·前沿聚焦：优化技术与应用专栏·　　DOI：10. 13471/j. cnki. acta. snus. 2023A035

关于 Anderson 混合的研究进展*

包承龙[1]，韦福超[2]

1. 清华大学丘成桐数学科学中心，北京 100084
2. 清华大学计算机科学与技术系，北京 100084

摘　要： Anderson 混合是一种经典的外推方法，它能利用历史迭代信息加速定点迭代的收敛，在科学计算和机器学习中得到了成功的应用. 由于 Anderson 混合在实践中经常表现出优越的数值性能，在各类应用中围绕 Anderson 混合的算法设计和理论分析成为近几年的研究热点. 本文综述关于 Anderson 混合的研究进展，重点介绍基于 Anderson 混合的新算法.

关键词： Anderson 混合；定点迭代；Krylov 子空间方法；拟 Newton 法

中图分类号： O221.2　　**文献标志码：** A　　**文章编号：** 2097－0137（2023）05－0059－08

Research advance on Anderson mixing

BAO Chenglong[1]，*WEI Fuchao*[2]

1. Yau Mathematical Sciences Center，Tsinghua University，Beijing 100084，China
2. Department of Computer Science and Technology，Tsinghua University，Beijing 100084，China

Abstract： Anderson mixing is a classical extrapolation method. It can make use of the information in historical iterations to accelerate the convergence of fixed-point iterations，and has been successfully applied in scientific computing and machine learning. Since Anderson mixing often exhibits superior numerical performance in practice，the algorithm design and theoretical analysis around Anderson mixing in various applications have become hot topics in recent years. This article reviews the research advance on Anderson mixing，and highlights new algorithms based on Anderson mixing.

Key words： Anderson mixing；fixed-point iteration；Krylov subspace method；quasi-Newton method

由 Anderson（1965）提出的 Anderson 混合（AM，Anderson mixing）最早被用于非线性积分方程的计算，现已成为加速定点迭代的一种经典算法. 在量子化学领域，AM 又被称为 Pulay 混合（Pulay，1980）或 DIIS 方法（Rohwedder et al.，2011），在自洽场迭代的加速中发挥了重要作用（Arora et al.，2017）. AM 本质是一种外推方法（Brezinski et al.，2018；Anderson，2019），它通过对历史迭代外推，生成与定点迭代不同的新迭代序列. 由于 AM 通常能够显著减少迭代过程收敛到定点的迭代次数，因此和定点迭代相较，当定点算子的计算开销很大时，AM 算法能够节省大量的计算时间. 所以，AM 在科学计算中也常被称为 Anderson 加速（Walker et al.，2011）. 近几年来，得益于其实现的简易性和优异的数值表现，AM 在科学计算和机器学习领域得到了广泛的关注，研究者们成功地将 AM 应用于各种定点问题的求解当中，例如，解 Navier-Stokes 方程（Pollock et al.，2019）、解地震波反演问题（Yang，2021）、加速机器学习中的 EM 算法（Henderson et al.，2019）、强化学习训练（Sun et al.，2021）等. 此外，AM 的理论性质也引起计算数学界的极大兴趣，对 AM 在定点问题中的收敛性给出理论分析仍是目前一个重要的研究问题（Toth et al.，2015；Evans et al.，2020；

* **收稿日期：** 2023－05－11　　**录用日期：** 2023－06－21　　**网络首发日期：** 2023－08－30
基金项目： 国家重点研发计划（2021YFA1001300）；国家自然科学基金（12271291）
作者简介： 包承龙（1989 年生），男；**研究方向：** 图像处理模型与优化算法；E-mail：clbao@mail.tsinghua.edu.cn

Bian et al.,2021).

