g:*
(a). $\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \le \kappa\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall\ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ with $\kappa \le 1$.
(b). The Jacobian of g, denoted as g', is Lipschitz continuous with constant $\gamma > 0$,

$$\|g'(\boldsymbol{x}) - g'(\boldsymbol{y})\| \le \gamma\|\boldsymbol{x} - \boldsymbol{y}\|.$$

2. Algorithm and related work. This section provides the details of the AAP algorithm, along with the necessary preliminary concepts and assumptions. In addition, existing work on the convergence analysis of AA is discussed.

2.1. The AAP algorithm. At global iteration t, AAP first takes m Picard iterations from current iterate $\boldsymbol{x}_t$ to generate new points $\{\boldsymbol{x}_t^1, \cdots, \boldsymbol{x}_t^m\}$ and compute their residuals. AAP next computes the mixing coefficients by solving a constrained LS problem based on the residuals of the history points $\{\boldsymbol{x}_t, \boldsymbol{x}_t^1, \cdots, \boldsymbol{x}_t^m\}$, and then updates the iterate to $\boldsymbol{x}_{t+1}$ by mixing the fixed-point values of these points. See Algorithm 2.1 for details. We denote the mixing coefficients at global iteration t as $\boldsymbol{\alpha}_t = \{\alpha_t^0, \alpha_t^1, ..., \alpha_t^m\} \in \mathbb{R}^{m+1}$, and denote the residual $f(\boldsymbol{x}_t)$ as either $\boldsymbol{f}_t$ or $\boldsymbol{f}_t^0$, depending on the context. When relevant, we refer to this algorithm as AAP(m), in which $m+1$ Picard iterations are performed in each global iteration.

Here the damping ratio β_t balances the weighted average of history points and the weighted average of history residuals, i.e.,

$$\boldsymbol{x}_{t+1} = (1-\beta_t)\sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{x}_t^{\ell} + \beta_t \sum_{\ell=0}^{m} \alpha_t^{\ell} g(\boldsymbol{x}_t^{\ell}) = \sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{x}_t^{\ell} + \beta_t \sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{f}_t^{\ell}. \tag{2.3}$$

Algorithm 2.1 Alternating Anderson-Picard (AAP) Method

Initialization: $\boldsymbol{x}_0 \in \mathbb{R}^d$, *number of Picard steps* $m \leq d$, *damping ratio* $\beta_t \geq 0$
for *global iteration* $t = 0$ **to** T **do**
    `/* Step 1: takes Picard iteration to generate` m `points and compute their residuals */`
    $\boldsymbol{x}_t^0 \leftarrow \boldsymbol{x}_t$
    for $\ell = 1$ **to** m **do**
        $\boldsymbol{x}_t^\ell \leftarrow g(\boldsymbol{x}_t^{\ell-1})$
        $\boldsymbol{f}_t^{\ell-1} \leftarrow \boldsymbol{x}_t^\ell - \boldsymbol{x}_t^{\ell-1}$
    end for
    $\boldsymbol{f}_t^m \leftarrow g(\boldsymbol{x}_t^m) - \boldsymbol{x}_t^m$
    `/* Step 2: Take one Anderson Acceleration step */`
    Solve

$$\boldsymbol{\alpha}_t \leftarrow \underset{\alpha \in \mathbb{R}^{m+1}}{\arg\min} \big\| \sum_{\ell=0}^{m} \alpha^\ell \boldsymbol{f}_t^\ell \big\|^2 \quad \text{s.t.} \quad \sum_{\ell=0}^{m} \alpha^\ell = 1 \tag{2.1}$$

    Update

$$\boldsymbol{x}_{t+1} \leftarrow (1-\beta_t) \sum_{\ell=0}^{m} \alpha_t^\ell \boldsymbol{x}_t^\ell + \beta_t \sum_{\ell=0}^{m} \alpha_t^\ell g(\boldsymbol{x}_t^\ell) \tag{2.2}$$

end for

Let $\boldsymbol{S}_t := [\boldsymbol{s}_t^0, \ldots, \boldsymbol{s}_t^{m-1}] \in \mathbb{R}^{d\times m}$ and $\boldsymbol{Y}_t := [\boldsymbol{y}_t^0, \ldots, \boldsymbol{y}_t^{m-1}] \in \mathbb{R}^{d\times m}$ with $\boldsymbol{s}_t^\ell = \boldsymbol{x}_t^{\ell+1} - \boldsymbol{x}_t^\ell$ and $\boldsymbol{y}_t^\ell = \boldsymbol{f}_t^{\ell+1} - \boldsymbol{f}_t^\ell$, for $\ell = 0, \ldots, m-1$. When $\boldsymbol{Y}_t$ is of full column rank, the constrained LS problem (2.1) can be written as an unconstrained LS problem, i.e.,

$$\boldsymbol{z}_t := \underset{\boldsymbol{z}\in\mathbb{R}^m}{\arg\min} \|\boldsymbol{Y}_t \boldsymbol{z} - \boldsymbol{f}_t\|^2 (= \boldsymbol{Y}_t^\dagger \boldsymbol{f}_t), \tag{2.4}$$

and $\boldsymbol{\alpha}_t$ can be recovered from $\boldsymbol{z}_t$ by setting $\alpha_t^0 = 1 + z_t^0$, $\alpha_t^\ell = z_t^\ell - z_t^{\ell-1}$ for $0 < \ell < m$, and $\alpha_t^m = -z_t^{m-1}$. This transformation leads to an equivalent multisecant quasi-Newton formulation of AAP, which updates $\boldsymbol{x}_{t+1}$ based on an approximate Jacobian inverse $\boldsymbol{H}_t$. Specifically,

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \boldsymbol{H}_t \boldsymbol{f}_t, \tag{2.5}$$

$$\text{where} \qquad \boldsymbol{H}_t \leftarrow -\beta_t \boldsymbol{I} + (\boldsymbol{S}_t + \beta_t \boldsymbol{Y}_t)(\boldsymbol{Y}_t^T \boldsymbol{Y}_t)^{-1} \boldsymbol{Y}_t^T. \tag{2.6}$$

It is straightforward to verify that $\boldsymbol{H}_t$ satisfies the inverse multisecant equation $\boldsymbol{H}_t \boldsymbol{Y}_t = \boldsymbol{S}_t$.

The following assumptions regarding the iterations generated by AAP are important for this paper.

ASSUMPTION 2.1. *For all* $t \geq 0$, *both* $\boldsymbol{S}_t$ *and* $\boldsymbol{Y}_t$ *have full column rank.*

Assumption 2.1 is fundamental for all the results presented in this paper and will be assumed without further mention. This assumption is reasonable as we can adaptively adjust m in each global iteration and stop the Picard iterations when the columns of $\boldsymbol{S}_t$ or $\boldsymbol{Y}_t$ become linearly independent. Alternatively, we can remove the

linearly dependent columns directly, though this approach may affect the analysis presented in this paper.

ASSUMPTION 2.2 (Uniform Boundedness of $\mathrm{cond}(\boldsymbol{S}_t)$). *There exist constants* $T > 0$ *and* $M > 0$ *such that* $\mathrm{cond}(\boldsymbol{S}_t) \leq M$ *for all* $t > T$.

Assumption 2.2 is a general assumption in the convergence analysis of AA (see Section 2.3). We will show that $\boldsymbol{S}_t$ converges to a Krylov matrix in Subsection 4.2, and discuss this assumption in more detail in Appendix A.

2.2. Multisecant matrices. In this paper, we define the multisecant matrix $\boldsymbol{B}_t$ as a matrix that satisfies the multisecant equation $\boldsymbol{B}_t\boldsymbol{S}_t = \boldsymbol{Y}_t$. It follows from $\boldsymbol{s}_t^\ell = \boldsymbol{x}_t^{\ell+1} - \boldsymbol{x}_t^\ell = \boldsymbol{f}_t^\ell$ that

$$\boldsymbol{B}_t\boldsymbol{S}_t = \boldsymbol{Y}_t, \quad \Longleftrightarrow \quad \boldsymbol{B}_t\boldsymbol{f}_t^\ell = \boldsymbol{f}_t^{\ell+1} - \boldsymbol{f}_t^\ell, \quad \forall\ 0 \leq \ell \leq m-1. \tag{2.7}$$

Furthermore, we define the set

$$\mathbb{B}_t := \{\boldsymbol{B} \in \mathbb{R}^{d\times d} \mid \boldsymbol{B}\boldsymbol{S}_t = \boldsymbol{Y}_t\}. \tag{2.8}$$

It is evident that $\mathbb{B}_t$ is non-empty and contains invertible matrices under Assumption 2.1. For instance, elements of $\mathbb{B}_t$ include $\boldsymbol{Y}_t\boldsymbol{S}_t^\dagger$ and $-\boldsymbol{I} + (\boldsymbol{Y}_t + \boldsymbol{S}_t)\boldsymbol{S}_t^\dagger$. Moreover, an invertible matrix in $\mathbb{B}_t$ can be constructed by taking $\boldsymbol{Y}\boldsymbol{S}^{-1}$, where $\boldsymbol{Y} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ are arbitrary invertible matrices with the first m columns being $\boldsymbol{Y}_t$ and $\boldsymbol{S}_t$, respectively. For any $\boldsymbol{B}_t \in \mathbb{B}_t$, we call $\widehat{f}_t(\boldsymbol{x}) = \boldsymbol{B}_t(\boldsymbol{x} - \boldsymbol{x}_t) + f(\boldsymbol{x}_t)$ a multisecant linear approximation of f at $\boldsymbol{x}_t$, which satisfies $\widehat{f}_t(\boldsymbol{x}_t^\ell) = f(\boldsymbol{x}_t^\ell)$ for $0 \leq \ell \leq m$. At global iteration t, this approximation is based on the m Picard iteration points generated in AAP, as seen in the definition of $\mathbb{B}_t$.

2.3. Existing convergence analysis of Anderson acceleration. The local convergence rate of AA is established in [36, 19]. Subsequently, [28, 12, 29] demonstrated how AA improves the convergence rate over fixed-point (Picard) iterations. The state-of-the-art convergence bound consists of a linear term and a higher-order term. A simplified result from [29, Theorem 5.1] is given by $\|f(\boldsymbol{x}_{t+1})\| \leq \kappa\theta_t\|f(\boldsymbol{x}_t)\| + c\sqrt{1-\theta_t^2}\|f(\boldsymbol{x}_t)\|\sum_{\ell=0}^m \|f(\boldsymbol{x}_{t-\ell})\|$ where $\theta_t \leq 1$ is called the *optimization gain*. This result indicates that the local convergence rate of AA, $\kappa\theta_t$, is superior to the local convergence rate of the Picard iteration, κ.

The assumptions to guarantee the convergence of AA include a smoothness condition on g and a uniform boundedness assumption on the coefficients $\boldsymbol{\alpha}$ [36, 12, 29, 25]. Alternatives to the uniform boundedness assumption on $\boldsymbol{\alpha}$ is to assume sufficient linear independence among the columns of $\boldsymbol{Y}_t$ [29] and the boundedness of the condition number of $\boldsymbol{S}_t$ or $\boldsymbol{Y}_t$ [25]. The equivalence between these conditions is discussed in [25]. These conditions are also directly related to the conditioning of the constrained LS problem in AA. In [30], a filtering strategy is proposed to enforce these conditions by removing nearly linearly dependent columns from the LS problem.

An important property of AA is its equivalence to GMRES on linear problems. When $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, [37] shows that AA with no truncation solving $\boldsymbol{x} = g(\boldsymbol{x})$ is "essentially equivalent" to GMRES applied to $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{b}$, in the sense that the weighted sum of the history points in each AA step equals the GMRES iterate $\boldsymbol{x}_t^{\mathrm{GMRES}}$ and the AA iterate $\boldsymbol{x}_{t+1}^{\mathrm{AA}} = g\left(\boldsymbol{x}_t^{\mathrm{GMRES}}\right)$. It is also mentioned in [37] that a "restarted" variant of AA, in which the method proceeds without truncation for m steps and then is restarted, is equivalent to GMRES(m) applied to $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{b}$ followed by a fixed-point iteration. In [23], the equivalence between Alternating Anderson-Richardson and GMRES is established. For nonlinear problems, AA is closely related

to the Nonlinear Generalized Minimal Residual Method (NGMRES) [33]. This work will establish the equivalence bettwen AAP and a multisecant-GMRES method for nonlinear case.

3. Equivalence between AAP and multisecant-GMRES. This section gives the equivalence between AAP and a multisecant-GMRES method. We show that, at global iteration t, AAP implicitly solves a multisecant linear system, followed by one additional linearized Picard iteration.

THEOREM 3.1 (Equivalence between AAP(m) and multisecant-GMRES(m)). *At global iteration t, let $\{\boldsymbol{x}_t^\ell\}_{\ell=0}^m$ be generated by AAP(m), then, for any $\boldsymbol{B}_t \in \mathbb{B}_t$,*

$$\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell = \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t, \qquad \text{where} \quad \widehat{\boldsymbol{p}}_t = \underset{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{B}_t,\boldsymbol{f}_t)}{\arg\min} \|\boldsymbol{B}_t\boldsymbol{p} - \boldsymbol{f}_t\|^2. \tag{3.1}$$

Further, $\boldsymbol{x}_{t+1} = \widehat{g}_t(\boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t)$ where $\widehat{g}_t(\boldsymbol{x}) := (1-\beta_t)\boldsymbol{x} + \beta_t[(\boldsymbol{I}+\boldsymbol{B}_t)(\boldsymbol{x}-\boldsymbol{x}_t) + g(\boldsymbol{x}_t)]$ is a (damped) linear approximation of g.

Proof. Let $\boldsymbol{B}_t \in \mathbb{B}_t$ be any multisecant matrix. By expressing $\boldsymbol{\alpha}$ in terms of $\boldsymbol{z}$ as given in (2.4),

$$\sum_{\ell=0}^m \alpha^\ell \boldsymbol{f}_t^\ell = \boldsymbol{f}_t^0 - z^0(\boldsymbol{f}_t^1 - \boldsymbol{f}_t^0) - \cdots - z^{m-1}(\boldsymbol{f}_t^m - \boldsymbol{f}_t^{m-1}) = \boldsymbol{f}_t - \boldsymbol{Y}_t\boldsymbol{z}.$$

From (2.7), $\boldsymbol{B}_t\boldsymbol{f}_t^\ell = \boldsymbol{f}_t^{\ell+1} - \boldsymbol{f}_t^\ell$ for $0 \le \ell \le m-1$. This implies that the residual sequence $\{\boldsymbol{f}_t^\ell\}_{\ell=0}^m$ is a Krylov sequence, i.e.,

$$\boldsymbol{f}_t^{\ell+1} = (\boldsymbol{I}+\boldsymbol{B}_t)\boldsymbol{f}_t^\ell = \cdots = (\boldsymbol{I}+\boldsymbol{B}_t)^{(\ell+1)}\boldsymbol{f}_t \in \mathcal{K}_{\ell+2}(\boldsymbol{I}+\boldsymbol{B}_t, \boldsymbol{f}_t). \tag{3.2}$$

Then, the columns of $\boldsymbol{Y}_t$ take the form $\boldsymbol{y}_t^\ell = \boldsymbol{f}_t^{\ell+1} - \boldsymbol{f}_t^\ell = (\boldsymbol{I}+\boldsymbol{B}_t)^{(\ell+1)}\boldsymbol{f}_t - (\boldsymbol{I}+\boldsymbol{B}_t)^{(\ell)}\boldsymbol{f}_t = \boldsymbol{B}_t(\boldsymbol{I}+\boldsymbol{B}_t)^{(\ell)}\boldsymbol{f}_t$, and thus

$$\operatorname{span}(\boldsymbol{Y}_t) = \operatorname{span}\{\boldsymbol{y}_t^0, \boldsymbol{y}_t^1, \cdots, \boldsymbol{y}_t^{m-1}\} = \boldsymbol{B}_t\mathcal{K}_m(\boldsymbol{I}+\boldsymbol{B}_t, \boldsymbol{f}_t) = \boldsymbol{B}_t\mathcal{K}_m(\boldsymbol{B}_t, \boldsymbol{f}_t),$$

where the last equality is due to the shift invariance property of the Krylov subspace. Therefore, we have the equivalence between the following three problems:

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^{m+1},\\ \sum_{\ell=0}^m \alpha^\ell = 1}} \|\sum_{\ell=0}^m \alpha^\ell \boldsymbol{f}_t^\ell\|^2 \quad \Leftrightarrow \min_{\boldsymbol{z}\in\mathbb{R}^m} \|\boldsymbol{Y}_t\boldsymbol{z} - \boldsymbol{f}_t\|^2 \quad \Leftrightarrow \min_{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{B}_t,\boldsymbol{f}_t)} \|\boldsymbol{B}_t\boldsymbol{p} - \boldsymbol{f}_t\|^2. \tag{3.3}$$

It is straightforward to verify that the solutions to each of the problems satisfy

$$\textstyle\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{f}_t^\ell = \boldsymbol{f}_t - \boldsymbol{Y}_t\boldsymbol{z}_t = \boldsymbol{f}_t - \boldsymbol{B}_t\widehat{\boldsymbol{p}}_t. \tag{3.4}$$

As $\boldsymbol{B}_t\boldsymbol{S}_t = \boldsymbol{Y}_t$, we have $\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t = \boldsymbol{Y}_t\boldsymbol{z}_t = \boldsymbol{B}_t\boldsymbol{S}_t\boldsymbol{z}_t$ holds for any $\boldsymbol{B}_t \in \mathbb{B}_t$. Since $\mathbb{B}_t$ contains invertible elements, we have $\widehat{\boldsymbol{p}}_t = \boldsymbol{S}_t\boldsymbol{z}_t$. The equivalence between $\boldsymbol{\alpha}_t$ and $\boldsymbol{z}_t$ then implies $\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell = \boldsymbol{x}_t - \boldsymbol{S}_t\boldsymbol{z}_t = \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t$, and proves (3.1).

Furthermore, let $\widehat{f}_t$ be the multisecant linear approximation of f at $\boldsymbol{x}_t$ as $\widehat{f}_t(\boldsymbol{x}) = \boldsymbol{B}_t(\boldsymbol{x}-\boldsymbol{x}_t) + f(\boldsymbol{x}_t)$. Then, $\widehat{g}_t(\boldsymbol{x}) = (1-\beta_t)\boldsymbol{x} + \beta_t(\boldsymbol{x} + \widehat{f}_t(\boldsymbol{x}))$. In addition, for all $0 \le \ell \le m$, $g(\boldsymbol{x}_t^\ell) = \boldsymbol{x}_t^\ell + f(\boldsymbol{x}_t^\ell)$ and $f(\boldsymbol{x}_t^\ell) = \widehat{f}_t(\boldsymbol{x}_t^\ell)$. It follows that

$$\textstyle\sum_{\ell=0}^m \alpha_t^\ell g(\boldsymbol{x}_t^\ell) = \sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell + \sum_{\ell=0}^m \alpha_t^\ell \widehat{f}_t(\boldsymbol{x}_t^\ell) = \sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell + \widehat{f}_t(\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell),$$

where the last equation is due to the linearity of $\widehat{f}_t$. Finally, we have $\boldsymbol{x}_{t+1} = (1-\beta_t)\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell + \beta_t[\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell + \widehat{f}_t(\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell)] = \widehat{g}_t(\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell) = \widehat{g}_t(\boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t)$, which completes the proof. ∎

We have shown that the LS problem (2.1) in AAP(m) is equivalent to applying GMRES(m) to the multisecant linear system $\boldsymbol{B}_t\boldsymbol{p} = \boldsymbol{f}_t$ for any $\boldsymbol{B}_t \in \mathbb{B}_t$. We call $\widehat{\boldsymbol{p}}_t$ the *multisecant-GMRES direction*, and $\widehat{\boldsymbol{p}}_t$ is equal to $\boldsymbol{S}_t\boldsymbol{Y}_t^{\dagger}\boldsymbol{f}_t$. As shown in Theorem 3.1, the weighted average of the history points equals the multisecant-GMRES update $\boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t$. The above theorem holds for any $\boldsymbol{B}_t \in \mathbb{B}_t$ because they generate the same Krylov (residual) subspaces

$$\mathcal{K}_m(\boldsymbol{B}_t, \boldsymbol{f}_t) = \operatorname{span}(\boldsymbol{S}_t), \qquad \boldsymbol{B}_t\mathcal{K}_m(\boldsymbol{B}_t, \boldsymbol{f}_t) = \operatorname{span}(\boldsymbol{Y}_t), \quad \forall\, \boldsymbol{B}_t \in \mathbb{B}_t. \tag{3.5}$$

We refer to [17] for the properties of matrices that generate the same Krylov residual spaces.

The equivalence between AAP and multisecant-GMRES resembles the one given in [37] for linear g. A similar result was also mentioned for a restarted variant of AA applied on gradient descent [26]. The equivalence between other variants of AA and the multisecant methods is discussed in [14].

Let us briefly take a closer look at the damping ratio. To that end, we define the *multisecant-GMRES residual* at iteration t as $\widehat{\boldsymbol{r}}_t$. According to the equivalence between the three minimization problems in (3.3), we have

$$\widehat{\boldsymbol{r}}_t = \boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t = \boldsymbol{Y}_t\boldsymbol{z}_t - \boldsymbol{f}_t = -\sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{f}_t^{\ell}. \tag{3.6}$$

The following corollary shows that the damping ratio acts as a stepsize parameter in the direction of the residual.

COROLLARY 3.2. *The iterate $\boldsymbol{x}_{t+1}$ generated from (2.2) in AAP satisfies*

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t - \beta_t\widehat{\boldsymbol{r}}_t. \tag{3.7}$$

Proof. The result follows from $\boldsymbol{x}_{t+1} = \sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{x}_t^{\ell} + \beta_t \sum_{\ell=0}^{m} \alpha_t^{\ell}\boldsymbol{f}_t^{\ell} = \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t - \beta_t\widehat{\boldsymbol{r}}_t$. ∎

When β_t satisfies $\|\boldsymbol{I} - \beta_t\boldsymbol{B}_t\| \le 1$, the "damped" direction $\widehat{\boldsymbol{p}}_t + \beta_t\widehat{\boldsymbol{r}}_t$ could be a better solution to the multisecant linear system $\boldsymbol{B}_t\boldsymbol{p} = \boldsymbol{f}_t$ than the original direction $\widehat{\boldsymbol{p}}_t$. Indeed, in this case, $\|\boldsymbol{B}_t(\widehat{\boldsymbol{p}}_t + \beta_t\widehat{\boldsymbol{r}}_t) - \boldsymbol{f}_t\| = \|(\boldsymbol{I} - \beta_t\boldsymbol{B}_t)(\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t)\| \le \|\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t\|$.

4. Connection Between AAP and Newton-GMRES. Newton's method finds the roots of a nonlinear function $f(\boldsymbol{x})$ by using its Jacobian information. At iteration t, a linear approximation of f at $\boldsymbol{x}_t$ is given by

$$f(\boldsymbol{x}) \approx \boldsymbol{J}_t(\boldsymbol{x} - \boldsymbol{x}_t) + f(\boldsymbol{x}_t), \tag{4.1}$$

where $\boldsymbol{J}_t := f'(\boldsymbol{x}_t)$ is the Jacobian of f at $\boldsymbol{x}_t$. The Newton-GMRES(m) method is a variant of the inexact-Newton method that uses GMRES(m) to approximately solve this linearized system for a search direction $\boldsymbol{p}_t^{\mathrm{N}}$ at each iteration as follows

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \boldsymbol{p}_t^{\mathrm{N}} \quad \text{with} \quad \boldsymbol{p}_t^{\mathrm{N}} := \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathcal{K}_m(\boldsymbol{J}_t, \boldsymbol{f}_t)} \|\boldsymbol{J}_t\boldsymbol{p} - \boldsymbol{f}_t\|. \tag{4.2}$$

However, the Jacobian is not always available or easy to compute directly.

As shown in Corollary 3.2, at iteration $\boldsymbol{x}_t$, the search direction given by AAP is equivalent to the multisecant-GMRES direction based on $\boldsymbol{B}_t \in \mathbb{B}_t$ with an additional damping term. In this section, we establish the connection between AAP and the Newton-GMRES method by showing that, (i) when the residual $\|\boldsymbol{f}_t\|$ goes to zero

as $t \to \infty$, the distance between the set of multisecant matrix $\mathbb{B}_t$ given by AAP and the Jacobian matrix $\boldsymbol{J}_t$ converges to zero (Section 4.1), (ii) the matrices $\boldsymbol{S}_t$, $\boldsymbol{Y}_t$, and $\boldsymbol{H}_t$ constructed by AAP converge respectively to their counterparts in the Newton-GMRES method (Section 4.2), and (iii) the optimization gain θ_t in AAP converges to an analogous quantities in Newton-GMRES (Section 4.3).

One of the key properties of AAP that enables the analysis in this section is the m Picard iteration steps taken in a global iteration. The Picard iteration steps often progress more conservatively than the acceleration step. Therefore, the multisecant matrix given by these history points could be a good approximation to the local Jacobian $\boldsymbol{J}_t$. In fact, the following lemma provides bounds on the Picard iteration updates and residuals, which are essential in the analysis in this section.

LEMMA 4.1. *Under Assumption* 1.1*(a), at global iteration* t *of AAP, for all* $0 \leq \ell \leq m$,

$$\|\boldsymbol{f}_t^\ell\| \leq \|\boldsymbol{f}_t\|, \quad \textit{and} \quad \|\boldsymbol{x}_t^\ell - \boldsymbol{x}_t\| \leq \ell \|\boldsymbol{f}_t\|.$$

Proof. Since $\boldsymbol{x}_t^\ell = g(\boldsymbol{x}_t^{\ell-1})$ and $\boldsymbol{x}_t^0 = \boldsymbol{x}_t$, we have $\boldsymbol{x}_t^\ell = \boldsymbol{x}_t^{\ell-1} + f(\boldsymbol{x}_t^{\ell-1})$ and thus $\|\boldsymbol{x}_t^\ell - \boldsymbol{x}_t\| = \|\sum_{i=0}^{\ell-1} f(\boldsymbol{x}_t^i)\| \leq \sum_{i=0}^{\ell-1} \|f(\boldsymbol{x}_t^i)\|$. The non-expansive property of g gives that, for $i = 1, \ldots, \ell$,

$$\|f(\boldsymbol{x}_t^i)\| = \|g(\boldsymbol{x}_t^i) - \boldsymbol{x}_t^i\| = \|g(\boldsymbol{x}_t^i) - g(\boldsymbol{x}_t^{i-1})\| \leq \kappa \|\boldsymbol{x}_t^i - \boldsymbol{x}_t^{i-1}\| = \kappa \|f(\boldsymbol{x}_t^{i-1})\|.$$

Since $\kappa \leq 1$, $\|\boldsymbol{f}_t^\ell\| \leq \|\boldsymbol{f}_t\|$ and $\|\boldsymbol{x}_t^\ell - \boldsymbol{x}_t\| \leq \sum_{i=0}^{\ell-1} \|f(\boldsymbol{x}_t)\| \leq \ell \|\boldsymbol{f}_t\|$. ∎

Another useful technique in the following analysis is the fundamental theorem for line integrals, i.e.,

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 f'(\boldsymbol{x} + r(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) \, \mathrm{d}r. \tag{4.3}$$

With these results, we show in the following subsections that, as the AAP residual $\boldsymbol{f}_t$ approaches zero, $\boldsymbol{B}_t$, $\boldsymbol{S}_t$, $\boldsymbol{Y}_t$, $\boldsymbol{H}_t$, and θ_t in AAP converge to their counterparts in Newton-GMRES.