先简要介绍AM的基本迭代格式. 考虑定点问题

$$\boldsymbol{x} = g(\boldsymbol{x}), \tag{1}$$

其中$\boldsymbol{x} \in \mathbb{R}^d$，$g$：$\mathbb{R}^d \to \mathbb{R}^d$. 如果$g$是收缩算子，那么由压缩映射原理，定点迭代

$$\boldsymbol{x}_{k+1} = g(\boldsymbol{x}_k),\ k = 0, 1, \cdots, \tag{2}$$

收敛. 为加速迭代(2)，AM对历史迭代序列进行外推来生成新的迭代值. 具体地，设第k次迭代用到的历史序列的长度为m_k，AM使用下式更新得到

$$\boldsymbol{x}_{k+1} = (1-\beta_k)\sum_{j=0}^{m_k}\alpha_k^{(j)}\boldsymbol{x}_{k-m_k+j} + \beta_k\sum_{j=0}^{m_k}\alpha_k^{(j)}g(\boldsymbol{x}_{k-m_k+j}), \tag{3}$$

其中β_k为混合参数；$\{\alpha_k^{(j)}\}_{j=0}^{m_k}$为外推系数，由求解以下带约束的最小二乘问题得到：

$$\min_{\{\alpha_k^{(j)}\}_{j=0}^{m_k}}\left\|\sum_{j=0}^{m_k}\alpha_k^{(j)}\left(g(\boldsymbol{x}_{k-m_k+j})-\boldsymbol{x}_{k-m_k+j}\right)\right\|_2,\ \text{s.t.}\ \sum_{j=0}^{m_k}\alpha_k^{(j)} = 1. \tag{4}$$

随后将看到问题(4)实际是一个残差极小化问题，能够让外推系数的确定符合某个最优条件. 如果限制外推系数均非负，即$\alpha_k^{(j)} \geqslant 0$，$j = 0$，$\cdots$，$m_k$，就得到EDIIS方法(Kudin et al.,2002).

相较于定点迭代，AM的迭代更新过程的主要开销在于存储历史序列并对之完成外推计算，而并不需要多余的关于g的计算. 在之后的研究中，人们基于AM的迭代格式发展得到更多的算法，这些算法在一些更具体的问题求解中比原始的AM有更好的性质和表现.

由于定点问题广泛存在于科学与工程的各个领域，AM有相当广阔的适用场景，因此在各类应用中围绕AM的算法设计和理论分析是目前科学界的前沿热点，本文将对关于AM的研究进展加以介绍. 接下来，本文深入剖析AM的迭代格式，随后重点介绍关于AM的几个改进算法，算法包括正则化的AM、随机AM、短递归AM和具有极小内存开销的AM算法，这些算法能够在很大程度上拓展AM的应用范围.

这里给出文中的符号定义：符号Δ为前向差分符号，例如$\Delta\boldsymbol{x}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$；符号$\dagger$表示取Penrose-Moore逆. 对任意矩阵$\boldsymbol{A}$，range($\boldsymbol{A}$)表示由$A$的列向量张成的子空间. 矩阵范数$\|\cdot\|_{F(W)}$的定义为对任意$\boldsymbol{X} \in \mathbb{R}^{d\times d}$，有$\|\boldsymbol{X}\|_{F(W)} = \|W^{1/2}\boldsymbol{X}W^{1/2}\|_F$.

1 基础算法

本节在投影-混合框架下分析AM的迭代格式，给出第一类AM方法，并介绍已有的结论.

对于求解问题(4)，一种方式是通过Lagrange乘子法求解，另一种方式是将其转换为无约束问题. 具体地，定义$\boldsymbol{x}_k$处的残差为$\boldsymbol{r}_k = g(\boldsymbol{x}_k) - \boldsymbol{x}_k$，历史序列被存储为两个矩阵$\boldsymbol{X}_k$，$\boldsymbol{R}_k \in \mathbb{R}^{d\times m_k}(m_k \geqslant 1)$：

$$\boldsymbol{X}_k = (\Delta\boldsymbol{x}_{k-m_k}, \Delta\boldsymbol{x}_{k-m_k+1}, \cdots, \Delta\boldsymbol{x}_{k-1}),\qquad \boldsymbol{R}_k = (\Delta\boldsymbol{r}_{k-m_k}, \Delta\boldsymbol{r}_{k-m_k+1}, \cdots, \Delta\boldsymbol{r}_{k-1}). \tag{5}$$

AM的迭代格式可以被分解为**投影步**和**混合步**：

$$\bar{\boldsymbol{x}}_k = \boldsymbol{x}_k - \boldsymbol{X}_k\boldsymbol{\Gamma}_k\ (\text{投影步}),\qquad \bar{\boldsymbol{r}}_k = \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}_k,\qquad \boldsymbol{x}_{k+1} = \bar{\boldsymbol{x}}_k + \beta_k\bar{\boldsymbol{r}}_k(\text{混合步}), \tag{6}$$