4.1. Convergence of the multisecant matrix. We start the analysis with showing that the distance between the set of multisecant matrices $\mathbb{B}_t$ generated by AAP and the Jacobian $\boldsymbol{J}_t$ goes to zero provided that the residual approaches zero as $t \to \infty$. Here, the distance between $\mathbb{B}_t$ and $\boldsymbol{J}_t$ is defined as the norm of

$$\boldsymbol{E}_t := \boldsymbol{B}_t^* - \boldsymbol{J}_t, \quad \text{where} \quad \boldsymbol{B}_t^* := \operatorname*{arg\,min}_{\boldsymbol{B} \in \mathbb{B}_t} \|\boldsymbol{B} - \boldsymbol{J}_t\|. \tag{4.4}$$

Because $\boldsymbol{S}_t$ is full column rank, the minimal-norm property [37] gives

$$\boldsymbol{E}_t = (\boldsymbol{J}_t \boldsymbol{S}_t - \boldsymbol{Y}_t) \boldsymbol{S}_t^\dagger. \tag{4.5}$$

Note that $\boldsymbol{J}_t + \boldsymbol{E}_t = \boldsymbol{B}_t^* \in \mathbb{B}_t$. In addition, when g is linear, $\boldsymbol{J}_t$ satisfies $\boldsymbol{J}_t \boldsymbol{S}_t = \boldsymbol{Y}_t$ and thus $\|\boldsymbol{E}_t\| = 0$. An upper bound on $\|\boldsymbol{E}_t\|$ for the nonlinear case is provided in the following theorem.

THEOREM 4.2. *Under Assumption* 1.1,

$$\|\boldsymbol{E}_t\| \leq \gamma m^{\frac{3}{2}} \operatorname{cond}(\boldsymbol{S}_t) \|\boldsymbol{f}_t\|. \tag{4.6}$$

Furthermore, if Assumption 2.2 *holds, then*

$$\|\boldsymbol{E}_t\| \leq C_E \|\boldsymbol{f}_t\|, \qquad \forall \; t > T, \tag{4.7}$$

where $C_E > 0$ *is a constant that is independent of* t.

Proof. It follows from (4.5) that

$$\|\boldsymbol{E}_t\| \le \|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{S}_t\|\|\boldsymbol{S}_t^\dagger\|.$$

Let $\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{S}_t = [\boldsymbol{v}^0, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^{m-1}] \in \mathbb{R}^{d\times m}$. For $0 \le \ell \le m-1$,

$$\boldsymbol{v}^\ell = \boldsymbol{y}^\ell - \boldsymbol{J}_t\boldsymbol{s}^\ell = f(\boldsymbol{x}_t^{\ell+1}) - f(\boldsymbol{x}_t^\ell) - f'(\boldsymbol{x}_t)(\boldsymbol{x}_t^{\ell+1} - \boldsymbol{x}_t^\ell).$$

By using (4.3) and $\boldsymbol{s}_t^\ell = \boldsymbol{x}_t^{\ell+1} - \boldsymbol{x}_t^\ell = \boldsymbol{f}_t^\ell$,

$$\boldsymbol{v}^\ell = \int_0^1 f'(\boldsymbol{x}_t^\ell + r\boldsymbol{f}_t^\ell)\boldsymbol{f}_t^\ell \,\mathrm{d}r - f'(\boldsymbol{x}_t)\boldsymbol{f}_t^\ell = \int_0^1 [f'(\boldsymbol{x}_t^\ell + r\boldsymbol{f}_t^\ell) - f'(\boldsymbol{x}_t)]\boldsymbol{f}_t^\ell \,\mathrm{d}r.$$

Applying $f'(\boldsymbol{x}) = g'(\boldsymbol{x}) - \boldsymbol{I}$ and the Lipschitz continuity of g' gives, for $r \in [0,1]$,

$$\|f'(\boldsymbol{x}_t^\ell + r\boldsymbol{f}_t^\ell) - f'(\boldsymbol{x}_t)\| \le \gamma\|\boldsymbol{x}_t^\ell + r\boldsymbol{f}_t^\ell - \boldsymbol{x}_t\| \le \gamma(\|\boldsymbol{x}_t^\ell - \boldsymbol{x}_t\| + \|\boldsymbol{f}_t^\ell\|).$$

By Lemma 4.1, $\|\boldsymbol{x}_t^\ell - \boldsymbol{x}_t\| + \|\boldsymbol{f}_t^\ell\| \le (\ell+1)\|\boldsymbol{f}_t\| \le \gamma m\|\boldsymbol{f}_t\|$ as $\ell < m$, which leads to

$$\|\boldsymbol{v}^\ell\| \le \int_0^1 \|f'(\boldsymbol{x}_t^\ell + r\boldsymbol{f}_t^\ell) - f'(\boldsymbol{x}_t)\|\|\boldsymbol{f}_t^\ell\| \,\mathrm{d}r \le \gamma m\|\boldsymbol{f}_t\|\|\boldsymbol{s}_t^\ell\|. \tag{4.8}$$

Using the matrix norm inequality $\|\cdot\| \le \|\cdot\|_F$,

$$\|\boldsymbol{J}_t\boldsymbol{S}_t - \boldsymbol{Y}_t\| \le \|\boldsymbol{J}_t\boldsymbol{S}_t - \boldsymbol{Y}_t\|_F = \sqrt{\sum_\ell \|\boldsymbol{v}^\ell\|^2} \le \gamma m\|\boldsymbol{f}_t\|\sqrt{\sum_\ell \|\boldsymbol{s}_t^\ell\|^2} = \gamma m\|\boldsymbol{f}_t\|\|\boldsymbol{S}_t\|_F.$$

Since $\boldsymbol{S}_t$ has full column rank and $\|\boldsymbol{S}_t\|_F \le \sqrt{m}\|\boldsymbol{S}_t\|$, it follows that

$$\|\boldsymbol{E}_t\| \le \sqrt{m}\gamma m\|\boldsymbol{f}_t\|\|\boldsymbol{S}_t\|\|\boldsymbol{S}_t^\dagger\| = \gamma m^{\frac{3}{2}}\|\boldsymbol{f}_t\|\mathrm{cond}(\boldsymbol{S}_t).$$

The second claim is then a direct consequence of Assumption 2.2. ∎

The bound given in (4.7) shows that $\|\boldsymbol{E}_t\| \le \mathcal{O}(\|\boldsymbol{f}_t\|)$, which leads to the following convergence result.

COROLLARY 4.3. *Under Assumptions 1.1 and 2.2, suppose that $\boldsymbol{f}_t \to 0$ as $t \to \infty$, then, as $t \to \infty$,*

$$\boldsymbol{E}_t \to 0 \quad \textit{and} \quad \boldsymbol{B}_t^* \to \boldsymbol{J}_t,$$

where $\boldsymbol{B}_t^* \in \mathbb{B}_t$ is defined in (4.4).

The convergence of $\boldsymbol{B}_t$ to $\boldsymbol{J}_t$ suggests that the multisecant direction $\widehat{\boldsymbol{p}}_t$ could converge to the Newton-GMRES direction $\boldsymbol{p}_t^{\mathrm{N}} := \arg\min_{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{J}_t,\boldsymbol{f}_t)} \|\boldsymbol{J}_t\boldsymbol{p} - \boldsymbol{f}_t\|$, which we prove later in Corollary 4.6.

4.2. Limits of $\boldsymbol{S}_t$, $\boldsymbol{Y}_t$, and $\boldsymbol{H}_t$. In this section, we establish the limits of $\boldsymbol{S}_t$ and $\boldsymbol{Y}_t$ as $t \to \infty$ under the assumption that $\boldsymbol{f}_t$ approaches zero. These limits are related to a Krylov matrix spanned by $g'(\boldsymbol{x}_t)$ and $\boldsymbol{f}_t$, which further implies the convergence of the approximate Jacobian inverse $\boldsymbol{H}_t$ of AAP defined in (2.6).

We first review the Newton-GMRES method. At iteration $\boldsymbol{x}_t$, the Krylov matrix associated with $\mathcal{K}_m(\boldsymbol{J}_t, \boldsymbol{f}_t)$ can be written as

$$\boldsymbol{F}_t := [\boldsymbol{f}_t, \boldsymbol{J}_t^{(1)}\boldsymbol{f}_t, \boldsymbol{J}_t^{(2)}\boldsymbol{f}_t, \cdots, \boldsymbol{J}_t^{(m-1)}\boldsymbol{f}_t] \in \mathbb{R}^{d\times m}. \tag{4.9}$$

The Newton-GMRES direction $\boldsymbol{p}_t^{\mathrm{N}} := \arg\min_{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{J}_t,\boldsymbol{f}_t)} \|\boldsymbol{J}_t\boldsymbol{p}-\boldsymbol{f}_t\|$ satisfies $\boldsymbol{J}_t\boldsymbol{p}_t^{\mathrm{N}} = \mathcal{P}_{\boldsymbol{J}_t\boldsymbol{F}_t}(\boldsymbol{f}_t)$ where $\mathcal{P}_{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{A}^\dagger$ denotes the orthogonal projection onto the subspace spanned by the columns of $\boldsymbol{A}$. Define

$$\boldsymbol{G}_t := [\boldsymbol{f}_t, g'(\boldsymbol{x}_t)^{(1)}\boldsymbol{f}_t, g'(\boldsymbol{x}_t)^{(2)}\boldsymbol{f}_t, \cdots, g'(\boldsymbol{x}_t)^{(m-1)}\boldsymbol{f}_t] \in \mathbb{R}^{d\times m}. \tag{4.10}$$

Since $g'(\boldsymbol{x}_t) = \boldsymbol{I}+\boldsymbol{J}_t$, the shift invariance of the Krylov subspace leads to $\mathcal{P}_{\boldsymbol{G}_t} = \mathcal{P}_{\boldsymbol{F}_t}$. Therefore, $\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t} = \mathcal{P}_{\boldsymbol{J}_t\boldsymbol{F}_t}$. If $\boldsymbol{J}_t$ is invertible, we have

$$\boldsymbol{p}_t^{\mathrm{N}} = \boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}(\boldsymbol{f}_t). \tag{4.11}$$

We refer to $\boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}$ as the *Newton-GMRES update operator*.

As shown in (3.2), the AAP residuals $\{\boldsymbol{f}_t^\ell\}_{\ell=0}^{m-1}$ form a basis of the Krylov subspace $\mathcal{K}_m(\boldsymbol{B}_t, \boldsymbol{f}_t)$ as $\boldsymbol{f}_t^\ell = (\boldsymbol{I}+\boldsymbol{B}_t)^{(\ell)}\boldsymbol{f}_t$. Therefore,

$$\boldsymbol{S}_t = [\boldsymbol{f}_t^0, \boldsymbol{f}_t^1, \cdots, \boldsymbol{f}_t^{m-1}] = [\boldsymbol{f}_t, (\boldsymbol{I}+\boldsymbol{B}_t)^{(1)}\boldsymbol{f}_t, (\boldsymbol{I}+\boldsymbol{B}_t)^{(2)}\boldsymbol{f}_t, \cdots, (\boldsymbol{I}+\boldsymbol{B}_t)^{(m-1)}\boldsymbol{f}_t]$$

is a Krylov matrix associated with $\boldsymbol{I}+\boldsymbol{B}_t$ and $\boldsymbol{f}_t$. As $\boldsymbol{f}_t$ approaches zero, the matrices $\boldsymbol{S}_t$, $\boldsymbol{Y}_t$, and $\boldsymbol{G}_t$ all approach zero. Further, the following lemma gives estimates of the distances between $\boldsymbol{S}_t$, $\boldsymbol{Y}_t$ and their Newton-GMRES counterparts.

LEMMA 4.4. *Under Assumption 1.1,*

$$\|\boldsymbol{S}_t - \boldsymbol{G}_t\| \le \gamma m^{\frac{5}{2}}\|\boldsymbol{f}_t\|^2 \qquad \textit{and} \qquad \|\boldsymbol{Y}_t - \boldsymbol{J}_t\,\boldsymbol{G}_t\| \le 2\gamma m^{\frac{5}{2}}\|\boldsymbol{f}_t\|^2. \tag{4.12}$$

Proof. For a better presentation, we omit the subscript t in this proof unless necessary. Note that $\boldsymbol{x}^0 = \boldsymbol{x}_t$ is the starting point for the Picard iterations at the global iteration t, and $\boldsymbol{f}^\ell = f(\boldsymbol{x}_t^\ell)$ for all $0 \le \ell \le m-1$.

We first define $\boldsymbol{\Delta}_t = \boldsymbol{G}_t - \boldsymbol{S}_t$ and write $\boldsymbol{\Delta}_t = [\boldsymbol{\epsilon}^0, \boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \cdots, \boldsymbol{\epsilon}^{m-1}]$. By the definitions of $\boldsymbol{G}_t$ and $\boldsymbol{S}_t$, $\boldsymbol{\epsilon}^0 = 0$ and $\boldsymbol{\epsilon}^\ell = \boldsymbol{f}^\ell - g'(\boldsymbol{x}_t)^{(\ell)}\boldsymbol{f}_t$ for $\ell = 0, \ldots, m-1$. We next estimate $\boldsymbol{\epsilon}^\ell$. By (4.3), $f(\boldsymbol{x}^\ell) - f(\boldsymbol{x}^{\ell-1}) = \int_0^1 f'(\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1})\boldsymbol{f}^{\ell-1}\,\mathrm{d}r$ as $\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1} = \boldsymbol{f}^\ell$. Thus,

$$\begin{aligned}
\boldsymbol{f}^\ell = f(\boldsymbol{x}^\ell) &= \textstyle\int_0^1 [\boldsymbol{I} + f'(\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1})]\boldsymbol{f}^{\ell-1}\,\mathrm{d}r \\
&= \textstyle\int_0^1 g'(\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1})\boldsymbol{f}^{\ell-1}\,\mathrm{d}r \\
&= g'(\boldsymbol{x}_t)\boldsymbol{f}^{\ell-1} + \textstyle\int_0^1 [g'(\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1}) - g'(\boldsymbol{x}_t)]\boldsymbol{f}^{\ell-1}\,\mathrm{d}r \\
&= \boldsymbol{K}\boldsymbol{f}^{\ell-1} + \boldsymbol{v}^{\ell-1},
\end{aligned}$$

where $\boldsymbol{K} := g'(\boldsymbol{x}_t)$ and $\boldsymbol{v}^{\ell-1} := \int_0^1 \left[g'(\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1}) - g'(\boldsymbol{x}_t)\right]\boldsymbol{f}^{\ell-1}\,\mathrm{d}r$. Applying this relation iteratively gives $\boldsymbol{f}^\ell = \boldsymbol{K}^{(\ell)}\boldsymbol{f}_t + \sum_{i=0}^{\ell-1}\boldsymbol{K}^{(i)}\boldsymbol{v}^{\ell-1-i}$. Therefore, $\boldsymbol{\epsilon}^\ell = \sum_{i=0}^{\ell-1}\boldsymbol{K}^{(i)}\boldsymbol{v}^{\ell-1-i}$. Furthermore, under Assumption 1.1, $\|\boldsymbol{K}\| \le \kappa \le 1$ and

$$\|\boldsymbol{v}^{\ell-1}\| \le \gamma\|\boldsymbol{x}^{\ell-1} + r\boldsymbol{f}^{\ell-1} - \boldsymbol{x}_t\|\|\boldsymbol{f}^{\ell-1}\| \le \gamma\ell\|\boldsymbol{f}_t\|\|\boldsymbol{f}^{\ell-1}\| \le m\gamma\|\boldsymbol{f}_t\|^2,$$

where the second inequality follows from Lemma 4.1 and $0 \le r \le 1$. Then,

$$\|\boldsymbol{\epsilon}^\ell\| = \textstyle\sum_{i=0}^{\ell-1}\boldsymbol{K}^{(i)}\boldsymbol{v}^{\ell-1-i} \le \sum_{i=0}^{\ell-1}\|\boldsymbol{v}^{\ell-1-i}\| \le \ell(m\gamma\|\boldsymbol{f}_t\|^2) \le m^2\gamma\|\boldsymbol{f}_t\|^2.$$

Therefore,

$$\|\boldsymbol{S}_t - \boldsymbol{G}_t\| = \|\boldsymbol{\Delta}_t\| \le \|\boldsymbol{\Delta}_t\|_F = \sqrt{\textstyle\sum_{i=0}^{m-1}\|\boldsymbol{\epsilon}^i\|^2} \le \gamma m^{\frac{5}{2}}\|\boldsymbol{f}_t\|^2. \tag{4.13}$$

To estimate $\|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{G}_t\|$, we use the relations established above and write the ℓ-th column of $\boldsymbol{Y}_t$ as

$$\boldsymbol{y}_t^\ell = \boldsymbol{f}^{\ell+1} - \boldsymbol{f}^\ell = \boldsymbol{K}^{(\ell+1)}\boldsymbol{f}_t + \boldsymbol{\epsilon}^{\ell+1} - \boldsymbol{K}^{(\ell)}\boldsymbol{f}_t - \boldsymbol{\epsilon}^\ell = \boldsymbol{J}_t\boldsymbol{K}^{(\ell)}\boldsymbol{f}_t + \boldsymbol{\epsilon}^{\ell+1} - \boldsymbol{\epsilon}^\ell,$$

where the last equality is due to $\boldsymbol{K} = g'(\boldsymbol{x}_t) = \boldsymbol{I} + f'(\boldsymbol{x}_t) = \boldsymbol{I} + \boldsymbol{J}_t$. Then, we write $\boldsymbol{Y}_t = \boldsymbol{J}_t\boldsymbol{G}_t + \widetilde{\boldsymbol{\Delta}}_t - \boldsymbol{\Delta}_t$ where $\widetilde{\boldsymbol{\Delta}}_t := [\boldsymbol{\epsilon}^1, \boldsymbol{\epsilon}^2, \boldsymbol{\epsilon}^3, \cdots, \boldsymbol{\epsilon}^m] \in \mathbb{R}^{d\times m}$. Following a similar approach as above, one can obtain $\|\widetilde{\boldsymbol{\Delta}}_t\| \le \gamma m^{\frac{5}{2}}\|\boldsymbol{f}_t\|^2$, and thus

$$\|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{G}_t\| \le \|\widetilde{\boldsymbol{\Delta}}_t\| + \|\boldsymbol{\Delta}_t\| \le 2\gamma m^{\frac{5}{2}}\|\boldsymbol{f}_t\|^2. \tag{4.14}$$ ∎

When $\boldsymbol{f}_t$ approaches 0, despite $\boldsymbol{S}_t$ and $\boldsymbol{Y}_t$ converging to 0, the above lemma shows that $\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{S}_t \to \frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t$ and $\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t \to \frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t\boldsymbol{G}_t$. We note that, when g is linear, i.e., $\gamma = 0$, for all t,

$$\boldsymbol{S}_t = \boldsymbol{G}_t, \qquad \text{and} \qquad \boldsymbol{Y}_t = \boldsymbol{J}_t\boldsymbol{G}_t. \tag{4.15}$$

Now we are ready to give the convergence of the approximate Jacobian inverse $\boldsymbol{H}_t$ in AAP. As given in (2.6), $\boldsymbol{H}_t = \boldsymbol{S}_t\boldsymbol{Y}_t^\dagger - \beta_t(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t})$ where β_t is the damping ratio. The definitions of $\widehat{\boldsymbol{p}}_t$ and $\widehat{\boldsymbol{r}}_t$ in (3.1) and (3.6) lead to $\widehat{\boldsymbol{p}}_t = \boldsymbol{S}_t\boldsymbol{Y}_t^\dagger\boldsymbol{f}_t$ and $\widehat{\boldsymbol{r}}_t = -\beta_t(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t})\boldsymbol{f}_t$, respectively. Thus, $\boldsymbol{S}_t\boldsymbol{Y}_t^\dagger$ corresponds to the multisecant-GMRES update operator, and $\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t}$ corresponds to the extra linearized fixed-point iteration in AAP. The following corollary gives the convergence of $\boldsymbol{S}_t\boldsymbol{Y}_t^\dagger$ to the Newton-GMRES update operator defined in (4.11), as well as the convergence of $\boldsymbol{H}_t$.

ASSUMPTION 4.1. *There exist constants $T > 0$ and $M > 0$ such that, for all $t \ge T$, $\boldsymbol{J}_t$ is invertible, $\sigma_{\min}(\boldsymbol{Y}_t) \ge M\|\boldsymbol{f}_t\|$, and $\boldsymbol{J}_t\boldsymbol{G}_t$ has full column rank.*

THEOREM 4.5. *Under Assumptions 1.1 and 4.1, when $\boldsymbol{f}_t$ goes to 0,*

$$\boldsymbol{S}_t\boldsymbol{Y}_t^\dagger \to \boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}, \qquad \boldsymbol{H}_t \to \boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t} - \beta_t(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}). \tag{4.16}$$

Proof. When $\boldsymbol{f}_t$ goes to 0, Lemma 4.4 gives $\boldsymbol{Y}_t \to \boldsymbol{J}_t\boldsymbol{G}_t$. Because both $\boldsymbol{Y}_t$ and $\boldsymbol{J}_t\boldsymbol{G}_t$ have full column rank, it follows from [34, Theorem 3.4] that

$$\|\boldsymbol{Y}_t^\dagger - (\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger\| \le \sqrt{2}\|\boldsymbol{Y}_t^\dagger\|\,\|(\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger\|\,\|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{G}_t\|. \tag{4.17}$$

Furthermore, $\|\boldsymbol{Y}_t^\dagger\| = \frac{1}{\sigma_{\min}(\boldsymbol{Y}_t)} \le \frac{1}{M\|\boldsymbol{f}_t\|}$ and, for sufficiently small $\|\boldsymbol{f}_t\|$,

$$\|(\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger\| = \frac{1}{\sigma_{\min}(\boldsymbol{J}_t\boldsymbol{G}_t)} \le \frac{1}{\sigma_{\min}(\boldsymbol{Y}_t) - \|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{G}_t\|} \le \tau\frac{1}{M\|\boldsymbol{f}_t\|}, \tag{4.18}$$

where $\tau < 1$ is a constant and the first inequality follows from [16, Corollary 2.4.4] and the estimate in Lemma 4.4 on $\|\boldsymbol{Y}_t - \boldsymbol{J}_t\boldsymbol{G}_t\|$. With these estimates, the RHS of (4.17) is uniformly bounded and thus, as $\|\boldsymbol{f}_t\| \to 0$,

$$\|(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t)^\dagger - (\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger\| = \|\boldsymbol{f}_t\|\|\boldsymbol{Y}_t^\dagger - (\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger\| \to 0. \tag{4.19}$$

Therefore, since $\boldsymbol{J}_t$ is invertible, $\boldsymbol{G}_t = \boldsymbol{J}_t^{-1}\boldsymbol{J}_t\boldsymbol{G}_t$, and then

$$\boldsymbol{S}_t\boldsymbol{Y}_t^\dagger = (\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{S}_t)(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t)^\dagger \quad \to \quad \boldsymbol{J}_t^{-1}(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t\boldsymbol{G}_t)(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger = \boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}.$$

Further, $\mathcal{P}_{\boldsymbol{Y}_t} = \boldsymbol{Y}_t\boldsymbol{Y}_t^\dagger \to (\boldsymbol{J}_t\boldsymbol{G}_t)(\boldsymbol{J}_t\boldsymbol{G}_t)^\dagger = \mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}$. It follows that

$$\boldsymbol{H}_t = \boldsymbol{S}_t\boldsymbol{Y}_t^\dagger - \beta_t(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t}) \quad \to \quad \boldsymbol{J}_t^{-1}\mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t} - \beta_t(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{J}_t\boldsymbol{G}_t}),$$

which completes the proof. ∎

When $\beta_t = 0$, the above theorem demonstrates the convergence of the approximate Jacobian inverse $\boldsymbol{H}_t$ to the Newton-GMRES update operator. This suggests that the iteration sequences generated by AAP and Newton-GMRES are likely to be close to each other, provided that they start from the same initial point with a sufficiently small residual.

As $\widehat{\boldsymbol{p}}_t = \boldsymbol{S}_t \boldsymbol{Y}_t^\dagger \boldsymbol{f}_t$, the following convergence of $\widehat{\boldsymbol{p}}_t$ to the Newton-GMRES search direction $\boldsymbol{p}_t^N = \boldsymbol{J}_t^{-1} \mathcal{P}_{\boldsymbol{J}_t \boldsymbol{G}_t}(\boldsymbol{f}_t)$ is straightforward.

COROLLARY 4.6. *Under Assumptions 1.1 and 4.1, when $\boldsymbol{f}_t$ goes to 0, we have $\frac{1}{\|\boldsymbol{f}_t\|}\widehat{\boldsymbol{p}}_t \to \frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{p}_t^N$.*

Remark 4.7. For Assumption 4.1, the condition that $\boldsymbol{J}_t$ is invertible is satisfied when g is a contractive mapping, i.e., $\kappa < 1$.

4.3. Convergence of the optimization gain. As mentioned in Section 2.3, the optimization gain θ_t, which is defined as the ratio between the minimum of the constrained LS problem (2.1) and $\|\boldsymbol{f}_t\|$, is a crucial factor in the convergence analysis of Anderson acceleration. Based on (3.6), the optimization gain can be written as

$$\theta_t := \frac{\|\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{f}_t^\ell\|}{\|\boldsymbol{f}_t\|} = \frac{\|\boldsymbol{B}_t \widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t\|}{\|\boldsymbol{f}_t\|} = \frac{\|\boldsymbol{Y}_t \boldsymbol{z}_t - \boldsymbol{f}_t\|}{\|\boldsymbol{f}_t\|} = \frac{\|(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t})(\boldsymbol{f}_t)\|}{\|\boldsymbol{f}_t\|}, \tag{4.20}$$

In this section, we provide convergence estimates for θ_t to an analogous gain term in the Newton-GMRES method.

Recall that at iterate $\boldsymbol{x}_t$, the Newton-GMRES method uses GMRES(m) to solve $\boldsymbol{J}_t \boldsymbol{p} - \boldsymbol{f}_t = 0$ where $\boldsymbol{J}_t = f'(\boldsymbol{x}_t)$ is the Jacobian at $\boldsymbol{x}_t$. We define the *Jacobian-GMRES(m) gain* as $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ where $\boldsymbol{r}_t^J$ is the GMRES residual:

$$\|\boldsymbol{r}_t^J\| := \min_{\boldsymbol{p} \in \mathcal{K}_m(\boldsymbol{J}_t, \boldsymbol{f}_t)} \|\boldsymbol{J}_t \boldsymbol{p} - \boldsymbol{f}_t\| = \|(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{J}_t \boldsymbol{G}_t})(\boldsymbol{f}_t)\|, \tag{4.21}$$

as shown in (4.11). The term *Jacobian-GMRES(m) gain* is used to distinguish it from the Newton-GMRES algorithm since $\boldsymbol{r}_t^J$ is evaluated at iteration $\boldsymbol{x}_t$ generated by the AAP method. If g is a linear function, we have $\theta_t = \|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ according to (4.15). The following theorem provides the convergence of θ_t to $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ provided that $\|\boldsymbol{f}_t\| \to 0$ as $t \to \infty$.

THEOREM 4.8 (Limit of optimization gain). *Under Assumptions 1.1 and 4.1, when $\|\boldsymbol{f}_t\|$ is small enough,*

$$\left| \theta_t - \|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\| \right| \leq \mathcal{O}(\|\boldsymbol{f}_t\|), \tag{4.22}$$

implying that θ_t converges to $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ as $\|\boldsymbol{f}_t\|$ approaches zero.