其中$\bar{\boldsymbol{x}}_k$和$\bar{\boldsymbol{r}}_k$分别为中间步和中间残差，二者通过混合步导出$\boldsymbol{x}_{k+1}$. $\boldsymbol{\Gamma}_k$由求解以下的最小二乘问题确定：

$$\boldsymbol{\Gamma}_k = \mathop{\arg\min}_{\boldsymbol{\Gamma}\in\mathbb{R}^{m_k}}\|\boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}\|_2. \tag{7}$$

因为$\bar{\boldsymbol{r}}_k = \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}_k$，所以问题(7)可以被视作一个残差极小化问题. 令$\boldsymbol{\Gamma}_k = (\boldsymbol{\Gamma}_k^{(1)}, \cdots, \boldsymbol{\Gamma}_k^{(m_k)})^{\mathrm{T}} \in \mathbb{R}^{m_k}$，那么外推系数$\{\alpha_k^{(j)}\}_{j=0}^{m_k}$能由$\boldsymbol{\Gamma}_k$得到：$\boldsymbol{\alpha}_k^{(0)} = \boldsymbol{\Gamma}_k^{(1)}$，$\boldsymbol{\alpha}_k^{(j)} = \boldsymbol{\Gamma}_k^{(j+1)} - \boldsymbol{\Gamma}_k^{(j)}(j = 1, \cdots, m_k - 1)$，$\boldsymbol{\alpha}_k^{(m_k)} = 1 - \boldsymbol{\Gamma}_k^{(m_k)}$. 因为有$\boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}_k = \sum_{j=0}^{m_k}\boldsymbol{\alpha}_k^{(j)}\boldsymbol{r}_{k-m_k+j}$，所以通过计算可以验证由式(6)~(7)给出的迭代与由式(3)~(4)给出的迭代是完全相同的.

注意到求解问题(7)等价于要求$\bar{\boldsymbol{r}}_k = \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}_k \perp \text{range}(\boldsymbol{R}_k)$，因此$\boldsymbol{\Gamma}_k$的计算是一个投影过程. 如果使用投影条件$\bar{\boldsymbol{r}}_k = \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma}_k \perp \text{range}(\boldsymbol{X}_k)$，就导出**第一类AM方法**(Fang et al.,2009). 与之区分，原始的AM方法被称为**第二类AM方法**. 两类方法的迭代格式可以被写成

$$\boldsymbol{x}_{k+1} = \boldsymbol{G}_k\boldsymbol{r}_k = \boldsymbol{x}_k + \beta_k\boldsymbol{r}_k - (\boldsymbol{X}_k + \beta_k\boldsymbol{R}_k)\boldsymbol{\Gamma}_k,$$

其中 $\boldsymbol{G}_k$，$\boldsymbol{\Gamma}_k$ 由各自的投影条件确定. Fang et al.(2009)指出两类AM方法实际都为**多重割线的拟Newton法**. 具体而言，第一类AM的 $\boldsymbol{G}_k$ 满足 $\boldsymbol{G}_k = \boldsymbol{J}_k^{-1}$，其中 J_k 求解了 $\min_J \left\| \boldsymbol{J} - \beta_k^{-1}\boldsymbol{I} \right\|_F$ s.t. $\boldsymbol{J}\boldsymbol{X}_k = -\boldsymbol{R}_k$；第二类AM的 $\boldsymbol{G}_k$ 求解了 $\min_G \left\| \boldsymbol{G} - \beta_k \boldsymbol{I} \right\|_F$ s.t. $\boldsymbol{G}\boldsymbol{R}_k = -\boldsymbol{X}_k$.

在AM的使用中需要确定历史序列长度 m_k. 一种做法是取 $m_k = k$，即使用全部的历史信息，因此这被称为**全记忆的方法**；另一种做法是取 $m_k = \min\{m, k\}$，即使用最近 m 步的迭代信息，从而限制内存占用，这被称为**有限内存的方法**，将第一类AM和第二类AM分别记为AM-I(m)和AM-II(m). 由于AM的外推计算的开销在 m_k 较大时较为可观，因此一种节省开销并保留一定的AM加速效果的方式是使用定点迭代和AM交替迭代(Pratapa et al.,2016；Suryanarayana et al.,2019). 在该方案中，每 $\boldsymbol{p}$ 步迭代中先执行 $(\boldsymbol{p}-1)$ 步定点迭代，再执行1步AM迭代.