Proof. By the definitions,

$$\left| \theta_t - \frac{\|\boldsymbol{r}_t^J\|}{\|\boldsymbol{f}_t\|} \right| = \frac{\|(\boldsymbol{I} - \mathcal{P}_{\boldsymbol{Y}_t})(\boldsymbol{f}_t) - (\boldsymbol{I} - \mathcal{P}_{\boldsymbol{J}_t \boldsymbol{G}_t})(\boldsymbol{f}_t)\|}{\|\boldsymbol{f}_t\|} \leq \|\mathcal{P}_{\boldsymbol{Y}_t} - \mathcal{P}_{\boldsymbol{J}_t \boldsymbol{G}_t}\|. \tag{4.23}$$

Lemma 4.4 shows that $\|\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t - \frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t \boldsymbol{G}_t\| \leq \mathcal{O}(\|\boldsymbol{f}_t\|)$. Meanwhile, when $\|\boldsymbol{f}_t\|$ is small enough, (4.19) in Theorem 4.5 shows that $\|(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t)^\dagger - (\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t \boldsymbol{G}_t)^\dagger\| \leq \mathcal{O}(\|\boldsymbol{f}_t\|)$, Therefore,

$$\|\mathcal{P}_{\boldsymbol{Y}_t} - \mathcal{P}_{\boldsymbol{J}_t \boldsymbol{G}_t}\| = \|(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t)(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Y}_t)^\dagger - (\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t \boldsymbol{G}_t)(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{J}_t \boldsymbol{G}_t)^\dagger\| \leq \mathcal{O}(\|\boldsymbol{f}_t\|),$$

which finished the proof. ∎

We showed that the optimization gain converges to the Jacobian-GMRES(m) gain $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ as the residual $\boldsymbol{f}_t$ approaches zero. Thus, $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ can serve as an estimator of θ_t when the residual is small. For a general $\boldsymbol{J}_t$, the bounds on the m-th residual produced by GMRES applied to $\boldsymbol{J}_t\boldsymbol{p} = \boldsymbol{f}_t$ can vary widely. However, when $\boldsymbol{J}_t$ is symmetric and positive definite, the norm of the m-th GMRES residual is well-known to be bounded as follows

$$\frac{\|\boldsymbol{r}_t^J\|}{\|\boldsymbol{f}_t\|} \leq 2\left(\frac{\sqrt{\mathrm{cond}(\boldsymbol{J}_t)}-1}{\sqrt{\mathrm{cond}(\boldsymbol{J}_t)}+1}\right)^k \approx 2\left(1-\frac{2}{\sqrt{\mathrm{cond}(\boldsymbol{J}_t)}}\right)^k, \tag{4.24}$$

showing exponential decay with respect to m. However, according to (4.12), when m is large, the upper bound in (4.22) could have a large constant, which makes the bound loose. Numerical examples are provided in Figure 2 in Section 6.1.

Remark 4.9. Another interpretation that illustrates the relationship between θ_t and $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ is based on the equivalence between AAP and multisecant-GMRES. Specifically, from (3.6), one can write $\theta_t\|\boldsymbol{f}_t\| = \|\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t\|$, for any $\boldsymbol{B}_t \in \mathbb{B}_t$. Since $(\boldsymbol{J}_t + \boldsymbol{E}_t)\boldsymbol{S}_t = \boldsymbol{Y}_t$, we have

$$\theta_t\|\boldsymbol{f}_t\| = \min_{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{J}_t+\boldsymbol{E}_t,\boldsymbol{f}_t)} \|(\boldsymbol{J}_t+\boldsymbol{E}_t)\boldsymbol{p} - \boldsymbol{f}_t\|.$$

Combining this with (4.21), the relation between θ_t and $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ can be regarded as the change in GMRES residuals under perturbation $\boldsymbol{E}_t$. Interested readers can refer to [32], where the GMRES residual change with respect to perturbations on the coefficient matrix is estimated using spectral perturbation theory and resolvent estimates. A simple application of Theorem 2.1 therein gives $\left|\,\theta_t - \|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|\,\right| \leq \mathcal{O}(\|\boldsymbol{E}_t\|)$ under some conditions. This bound is essentially identical as (4.22).

5. Convergence analysis. This section presents the convergence analysis of AAP. We provide a one-step convergence analysis to establish the bounds on $\|\boldsymbol{f}_{t+1}\|$, followed by a local convergence result. These analyses are based on the equivalence between AAP and multisecant-GMRES. We present some preliminary results before starting the convergence analysis.

First, recall Corollary 3.2 which states that the update of $\boldsymbol{x}_{t+1}$ in AAP satisfies

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t - \beta_t\widehat{\boldsymbol{r}}_t,$$

where $\widehat{\boldsymbol{r}}_t = \boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t$ is the residual of the multisecant-GMRES. It follows from the optimization gain $\theta_t = \|\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t\|/\|\boldsymbol{f}_t\|$ that $\|\widehat{\boldsymbol{r}}_t\| = \theta_t\|\boldsymbol{f}_t\|$. Meanwhile, the following lemma characterizes the lengths of $\|\widehat{\boldsymbol{p}}_t\|$.

LEMMA 5.1. *For any invertible $\boldsymbol{B}_t \in \mathbb{B}_t$, we have $\widehat{\boldsymbol{p}}_t$ satisfies*

$$\|\widehat{\boldsymbol{p}}_t\| \leq \sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\|\boldsymbol{f}_t\|.$$

Proof. Since $\widehat{\boldsymbol{p}}_t = \arg\min_{\boldsymbol{p}\in\mathcal{K}_m(\boldsymbol{B}_t,\boldsymbol{f}_t)}\|\boldsymbol{B}_t\boldsymbol{p} - \boldsymbol{f}_t\|$, $\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t \perp (\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t)$. Thus, $\|\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t\| = \sqrt{1-\theta_t^2}\|\boldsymbol{f}_t\|$ as $\|\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \boldsymbol{f}_t\| = \theta_t\|\boldsymbol{f}_t\|$. The result follows from the nonsingularity of $\boldsymbol{B}_t$. ∎

We note that assuming $\boldsymbol{B}_t$ nonsingular does not impose additional restrictions, since the equivalence of AAP and multi-secant GMRES holds for any $\boldsymbol{B}_t \in \mathbb{B}_t$ (Section 3) and $\mathbb{B}_t$ has nonsingular elements under Assumption 2.1 as shown in Section 2.2.

The above bound holds for any invertible $\boldsymbol{B}_t \in \mathbb{B}_t$. We give a lower bound estimate of $\|\boldsymbol{B}_t^{-1}\|$: considering that $\boldsymbol{B}_t^{-1}$ satisfies $\boldsymbol{B}_t^{-1}\boldsymbol{Y}_t = \boldsymbol{S}_t$, we have, according to the minimal norm properties of pseudoinverse, $\|\boldsymbol{B}_t^{-1}\| \geq \|\boldsymbol{S}_t\boldsymbol{Y}_t^{\dagger}\|, \forall\, \boldsymbol{B}_t \in \mathbb{B}_t$ being invertible. On the other hand, if $\boldsymbol{J}_t$ is invertible and $\|\boldsymbol{J}_t^{-1}\boldsymbol{E}_t\| < 1$, Theorem 2.3.4 of [16] gives that $\boldsymbol{J}_t + \boldsymbol{E}_t$ is invertible and

$$\|(\boldsymbol{J}_t + \boldsymbol{E}_t)^{-1} - \boldsymbol{J}_t^{-1}\| \leq \frac{\|\boldsymbol{E}_t\|\,\|\boldsymbol{J}_t^{-1}\|^2}{1 - \|\boldsymbol{J}_t^{-1}\boldsymbol{E}_t\|}. \tag{5.1}$$

When $\|\boldsymbol{E}_t\|$ is small enough, the denominator is negligible. It follows that $\|(\boldsymbol{J}_t + \boldsymbol{E}_t)^{-1} - \boldsymbol{J}_t^{-1}\| \leq \mathcal{O}(\|\boldsymbol{E}_t\|)$ and $\|\boldsymbol{J}_t^{-1}\|$ can be used to estimate $\|\boldsymbol{B}_t^{-1}\|$ when $\|\boldsymbol{E}_t\|$ is small enough.

Now, we are ready to give the bound of $\|\boldsymbol{f}_{t+1}\|$.

THEOREM 5.2. *Under Assumptions 1.1 and 2.2, for any invertible* $\boldsymbol{B}_t \in \mathbb{B}_t$,

$$\|\boldsymbol{f}_{t+1}\| \leq [(1-\beta_t) + \beta_t\kappa]\,\theta_t\|\boldsymbol{f}_t\| + \sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\left[\frac{\gamma}{2}\sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\| + C_E\right]\|\boldsymbol{f}_t\|^2. \tag{5.2}$$

Proof. The update can be expressed as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \boldsymbol{p}_t = \boldsymbol{x}_t - (\widehat{\boldsymbol{p}}_t + \beta_t\widehat{\boldsymbol{r}}_t) = \widehat{\boldsymbol{x}}_t - \beta_t\widehat{\boldsymbol{r}}_t,$$

where $\widehat{\boldsymbol{x}}_t := \boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t = \sum_\ell \alpha^\ell \boldsymbol{x}_t^\ell$. By (4.3),

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &= f(\boldsymbol{x}_t) + [f(\widehat{\boldsymbol{x}}_t) - f(\boldsymbol{x}_t)] + [f(\boldsymbol{x}_{t+1}) - f(\widehat{\boldsymbol{x}}_t)] \\ &= \boldsymbol{f}_t - \textstyle\int_0^1 f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)\widehat{\boldsymbol{p}}_t\,\mathrm{d}r - \int_0^1 f'(\widehat{\boldsymbol{x}}_t - s\beta_t\widehat{\boldsymbol{r}}_t)\beta_t\widehat{\boldsymbol{r}}_t\,\mathrm{d}s \\ &= -\widehat{\boldsymbol{r}}_t + \boldsymbol{B}_t\widehat{\boldsymbol{p}}_t - \textstyle\int_0^1 f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)\widehat{\boldsymbol{p}}_t\,\mathrm{d}s - \int_0^1 f'(\widehat{\boldsymbol{x}}_t - s\beta_t\widehat{\boldsymbol{r}}_t)\beta_t\widehat{\boldsymbol{r}}_t\,\mathrm{d}s \\ &= \underbrace{-\textstyle\int_0^1[\boldsymbol{I} + \beta_t f'(\widehat{\boldsymbol{x}}_t - s\beta_t\widehat{\boldsymbol{r}}_t)]\widehat{\boldsymbol{r}}_t\,\mathrm{d}s}_{:=\mathcal{L}_t} + \underbrace{\textstyle\int_0^1[\boldsymbol{B}_t - f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)]\widehat{\boldsymbol{p}}_t\,\mathrm{d}s}_{:=\mathcal{H}_t}. \end{aligned}$$

For the term $\mathcal{L}_t$, we have $\boldsymbol{I} + \beta_t f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{r}}_t) = (1-\beta_t)\boldsymbol{I} + \beta_t[\boldsymbol{I} + f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{r}}_t)]$. Note that $\boldsymbol{I} + f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{r}}_t) = g'(\boldsymbol{x}_t - s\widehat{\boldsymbol{r}}_t)$ and $\|g'(\boldsymbol{x})\| \leq \kappa, \forall\, \boldsymbol{x}$. Thus,

$$\|\mathcal{L}_t\| \leq \int_0^1 \|\boldsymbol{I} + \beta_t f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{r}}_t)\|\|\widehat{\boldsymbol{r}}_t\|\,\mathrm{d}s \leq [(1-\beta_t) + \beta_t\kappa]\theta_t\|\boldsymbol{f}_t\|. \tag{5.3}$$

The term $\mathcal{H}_t = \int_0^1[\boldsymbol{B}_t - f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)]\widehat{\boldsymbol{p}}_t\,\mathrm{d}s$ is equal to $\int_0^1[\boldsymbol{J}_t + \boldsymbol{E}_t - f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)]\widehat{\boldsymbol{p}}_t\,\mathrm{d}s$ since $\boldsymbol{B}_t\widehat{\boldsymbol{p}}_t = (\boldsymbol{J}_t + \boldsymbol{E}_t)\widehat{\boldsymbol{p}}_t$ for all $\boldsymbol{B}_t \in \mathbb{B}_t$. By telescoping, we have

$$\begin{aligned} \|\boldsymbol{J}_t + \boldsymbol{E}_t - f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t)\| &\leq \|f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t) - f'(\boldsymbol{x}_t)\| + \|f'(\boldsymbol{x}_t) - (\boldsymbol{J}_t + \boldsymbol{E}_t)\| \\ &\leq s\gamma\|\widehat{\boldsymbol{p}}_t\| + \|\boldsymbol{E}_t\| \\ &\leq s\gamma\sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\|\boldsymbol{f}_t\| + \|\boldsymbol{E}_t\|, \end{aligned}$$

where the second inequality holds because g' is Lipschitz continuous and the last inequality is due to the Lemma 5.1. Thus, we have

$$\begin{aligned} \|\mathcal{H}_t\| &\leq \int_0^1 \|f'(\boldsymbol{x}_t - s\widehat{\boldsymbol{p}}_t) - (\boldsymbol{J}_t + \boldsymbol{E}_t)\|\,\|\widehat{\boldsymbol{p}}_t\|\,\mathrm{d}s \\ &\leq \sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\left[\frac{\gamma}{2}\sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\|\boldsymbol{f}_t\| + \|\boldsymbol{E}_t\|\right]\|\boldsymbol{f}_t\|. \end{aligned} \tag{5.4}$$

The claim then follows from (5.3), (5.4), and the estimate $\|\boldsymbol{E}_t\| \le C_E\|\boldsymbol{f}_t\|$ given in Theorem 4.2 under Assumption 2.2.

∎

The bound given in (5.2) includes a linear term and a higher-order term, which is similar to the one for classical AA given in [29]. The higher-order term here is simpler because the history points in AAP are generated by Picard iterations and are well bounded around $\boldsymbol{x}_t$. Furthermore, the subsequent result is derived directly by applying (5.1) and $\|\boldsymbol{E}_t\| \le C_E\|\boldsymbol{f}_t\|$.

COROLLARY 5.3. *Under Assumptions 1.1 and 2.2, assume that $\boldsymbol{J}_t$ is invertible and $\boldsymbol{f}_t$ is small enough. Then,*

$$\begin{aligned}\|\boldsymbol{f}_{t+1}\| \le{}& [(1-\beta_t)+\beta_t\kappa]\,\|\boldsymbol{f}_t\| \\ &+ \sqrt{1-\theta_t^2}\,\left\|\boldsymbol{J}_t^{-1}\right\|\left[\tfrac{\gamma}{2}\sqrt{1-\theta_t^2}\left\|\boldsymbol{J}_t^{-1}\right\|\|\boldsymbol{f}_t\| + C_E\right]\|\boldsymbol{f}_t\|^2 + o(\|\boldsymbol{f}_t\|^2).\end{aligned}$$

Remark 5.4. The higher-order term $\mathcal{H}_t$ in the proof of Theorem 5.2 equals

$$\mathcal{H}_t = \boldsymbol{B}_t\widehat{\boldsymbol{p}}_t + f(\boldsymbol{x}_t - \widehat{\boldsymbol{p}}_t) - f(\boldsymbol{x}_t) = f\left(\textstyle\sum_{\ell=0}^m \alpha_t^\ell \boldsymbol{x}_t^\ell\right) - \sum_{\ell=0}^m \alpha_t^\ell f(\boldsymbol{x}_t^\ell),$$

which might be positive, potentially resulting in $\|\boldsymbol{f}_{t+1}\| > \|\boldsymbol{f}_t\|$. However, if f is convex and $\alpha_t^\ell \ge 0$ for $\ell = 0,\dots,m$, then $\mathcal{H}_t \le 0$ and the monotonic decay of $\|\boldsymbol{f}_t\|$ is guaranteed, i.e., $\|\boldsymbol{f}_{t+1}\| \le \|\boldsymbol{f}_t\|$. The condition $\boldsymbol{\alpha}_t \ge 0$ can be satisfied by imposing nonnegative constraints on the LS problem in (2.1), as in the EDIIS method [7], but the convergence might become slow.

Finally, we present the local convergence of AAP under the condition that g is a contractive mapping, i.e., $\kappa < 1$. Under this condition, g has a unique fixed point $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) = 0$. Since $\|g'(\boldsymbol{x}^*)\| \le \kappa < 1$, it follows that the Jacobian $f'(\boldsymbol{x}^*)$ is non-singular. Furthermore, an important property of contractive mappings is the relation between the residual and error. Specifically, for all $\boldsymbol{x} \in \mathbb{R}^d$, the error satisfies

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \le \|\boldsymbol{x} - g(\boldsymbol{x})\| + \|g(\boldsymbol{x}) - g(\boldsymbol{x}^*)\| \le \|f(\boldsymbol{x})\| + \kappa\|\boldsymbol{x} - \boldsymbol{x}^*\|,$$

which implies $\|\boldsymbol{x} - \boldsymbol{x}^*\| \le (1-\kappa)^{-1}\|f(\boldsymbol{x})\|$. Meanwhile, the residual satisfies

$$\|f(\boldsymbol{x})\| = \|(\boldsymbol{x} - \boldsymbol{x}^*) - [g(\boldsymbol{x}) - g(\boldsymbol{x}^*)]\| \le (1+\kappa)\|\boldsymbol{x} - \boldsymbol{x}^*\|.$$

THEOREM 5.5 (Local linear convergence of residual). *Assume Assumptions 1.1 and 2.2 hold, with $\beta_t \ge \beta > 0$ for some β, and the contraction constant κ in Assumption 1.1 satisfies $\kappa < 1$. Let $\boldsymbol{x}^*$ be the fixed-point solution, i.e., $\boldsymbol{x}^* = g(\boldsymbol{x}^*)$. If $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{x}^*$, then $\boldsymbol{f}_t$ converges linearly to $f(\boldsymbol{x}^*) = 0$, i.e., $\|\boldsymbol{f}_{t+1}\| \le \rho\|\boldsymbol{f}_t\|$ where ρ is a constant such that $\sup_t[(1-\beta_t)+\beta_t\kappa]\theta_t < \rho < 1$.*

Proof. Recall the bound on $\|\boldsymbol{f}_{t+1}\|$ given in (5.2). As $\kappa < 1$, $\beta_t \ge \beta > 0$ and $\theta_t \le 1$, the coefficient in the linear term satisfies $[(1-\beta_t)+\beta_t\kappa]\theta_t < 1$. Thus, there exists ρ that satisfies $\sup_t[(1-\beta_t)+\beta_t\kappa]\theta_t < \rho < 1$. Additionally, the coefficient in the higher-order term includes $\|\boldsymbol{B}_t^{-1}\|$ where $\boldsymbol{B}_t \in \mathbb{B}_t$ is invertible.

We first find the conditions to bound $\|\boldsymbol{B}_t^{-1}\|$. We choose a small $\epsilon > 0$. According to the continuity property of the Jacobian inverse [10, Lemma 2.1] and the fact that $f'(\boldsymbol{x}^*)$ is nonsingular, there exists $\delta_1 > 0$ such that $f'(\boldsymbol{x}_t) = \boldsymbol{J}_t$ is invertible and $\|f'(\boldsymbol{x}_t)^{-1} - f'(\boldsymbol{x}^*)^{-1}\| \le \epsilon$ provided that $\|\boldsymbol{x}_t - \boldsymbol{x}^*\| \le (1-\kappa)^{-1}\delta_1$, which can be satisfied by requiring $\|f(\boldsymbol{x}_t)\| \le \delta_1$. Furthermore, according to (5.1) and $\|\boldsymbol{E}_t\| \le C_E\|\boldsymbol{f}_t\|$ under

Assumptions 1.1 and 2.2, if $\boldsymbol{J}_t$ is invertible, there exists a sufficiently small $\delta_2 > 0$ such that $(\boldsymbol{J}_t + \boldsymbol{E}_t)$ is invertible and $\|(\boldsymbol{J}_t + \boldsymbol{E}_t)^{-1}\| \leq \|\boldsymbol{J}_t^{-1}\| + \epsilon$ when $\|\boldsymbol{f}_t\| \leq \delta_2$. Note that $\boldsymbol{J}_t + \boldsymbol{E}_t \in \mathbb{B}_t$. Therefore, if $\|\boldsymbol{f}_t\| \leq \min\{\delta_1, \delta_2\}$, there exists $\boldsymbol{B}_t \in \mathbb{B}_t$ which is invertible and satisfies $\|\boldsymbol{B}_t^{-1}\| \leq M^*$ where $M^* := \|f'(\boldsymbol{x}^*)^{-1}\| + 2\epsilon$.

Next, we find conditions to guarantee the monotonic decay of the residual, i.e., $\|\boldsymbol{f}_{t+1}\| \leq \rho\|\boldsymbol{f}_t\|$. According to (5.2),

$$\begin{aligned}\|\boldsymbol{f}_{t+1}\| &\leq [(1-\beta_t) + \beta_t\kappa]\,\theta_t\|\boldsymbol{f}_t\| + \sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\|\left[\frac{\gamma}{2}\sqrt{1-\theta_t^2}\|\boldsymbol{B}_t^{-1}\| + C_E\right]\|\boldsymbol{f}_t\|^2 \\ &\leq [(1-\beta_t) + \beta_t\kappa]\,\theta_t\|\boldsymbol{f}_t\| + \|\boldsymbol{B}_t^{-1}\|\left[\frac{\gamma}{2}\|\boldsymbol{B}_t^{-1}\| + C_E\right]\|\boldsymbol{f}_t\|^2 \\ &= \left[[(1-\beta_t) + \beta_t\kappa]\theta_t + \|\boldsymbol{B}_t^{-1}\|\left(\frac{\gamma}{2}\|\boldsymbol{B}_t^{-1}\| + C_E\right)\|\boldsymbol{f}_t\|\right]\|\boldsymbol{f}_t\|.\end{aligned}$$

We have that if $\|\boldsymbol{f}_t\| \leq \min\{\delta_1, \delta_2\}$,

$$\|\boldsymbol{f}_{t+1}\| \leq \left[[(1-\beta_t) + \beta_t\kappa]\theta_t + M^*\left(\tfrac{\gamma}{2}M^* + C_E\right)\|\boldsymbol{f}_t\|\right]\|\boldsymbol{f}_t\|. \qquad \square$$

It follows that $\|\boldsymbol{f}_{t+1}\| \leq \rho\|\boldsymbol{f}_t\|$ when $\|\boldsymbol{f}_t\| \leq \delta_3$ where $\delta_3 = \frac{\rho - \sup_t[(1-\beta_t)+\beta_t\kappa)]\theta_t}{M^*\left(\frac{\gamma}{2}M^* + C_E\right)} > 0$. Thus, if $\|\boldsymbol{f}_t\| \leq \min\{\delta_1, \delta_2, \delta_3\}$, we have $\|\boldsymbol{f}_{t+1}\| \leq \rho\|\boldsymbol{f}_t\|$.

Since $\rho < 1$, if $\|\boldsymbol{f}_0\| \leq \min\{\delta_1, \delta_2, \delta_3\}$, then $\|\boldsymbol{f}_{t+1}\| \leq \rho\|\boldsymbol{f}_t\|$ for all $t \geq 0$. This condition, $\|\boldsymbol{f}_0\| \leq \min\{\delta_1, \delta_2, \delta_3\}$, is satisfied when $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{x}^*$ due to the continuity of f. This concludes the proof.

$\square$

We have established the linear convergence of the residual, $\|\boldsymbol{f}_{t+1}\| \leq \rho\|\boldsymbol{f}_t\|$. The convergence of the iterates $\{\boldsymbol{x}_t\}$ follows as

$$\|\boldsymbol{x}_t - \boldsymbol{x}^*\| \leq (1+\kappa)\|\boldsymbol{f}_t\| \leq (1+\kappa)\rho^t\|\boldsymbol{f}_0\|.$$

Remark 5.6. AAP can be regarded as an inexact Newton method [10] whose search direction satisfies $\|\boldsymbol{J}_t\boldsymbol{p}_t - \boldsymbol{f}_t\| \leq \eta_t\|\boldsymbol{f}_t\|$ where $\boldsymbol{p}_t := \boldsymbol{x}_t - \boldsymbol{x}_{t+1}$. This η_t is called the forcing term of the inexact Newton method and assesses how well $\boldsymbol{p}_t$ approximates the exact Newton direction in each iteration. For any invertible $\boldsymbol{B}_t \in \mathbb{B}_t$, we have

$$\|\boldsymbol{J}_t\boldsymbol{p}_t - \boldsymbol{f}_t\| \leq \left[[(1-\beta_t) + \beta_t\kappa]\theta_t + \sqrt{1-\theta_t^2}\|\boldsymbol{E}_t\|\|\boldsymbol{B}_t^{-1}\|\right]\|\boldsymbol{f}_t\|.$$

Since $\boldsymbol{p}_t = \widehat{\boldsymbol{p}}_t + \beta_t\widehat{\boldsymbol{r}}_t$, the above bound comes from the decomposition of $\boldsymbol{J}_t\boldsymbol{p}_t - \boldsymbol{f}_t = (I + \beta_t\boldsymbol{J}_t)\widehat{\boldsymbol{r}}_t - \boldsymbol{E}_t\widehat{\boldsymbol{p}}_t$. Thus, the forcing term of AAP satisfies $\eta_t = [(1-\beta_t) + \beta_t\kappa]\theta_t + \sqrt{1-\theta_t^2}\|\boldsymbol{E}_t\|\|\boldsymbol{B}_t^{-1}\|$. If $\|\boldsymbol{E}_t\| \leq C_E\|\boldsymbol{f}_t\|$, then $\eta_t \leq [(1-\beta_t) + \beta_t\kappa]\theta_t + \mathcal{O}(\|\boldsymbol{f}_t\|)$. It follows that, when $\|\boldsymbol{f}_t\|$ approaches 0, η_t converges to $[(1-\beta_t) + \beta_t\kappa]\theta_t$, with the optimization gain θ_t converging to the Jacobian-GMRES gain as shown in Theorem 4.8. One sufficient condition to guarantee the local convergence of AAP, as an inexact Newton method, is that $\sup_t \eta_t < 1$ [10], which can be ensured by the same assumptions as in Theorem 5.5.

6. Numerical results. In this section, we present numerical experiments to demonstrate the performance of the AAP(m) against other algorithms, including:

- Picard: Picard iteration with $\boldsymbol{x}_{t+1} = g(\boldsymbol{x}_t)$.
- AA(m) [37]: Classical Anderson acceleration with a fixed window size m, taking an AA step at each iteration.

- resAA(m) [26]: Restarted Anderson acceleration, where an AA step is taken at each iteration. The number of history iterates used increases until a given threshold m is reached, after which the number of history iterates used is reset to 0.
- Newton-GMRES[8]: Solves $\boldsymbol{J}_t \boldsymbol{p}_t^{\mathrm{N}} = \boldsymbol{f}_t$ using GMRES(m) at each global iteration, updating $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \boldsymbol{p}_t^{\mathrm{N}}$. There is no line search applied.

Note that AAP(m) evaluates $m+1$ Picard iterations in each global iteration, whereas Picard, AA(m), and resAA(m) evaluate one Picard iteration per iteration. The performance of each algorithm may vary depending on the specific problem and the parameters. The examples provided are chosen to better illustrate the features of AAP. The code can be accessed at https://github.com/xue1993/AAP.git.

6.1. Logistic regression. The first example is a nonlinear fixed-point problem derived from the gradient descent (GD) algorithm to minimize a function $h(\boldsymbol{x})$. The GD step with stepsize η is defined by $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla h(\boldsymbol{x}_t)$, which can be reformulated as a fixed-point problem:

$$\boldsymbol{x} = g(\boldsymbol{x}) = \boldsymbol{x} - \eta \nabla h(\boldsymbol{x}).$$

Here, h is the loss function of a regularized logistic regression problem. Specifically,

$$h(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{x}^\top \boldsymbol{v}_i)) + \frac{\mu}{2} \|\boldsymbol{x}\|^2,$$

where $\boldsymbol{v}_i \in \mathbb{R}^d$ is a feature vector and $y_j \in \{-1, 1\}$ is the corresponding response, $\boldsymbol{x}$ represents the weights, and μ is the regularization parameter. We use two classical datasets: `w8a` ($n = 49,749$ and $d = 300$) and `covtype` ($n = 581,000$ and $d = 54$). Both datasets are available at the LIBSVM website [6]. The function h is strongly convex. For all examples, we use $\eta = 1$ and ensure that g is a contractive mapping. The initial point is $\boldsymbol{x}_0 = 0$, and the global minimizer is denoted as $\boldsymbol{x}^*$.

Figure 1 shows the results of different algorithms on the `w8a` and `covtype` datasets. From Figure 1(a) about `w8a` dataset, we observe that AAP outperforms other algorithms. The convergence plot for AAP in terms of the number of Picard iterations exhibits a stair-step effect, with flat regions in each global iteration from Picard steps and rapid decay from the AA step marked as red dot. Meanwhile, the errors of resAA and AA oscillate, especially in the first few iterations. This oscillation is also observed in the landscape of function h and optimization path plots in Figure 1(c), where AA and resAA overshoot around the global minimizer, slowing their convergence. This phenomenon can be explained as the approximate Jacobian is inaccurate when constructed using points that are far apart from each other.