关于AM的理论分析主要分为两部分，一个是线性定点问题中全记忆的AM和Krylov子空间方法的关系，另一个是非线性定点问题中有限内存的第二类AM的收敛性. 以下加以叙述.

对于求解线性方程组，两类AM与Arnoldi方法(Saad,1981)、GMRES方法(Saad et al.,1986)这两类典型的Krylov子空间方法有着本质的联系. 设问题(1)中 $g(\boldsymbol{x}) = (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} + \boldsymbol{b}$，$\boldsymbol{A} \in \mathbb{R}^{d\times d}$ 非奇异，$\boldsymbol{b} \in \mathbb{R}^d$. 令 $\{\boldsymbol{x}_k^{\mathrm{I}}\}$ 和 $\{\boldsymbol{x}_k^{\mathrm{II}}\}$ 分别为全记忆的第一类和全记忆的第二类AM生成的序列，令 $\{\boldsymbol{x}_k^{\mathrm{A}}\}$ 和 $\{\boldsymbol{x}_k^{\mathrm{G}}\}$ 为Arnoldi方法和GMRES生成的序列. 如果各算法的迭代初值相同，那么，在一定假设下，有关系式 $\bar{x}_k^{\mathrm{I}} = \boldsymbol{x}_k^{\mathrm{A}}$ 和 $\bar{x}_k^{\mathrm{II}} = \boldsymbol{x}_k^{\mathrm{G}}$ 成立，即AM的中间步和对应的一类Krylov子空间方法的迭代步相同. Walker et al.(2011)最早对此关系给出严格证明，并在文中称之为“基本等价”，本文沿用这样的说法. 简要来说，在线性情形，可以证明 $\mathrm{range}(\boldsymbol{X}_k)$ 是Krylov子空间，AM的投影条件实质对应于两类Galerkin条件(Saad,2003)，因此可以证明基本等价性. 这种等价性解释了AM在线性方程组求解中能快速收敛的原因，AM也因此被称为非线性Krylov子空间法(Calef et al.,2013). 同时，Walker et al.(2011)也指出AM在线性方程组求解中不如GMRES可靠. 此外，前述交替迭代方法也和GMRES在一定情形下有等价性(Lupo Pasini,2019).

AM在非线性定点问题中的收敛性分析直到2015年才有实质的突破，目前的主要工作集中在对第二类AM的分析上. Toth et al.(2015)在一定的假设条件下证明了AM-II(m)的局部线性收敛性. 设问题(1)中 g 在定点的一个邻域内Lipschitz连续可微，Lipschitz常数为 $\kappa \in (0,1)$. 对 $\tilde{\kappa} \in (\kappa, 1)$，如果初值足够好，并且迭代中保证 $\sum_{j=0}^{m_k} \left|\alpha_k^{(j)}\right|$ 一致有界，那么对残差 $\boldsymbol{r}_k$，有

$$\left\| \boldsymbol{r}_k \right\|_2 \leqslant \tilde{\kappa}^k \left\| \boldsymbol{r}_0 \right\|_2 . \tag{8}$$

之后，类似的收敛性结论对于EDIIS也被证明成立(Chen et al.,2019)，并于近期被推广到针对非光滑问题的分析中(Mai et al.,2020；Bian et al.,2021；Bian et al.,2022). 然而，这些理论结果还非常局限. Anderson(2019)在其评论中指出，结论(8)不能解释AM在实践中比定点迭代明显更好的收敛性，因为后者 $\boldsymbol{q}$-线性收敛并且 $\boldsymbol{q}$-因子为 κ. 为此，Evans et al.(2020)给出了一个改进的结论：

$$\left\| \boldsymbol{r}_{k+1} \right\|_2 \leqslant s_k(1 - \beta_k + \kappa\beta_k)\left\| \boldsymbol{r}_k \right\|_2 + \sum_{j=0}^{m} \mathcal{O}\left(\left\| \boldsymbol{r}_{k-j} \right\|_2^2 \right), \tag{9}$$

其中 $s_k = \left\| \bar{\boldsymbol{r}}_k \right\|_2 / \left\| \boldsymbol{r}_k \right\|_2$，$k \geqslant m$. 因为总有 $s_k \in [0,1]$，因此当 $s_k \ll 1$ 时，残差的一阶分量迅速衰减. Pollock et al.(2021)给出了更精细的分析结果. 由于 s_k 是在迭代过程中才能确定的，结论(9)不能在迭代之初预测AM的收敛情况.