For the `covtype` dataset, both AAP and resAA perform well, as shown in Figure 1(b). The optimization paths of all AA algorithms in Figure 1(d) also show less oscillation on this dataset. The third row in Figure 1 illustrates the impact of varying m on different AA algorithms. It is observed that increasing m does not necessarily improve the convergence rate for all AA variants. Compared to AA and resAA, the performance of AAP demonstrates less oscillation across different values of m. Nonetheless, the convergence rate of AAP tends to decrease over iterations. This is likely due to machine round-off error, as the points generated by Picard are very close to each other, making it difficult to retain useful curvature information in $\boldsymbol{Y}_t$.

The first row of Figure 2 presents the convergence of the residual $\|f(\boldsymbol{x}_t)\|$ in terms of global iterations t for AAP with varying m. In these examples, we chose $\gamma = 0.0001$ on the `w8a` dataset. With this small γ, the least singular value of the

Hessian of h is 0.0001, and the contraction constant κ of g is 0.9999. We can see that the convergence of the Picard method is slow. However, the performance of AAP is close to the Newton-GMRES method, both significantly improving the convergence of Picard iteration. From subplots (b) and (c), it is evident that the residual $\|f(\boldsymbol{x}_t)\|$ decrease monotonically when the residual is sufficiently small. The second row of Figure 2 shows the optimization gain θ_t of AAP in terms of global iterations t. By comparing each pair, for example, subplot (a) of the convergence plot and subplot (d) of the θ_t plot both in red, we observe that the convergence rate of AAP is largely influenced by θ_t, and the convergence is faster with smaller θ_t.

Additionally, subplots (d) and (e) of Figure 2 demonstrate that the optimization gain θ_t of AAP is close to the Jacobian-GMRES gain. However, for larger values of m (e.g., $m = 7$ as shown in subplot (f)), the Jacobian-GMRES gain is no longer a good estimate of the optimization gain. This discrepancy is likely due to large constant in the bound (4.22), and/or machine round-off errors.

6.2. Nonnegative matrix factorization. The second example is the Nonnegative Matrix Factorization (NMF) problem, which has been applied in fields such as text mining, image processing, and bioinformatics [11]. NMF decomposes a matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ into two nonnegative matrices $\boldsymbol{W} \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{H} \in \mathbb{R}^{r \times d_2}$, i.e., $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$ and $\boldsymbol{W} \geq 0, \boldsymbol{H} \geq 0$. The optimization problem associated with NMF can be formulated as

$$\min_{\boldsymbol{W},\boldsymbol{H}} \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_F^2 \quad \text{subject to} \quad \boldsymbol{W} \geq 0,\ \boldsymbol{H} \geq 0,$$

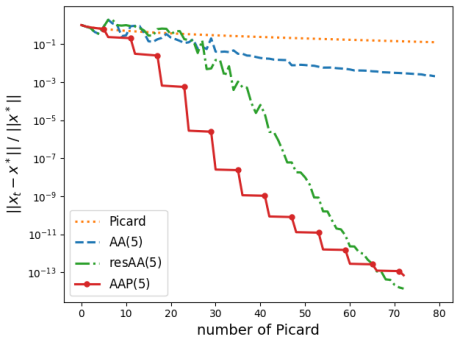
where $\|\cdot\|_F$ denotes the Frobenius norm. A widely used method to solve this problem is the alternating nonnegative least-squares (ANNLS) method: starting from $\boldsymbol{W}_0 \geq 0$, one generates a sequence of $(\boldsymbol{W}_t, \boldsymbol{H}_t)$ by alternately solving the standard nonnegatively constrained linear least-squares problems:

$$\boldsymbol{H}_t = \arg\min_{\boldsymbol{H} \geq 0} \|\boldsymbol{A} - \boldsymbol{W}_{t-1}\boldsymbol{H}\|_F \quad \text{and} \quad \boldsymbol{W}_t = \arg\min_{\boldsymbol{W} \geq 0} \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}_t\|_F.$$

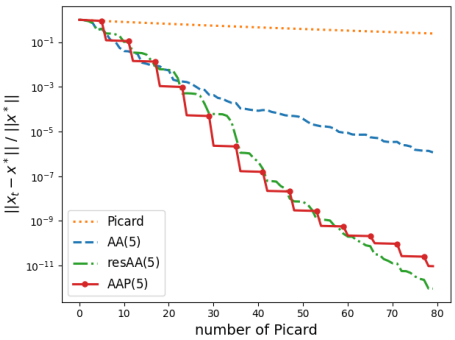
Additionally, a normalization of $\boldsymbol{W}_t$ is applied at the beginning of each iteration. Following [37], we consider ANNLS to be a fixed-point iteration through the assignment $(\boldsymbol{W}_t, \boldsymbol{H}_t) \to (\boldsymbol{W}_{t+1}, \boldsymbol{H}_{t+1})$. We perform the AA step on this problem by solving a constrained LS based on the vectorized variable, and we truncate $(\boldsymbol{W}_t, \boldsymbol{H}_t)$ after the AA step to guarantee nonnegativity of the sequence.

We construct synthetic NMF problems by generating random matrices $\boldsymbol{W} \in \mathbb{R}^{300 \times r}$ and $\boldsymbol{H} \in \mathbb{R}^{r \times 50}$, and then take $\boldsymbol{A} = \boldsymbol{W}\boldsymbol{H} \in \mathbb{R}^{300 \times 50}$. Consequently, the minimum of the NMF objective function is zero. The initial point $\boldsymbol{W}_0$ is chosen randomly. Figure 3 presents the results of different algorithms with the same initialization. It is evident that all AA variants, including AAP, AA, and resAA, significantly improve convergence compared to ANNLS. Since performance of each algorithm can be influenced by different initializations, we also provide the median and interquartile range plots of 15 runs.

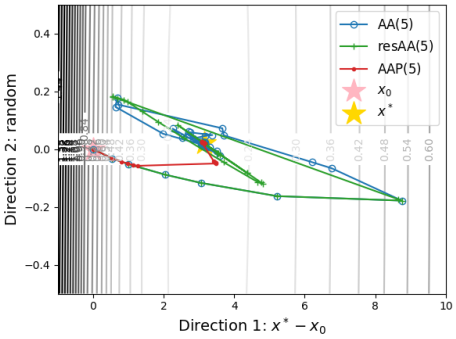
7. Conclusion. This work focuses on the analysis of the Alternating Anderson Picard method, in which an Anderson acceleration step is periodically applied after a number of Picard iterations. We established the equivalence between AAP and multisecant-GMRES and explored the relation between AAP and Newton-GMRES, with particular emphasis on convergence analysis. We proved that the AAP residual converges locally at an improved rate when compared to Picard iteration. Numerically, this method has demonstrated efficiency and robustness in our evaluations and



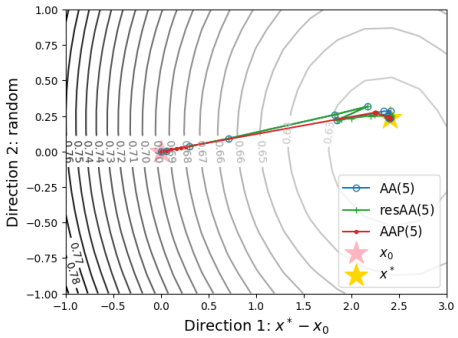
(a) `w8a`($m = 5$)



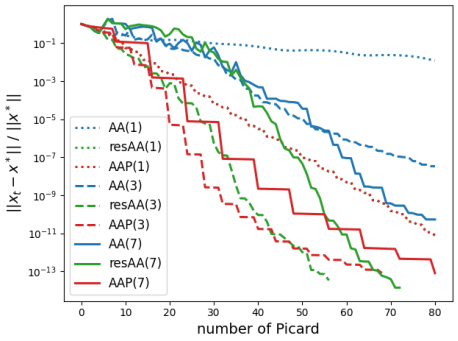
(b) `covtype`($m = 5$)



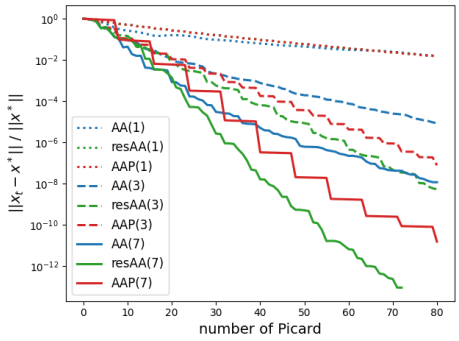
(c) `w8a`: optimization paths($m = 5$)



(d) `covtype`: optimization paths($m = 5$)

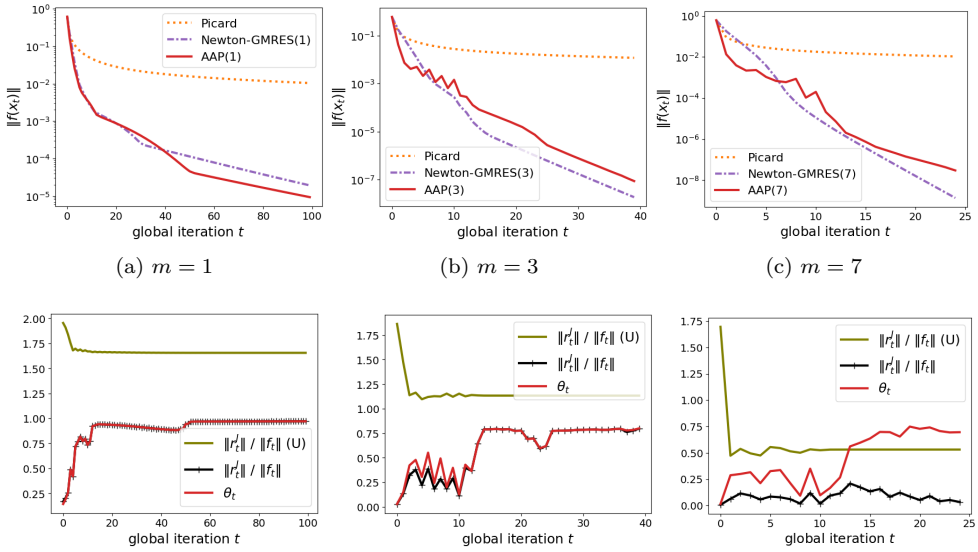


(e) `w8a`: varying m



(f) `covtype`: varying m

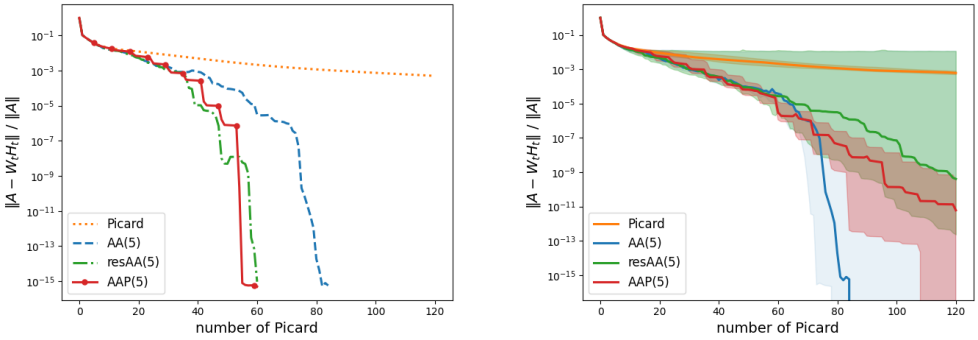
FIG. 1. *Comparison of different algorithms applied to logistic regression with* $\mu = 0.01$ *on both* `w8a` *and* `covtype` *datasets. First row: number of Picard iterations versus relative error. Second column: Landscape of the function* h *and the optimization paths of different algorithms. The true domain dimension of* h *is* 300 *for the* `w8a` *dataset and* 54 *for the* `covtype` *dataset. To visualize* h*, we show the level set of* h *along two directions starting from the initial point* $\boldsymbol{x}_0 = 0$*: normalized* $\boldsymbol{x}_0 - \boldsymbol{x}^*$ *and a random direction. The optimization paths are also projected onto these two directions, with each marker representing one Picard iteration. Due to the inefficiency of the history data points, the first* m *iterations of AA*(m) *are equivalent to those of resAA*(m)*, with their plots overlapping. Third row: effect of varying* m *on different AA algorithms. It is noteworthy that when* $m = 1$*, AAP*(1) *is equivalent to resAA*(1)*, resulting in the vanishing plot of resAA*(1)*.*



(a) $m = 1$ (b) $m = 3$ (c) $m = 7$

(d) $m = 1$ (e) $m = 3$ (f) $m = 7$

FIG. 2. *Convergence of AAP on logistic regression with* $\gamma = 0.0001$ *on the* `w8a` *dataset. The first row shows global iterations versus residual norm* $\|f(\boldsymbol{x}_t)\|$ *for different algorithms, including Picard and Newton-GMRES as reference. In each global iteration, Picard performs* $m + 1$ *steps to have a fair comparison with AAP. The second row presents global iterations versus optimization gain* θ_t*. We also display the value of the Jacobian-GMRES gain* $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$*, as defined in* (4.21)*. The Jacobian-GMRES gain* $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ *(U) represents the theoretical upper bound of* $\|\boldsymbol{r}_t^J\|/\|\boldsymbol{f}_t\|$ *given in* (4.24)*. Here the Jacobian is the Hessian of* h*, which is symmetric and positive definite.*



(a) (b)

FIG. 3. *Comparison of different algorithms on NMF problems with* $r = 4$*. All plot shows number of Picard iterations versus relative residual* $\|\boldsymbol{A}-\boldsymbol{W}_t\boldsymbol{H}_t\|_F/\|\boldsymbol{A}\|_F$*. (a): Different algorithms with the same initialization. (b): Median and interquartile range plots of 15 runs with random initializations.*

is comparable to the Newton-GMRES method without accessing the Jacobian. This underscores the potential of AAP as a practical alternative in scenarios where computational resources are limited or when explicit Jacobian computation is impractical.

Acknowledgments. X.F. would like to express her sincere gratitude to Professor Zhaojun Bai for his valuable suggestions and insightful feedback, which contributed

to the improvement of this work.

Appendix A. Uniform boundedness of $\mathrm{cond}(\boldsymbol{S}_t)$**.** The uniform boundedness of $\mathrm{cond}(\boldsymbol{S}_t)$ is important for the bound of $\boldsymbol{E}_t$ as shown in Theorem 4.2. Thus, in this section, we discuss $\mathrm{cond}(\boldsymbol{S}_t)$ when the AAP residual $\boldsymbol{f}_t$ approaches 0.

We have shown that $\boldsymbol{S}_t$ converges to $\boldsymbol{G}_t$ as $\|\boldsymbol{f}_t\|$ approaches zero. Consequently, $\mathrm{cond}(\boldsymbol{G}_t)$ can serve as an estimator for $\mathrm{cond}(\boldsymbol{S}_t)$ when $\|\boldsymbol{f}_t\|$ is sufficiently small. The matrix $\boldsymbol{G}_t$ is a Krylov matrix. The condition number and the least singular value of Krylov matrices are active research areas [5, 9, 4]. Generally speaking, $\boldsymbol{G}_t$ can be ill-conditioned even for a moderate number of columns [5].

In the following, we discuss $\mathrm{cond}(\boldsymbol{G}_t)$ in a simple case where $g'(\boldsymbol{x}_t)$ is symmetric, allowing $\boldsymbol{G}_t$ to be decomposed into the product of a diagonal matrix and a Vandermonde matrix.

LEMMA A.1. *Let $g'(\boldsymbol{x}_t)$be a symmetric matrix with eigendecomposition $g'(\boldsymbol{x}_t) = \boldsymbol{Q}^T\boldsymbol{\Lambda}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Denote $\boldsymbol{Q}\frac{\boldsymbol{f}_t}{\|\boldsymbol{f}_t\|} = [a_1, a_2, \ldots, a_d]^T$. If $\min_i |a_i| > 0$, then*

$$\mathrm{cond}(\boldsymbol{G}_t) \le \frac{\sqrt{m}}{(\min_i |a_i|)\ \sigma_{\min}\left(\boldsymbol{V}_m(\lambda_1, \lambda_2, \ldots, \lambda_d)\right)}, \tag{A.1}$$

where $\boldsymbol{V}_m(\lambda_1, \ldots, \lambda_d)$ is a $d \times m$ rectangular Vandermonde matrix.

Proof. Denote $\boldsymbol{K} = g'(\boldsymbol{x}_t)$. Since $\boldsymbol{K} = \boldsymbol{Q}^T\boldsymbol{\Lambda}\boldsymbol{Q}$, we have $\boldsymbol{K}^{(\ell)} = \boldsymbol{Q}^T\boldsymbol{\Lambda}^{(\ell)}\boldsymbol{Q}$. By the definition of $\boldsymbol{G}_t$, we have

$$\begin{aligned}
\boldsymbol{Q}\boldsymbol{G}_t &= \boldsymbol{Q}[\boldsymbol{f}_t, \boldsymbol{K}^{(1)}\boldsymbol{f}_t, \boldsymbol{K}^{(2)}\boldsymbol{f}_t, \ldots, \boldsymbol{K}^{(m-1)}\boldsymbol{f}_t] \\
&= [\boldsymbol{Q}\boldsymbol{f}_t, \boldsymbol{Q}\boldsymbol{K}^{(1)}\boldsymbol{f}_t, \boldsymbol{Q}\boldsymbol{K}^{(2)}\boldsymbol{f}_t, \ldots, \boldsymbol{Q}\boldsymbol{K}^{(m-1)}\boldsymbol{f}_t] \\
&= [\boldsymbol{Q}\boldsymbol{f}_t, \boldsymbol{\Lambda}^{(1)}\boldsymbol{Q}\boldsymbol{f}_t, \boldsymbol{\Lambda}^{(2)}\boldsymbol{Q}\boldsymbol{f}_t, \ldots, \boldsymbol{\Lambda}^{(m-1)}\boldsymbol{Q}\boldsymbol{f}_t].
\end{aligned}$$

It follows from $\boldsymbol{Q}\frac{\boldsymbol{f}_t}{\|\boldsymbol{f}_t\|} = [a_1, a_2, \ldots, a_d]^T$ that

$$\begin{aligned}
\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Q}\boldsymbol{G}_t &= \begin{bmatrix} a_1 & a_1\lambda_1 & a_1\lambda_1^2 & \cdots & a_1\lambda_1^{m-1} \\ a_2 & a_2\lambda_2 & a_2\lambda_2^2 & \cdots & a_2\lambda_2^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_d & a_d\lambda_d & a_d\lambda_d^2 & \cdots & a_d\lambda_d^{m-1} \end{bmatrix} \\
&= \mathrm{diag}(a_1, a_2, \ldots, a_d)\boldsymbol{V}_m(\lambda_1, \lambda_2, \ldots, \lambda_d).
\end{aligned}$$

Denote $\boldsymbol{D} = \mathrm{diag}(a_1, a_2, \ldots, a_d)$ and $\boldsymbol{V} = \boldsymbol{V}_m(\lambda_1, \lambda_2, \ldots, \lambda_d)$. Since $\boldsymbol{Q}$ is an orthogonal matrix, we have

$$\sigma_{\min}\left(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right) = \sigma_{\min}\left(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{Q}\boldsymbol{G}_t\right) = \sigma_{\min}\left(\boldsymbol{D}\boldsymbol{V}\right).$$

Assume $\sigma_{\min}(\boldsymbol{D}\boldsymbol{V}) = \|\boldsymbol{D}\boldsymbol{V}\bar{\boldsymbol{x}}\|$ with $\|\bar{\boldsymbol{x}}\|$=1. Since $\sigma_{\min}(\boldsymbol{V}) = \min_{\boldsymbol{x}\neq 0}\frac{\|\boldsymbol{V}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$,

$$\sigma_{\min}(\boldsymbol{V}) \le \frac{\|\boldsymbol{V}\bar{\boldsymbol{x}}\|}{\|\bar{\boldsymbol{x}}\|} = \frac{\|\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{V}\bar{\boldsymbol{x}}\|}{\|\bar{\boldsymbol{x}}\|} \le \|\boldsymbol{D}^{-1}\|\,\|\boldsymbol{D}\boldsymbol{V}\bar{\boldsymbol{x}}\| = \frac{1}{\min_i |a_i|}\|\boldsymbol{D}\boldsymbol{V}\bar{\boldsymbol{x}}\|.$$

Thus, $\sigma_{\min}\left(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right) \ge (\min_i |a_i|)\ \sigma_{\min}\left(\boldsymbol{V}\right)$. On the other hand, $\sigma_{\max}\left(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right) = \|\left(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right)\| \le \|\left(\frac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right)\|_F \le \sqrt{m}$. Combining these with

$$\mathrm{cond}(\boldsymbol{G}_t) = \mathrm{cond}(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t) = \sigma_{\max}\left(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right)/\sigma_{\min}\left(\tfrac{1}{\|\boldsymbol{f}_t\|}\boldsymbol{G}_t\right),$$

gives the result. ∎

In Lemma A.1, the condition that $\min_i |a_i| > 0$ is equivalent to the vector $\boldsymbol{f}_t$ not being parallel to any of the eigenvectors of the matrix $g'(\boldsymbol{x}_t)$. The least singular value of the Vandermonde matrix $\boldsymbol{V}_m(\lambda_1, \ldots, \lambda_d)$ depends on the distribution of its seeds λ_i [4, 21, 9]. Theorem 6 of [9] shows that when the largest seed $|\lambda_1| < 1$, the least singular value

$$\sigma_{\min}(\boldsymbol{V}_m(\lambda_1, \ldots, \lambda_d)) \leq \sigma_{\max}(\boldsymbol{V}_m(\lambda_1, \ldots, \lambda_d))|\lambda_1|^{(m)},$$

which decays exponentially with respect to m. It indicates that a large m may lead to a significant increase in $\text{cond}(\boldsymbol{G}_t)$, which in turn can cause a dramatic increase in $\|\boldsymbol{E}_t\|$. In addtion, since $\boldsymbol{S}_t \to \boldsymbol{G}_t$, and $\boldsymbol{Y}_t \to \boldsymbol{G}_t$, large m may cause instability of the least square problem in the AA step. However, when $\boldsymbol{f}_t$ is small, machine round-off errors will also affect the numerical value of $\text{cond}(\boldsymbol{S}_t)$ and $\text{cond}(\boldsymbol{Y}_t)$.

REFERENCES

[1] H.-B. An and Z.-Z. Bai, *A globally convergent newton-gmres method for large sparse systems of nonlinear equations*, Applied Numerical Mathematics, 57 (2007), pp. 235–252.

[2] D. G. Anderson, *Iterative procedures for nonlinear integral equations*, Journal of the ACM (JACM), 12 (1965), pp. 547–560.

[3] A. S. Banerjee, P. Suryanarayana, and J. E. Pask, *Periodic Pulay method for robust and efficient convergence acceleration of self-consistent field iterations*, Chemical Physics Letters, 647 (2016), pp. 31–35.

[4] F. S. Bazán, *Conditioning of rectangular Vandermonde matrices with nodes in the unit disk*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 679–693.

[5] B. Beckermann, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numerische Mathematik, 85 (2000), pp. 553–577.

[6] C.-C. Chang and C.-J. Lin, *Libsvm: a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology (TIST), 2 (2011), pp. 1–27.

[7] X. Chen and C. T. Kelley, *Convergence of the EDIIS algorithm for nonlinear equations*, SIAM Journal on Scientific Computing, 41 (2019), pp. A365–A379.

[8] R. Choquet and J. Erhel, *Some convergence results for the Newton-GMRES algorithm*, PhD thesis, INRIA, 1993.

[9] A. Dax, *The numerical rank of Krylov matrices*, Linear Algebra and Its Applications, 528 (2017), pp. 185–205.

[10] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact newton methods*, SIAM Journal on Numerical analysis, 19 (1982), pp. 400–408.

[11] L. Eldén, *Matrix methods in data mining and pattern recognition*, SIAM, 2019.

[12] C. Evans, S. Pollock, L. G. Rebholz, and M. Xiao, *A proof that Anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically)*, SIAM Journal on Numerical Analysis, 58 (2020), pp. 788–810.

[13] H.-r. Fang and Y. Saad, *Two classes of multisecant methods for nonlinear acceleration*, Numerical linear algebra with applications, 16 (2009), pp. 197–221.

[14] X. Feng, *Anderson Acceleration and Dynamic Optimal Transport in Optimization: Theoretical Analysis, Algorithms, and Applications in Machine Learning*, University of California, Davis, 2024.

[15] M. Geist and B. Scherrer, *Anderson acceleration for reinforcement learning*, arXiv preprint arXiv:1809.09501, (2018).

[16] G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU press, 2013.

[17] A. Greenbaum and Z. Strakos, *Matrices that generate the same Krylov residual spaces*, Springer, 1994.

[18] A. Han, B. Mishra, P. Jawanpuria, and J. Gao, *Riemannian accelerated gradient methods via extrapolation*, in International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 1554–1585.

[19] C. T. Kelley, *Numerical methods for nonlinear equations*, Acta Numerica, 27 (2018), pp. 207–287.

[20] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton–Krylov methods: a survey of approaches and applications*, Journal of Computational Physics, 193 (2004), pp. 357–397.
[21] S. Kunis and D. Nagel, *On the condition number of Vandermonde matrices with pairs of nearly-colliding nodes*, Numerical Algorithms, 87 (2021), pp. 473–496.
[22] 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*, Advances in Water Resources, 38 (2012), pp. 92–101.
[23] M. Lupo Pasini, *Convergence analysis of Anderson-type acceleration of Richardson's iteration*, Numerical Linear Algebra with Applications, 26 (2019), p. e2241.
[24] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 1999.
[25] W. Ouyang, Y. Liu, and A. Milzarek, *Descent properties of an Anderson accelerated gradient method with restarting*, SIAM Journal on Optimization, 34 (2024), pp. 336–365.
[26] W. Ouyang, Y. Peng, Y. Yao, J. Zhang, and B. Deng, *Anderson acceleration for nonconvex ADMM based on Douglas-Rachford splitting*, in Computer Graphics Forum, vol. 39, Wiley Online Library, 2020, pp. 221–239.
[27] Y. Peng, B. Deng, J. Zhang, F. Geng, W. Qin, and L. Liu, *Anderson acceleration for geometry optimization and physics simulation*, ACM Transactions on Graphics, 37 (2018), p. Article 42.
[28] S. Pollock, L. Rebholz, and M. Xiao, *Anderson-accelerated convergence of Picard iterations for incompressible Navier-Stokes equations*, SIAM Journal on Numerical Analysis, 57 (2019), pp. 615–637.
[29] S. Pollock and L. G. Rebholz, *Anderson acceleration for contractive and noncontractive operators*, IMA Journal of Numerical Analysis, 41 (2021), pp. 2841–2872.
[30] S. Pollock and L. G. Rebholz, *Filtering for Anderson acceleration*, SIAM Journal on Scientific Computing, 45 (2023), pp. A1571–A1590.
[31] P. P. Pratapa, P. Suryanarayana, and J. E. Pask, *Anderson acceleration of the Jacobi iterative method: An efficient alternative to Krylov methods for large, sparse linear systems*, Journal of Computational Physics, 306 (2016), pp. 43–54.
[32] J. A. Sifuentes, M. Embree, and R. B. Morgan, *GMRES convergence for perturbed coefficient matrices, with application to approximate deflation preconditioning*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 1066–1088.
[33] H. D. Sterck and Y. He, *On the asymptotic linear convergence speed of Anderson acceleration, Nesterov acceleration, and nonlinear GMRES*, SIAM Journal on Scientific Computing, 43 (2021), pp. S21–S46.
[34] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM review, 19 (1977), pp. 634–662.
[35] P. Suryanarayana, P. P. Pratapa, and J. E. Pask, *Alternating Anderson–Richardson method: An efficient alternative to preconditioned Krylov methods for large, sparse linear systems*, Computer Physics Communications, 234 (2019), pp. 278–285.
[36] A. Toth and C. T. Kelley, *Convergence analysis for Anderson acceleration*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 805–819.
[37] H. F. Walker and P. Ni, *Anderson acceleration for fixed-point iterations*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 1715–1735.
[38] D. Wang, Y. He, and H. De Sterck, *On the asymptotic linear convergence speed of Anderson acceleration applied to ADMM*, Journal of Scientific Computing, 88 (2021), p. 38.
[39] J. Zhang, B. O'Donoghue, and S. Boyd, *Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations*, SIAM Journal on Optimization, 30 (2020), pp. 3170–3197.