2　正则化的Anderson混合

正则化的Anderson混合(Scieur et al.,2020)是AM的一种改进算法，通常能够改善AM迭代的稳定性. 在AM的外推计算中，求解最小二乘问题可能会带来数值稳定性问题，因此Walker et al.(2011)建议使用QR分解求解问题(7)，这样在 R_k 的列接近线性相关时可以确保求解的精度. 即便如此，对于求解非线性问题，如果算得的 $\left|\alpha_k^{(j)}\right|$ 过大，也会使得迭代不稳定. 因此，正则化的AM在问题(7)中引入正则项以限制 $\left\| \boldsymbol{\Gamma}_k \right\|_2$ 的大小，$\boldsymbol{\Gamma}_k$ 的确定方式为

$$\boldsymbol{\Gamma}_k = \arg\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{m_k}} \left\| \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma} \right\|_2^2 + \delta \left\| \boldsymbol{\Gamma} \right\|_2^2, \tag{10}$$

其中$\delta \geqslant 0$为正则化参数. 从而得到$\boldsymbol{\Gamma}_k = (\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{R}_k + \delta\boldsymbol{I})^{-1}\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{r}_k$，缓解了$\boldsymbol{R}_k$非列满秩时的数值影响.

正则化的AM是人们应用AM求解实际问题的一种常用方案. Scieur et al.(2020)在交替迭代算法中引入该正则化，得到适用于无约束最优化的优化算法，并且通过正则化的Chebyshev多项式给出了收敛性分析. Fu et al.(2020)将正则化的AM用于Douglas-Rachford算子分裂迭代的加速，算法能够求解带线性等式约束的非光滑凸优化问题. Henderson et al.(2019)在正则化的AM基础上引入重启动和单调性检验，将算法应用于机器学习中的EM算法的加速. Sun et al.(2021)将正则化的AM用于强化学习的训练加速. 在这些实践中，正则化对算法的稳定性起到了重要作用.

3 随机Anderson混合

随机Anderson混合(Wei et al.,2021)是AM的一种随机版本，适用于求解随机优化问题. 问题描述为

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \coloneqq \mathbb{E}_{\xi}[F(\boldsymbol{x};\ \xi)], \tag{11}$$

其中函数F：$\mathbb{R}^d \times \Xi \to \mathbb{R}$是连续可微并且可能非凸的函数，$\xi \in \Xi$是随机变量. 因为通常情况下$F(\cdot;\ \xi)$的具体形式难以显式给出，如$\xi$服从一个未知的概率分布，或者显式计算$f$开销过大，所以实践中只能获得问题(11)的带噪声的一阶信息，即带噪声的梯度估计. 通过对ξ采样，得到问题(11)的一个特例，即**经验风险极小化问题**：

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \coloneqq \frac{1}{T}\sum_{i=1}^{T} f_i(\boldsymbol{x}), \tag{12}$$

其中f_i：$\mathbb{R}^d \to \mathbb{R}$是关于第$i$个数据样本的函数，$T$是样本总数. 问题(12)广泛存在于机器学习的各种算法之中. 通常T很大，造成遍历数据集得到全梯度$\nabla f(\boldsymbol{x})$的代价昂贵，因此实用的方式是在T个样本中随机采样，得到样本集上的梯度作为全梯度的估计.

使用AM求解优化问题是一个自然的想法，因为梯度下降法$\boldsymbol{x}_{k+1} = g(\boldsymbol{x}_k) \coloneqq \boldsymbol{x}_k - \nabla f(\boldsymbol{x}_k)$是一个定点迭代，可以尝试用AM加速，此时残差为$\boldsymbol{r}_k = g(\boldsymbol{x}_k) - \boldsymbol{x}_k = -\nabla f(\boldsymbol{x}_k)$. 然而，随机优化有本质的难度，由于不能得到精确的梯度，如果使用带噪声的梯度定义残差，传统的AM没有任何收敛性保证. 随机AM将AM推广到求解随机优化，拓展了AM的应用范围.

在随机AM算法中，定义残差$\boldsymbol{r}_k = -g(\boldsymbol{x}_k)$，其中$g(\boldsymbol{x}_k)$是无偏的梯度估计. 相应地，如式(5)所示可以得到历史序列$\boldsymbol{X}_k$，$\boldsymbol{R}_k \in \mathbb{R}^{d\times m_k}$，其中$m_k = \min\{m, k\}$. 随机AM在AM-II($m$)的基础上引入了阻尼投影和自适应正则化，其投影步和混合步为：

$$\bar{\boldsymbol{x}}_k = \boldsymbol{x}_k - \alpha_k\boldsymbol{X}_k\boldsymbol{\Gamma}_k, \text{(投影步)} \qquad \bar{\boldsymbol{r}}_k = \boldsymbol{r}_k - \alpha_k\boldsymbol{R}_k\boldsymbol{\Gamma}_k, \qquad \boldsymbol{x}_{k+1} = \bar{\boldsymbol{x}}_k + \beta_k\bar{\boldsymbol{r}}_k, \text{(混合步)} \tag{13}$$

其中$\alpha_k \in [0, 1]$为阻尼参数，系数$\boldsymbol{\Gamma}_k$由以下正则化的最小二乘问题确定：

$$\boldsymbol{\Gamma}_k = \arg\min_{\boldsymbol{\Gamma}\in\mathbb{R}^{m_k}} \left\| \boldsymbol{r}_k - \boldsymbol{R}_k\boldsymbol{\Gamma} \right\|_2^2 + \delta_k \left\| \boldsymbol{X}_k\boldsymbol{\Gamma} \right\|_2^2, \tag{14}$$

其中$\delta_k \geqslant 0$为正则化参数. 由于投影步的变化可能过大，导致中间步越过目标函数当前的信赖域，因此阻尼投影使得$\bar{\boldsymbol{x}}_k = (1-\alpha_k)\boldsymbol{x}_k + \alpha_k(\boldsymbol{x}_k - \boldsymbol{X}_k\boldsymbol{\Gamma}_k) = \boldsymbol{x}_k - \alpha_k\boldsymbol{X}_k\boldsymbol{\Gamma}_k$. 同时，正则化也起到限制$\left\| \boldsymbol{X}_k\boldsymbol{\Gamma}_k \right\|_2$的大小的作用. 这些操作可以改善算法在随机场景中的鲁棒性和稳定性，确保算法的收敛性.

随机AM在非凸随机优化中有全局收敛性. 假设f连续可微且有下界，∇f全局Lipschitz连续；各样本独立无关，并且与之前的迭代步无关，梯度估计是全梯度的无偏估计，且方差一致有界. 对于随机AM，使用递减的混合参数

$$\sum_{k=0}^{+\infty}\beta_k = +\infty, \qquad \sum_{k=0}^{+\infty}\beta_k^2 < +\infty,$$

并对α_k，δ_k施加必要的条件，有

$$\liminf_{k\to\infty} \left\| \nabla f(\boldsymbol{x}_k) \right\|_2 = 0 \quad \text{以概率1成立.}$$

如果还有梯度估计$g(\boldsymbol{x}_k)$一致有界，那么有$\lim_{k\to\infty} \left\| \nabla f(\boldsymbol{x}_k) \right\|_2 = 0$以概率1成立. 此外，如果从历史迭代步

里随机选取解$\bar{\boldsymbol{x}}$，那么为了确保$\mathbb{E}\left[\left\|\nabla f(\bar{\boldsymbol{x}})\right\|_2^2\right]\leqslant\epsilon$，所需的梯度采样总次数为$\mathcal{O}(\epsilon^{-2})$，这表明随机AM达到了一阶黑盒随机优化的最优复杂度.

此外，Wei et al.(2021)还给出随机AM的改进方案，方案包括预处理和方差减少技术.

预处理的随机AM使用了预处理的混合步：

$$\boldsymbol{x}_{k+1}=\bar{\boldsymbol{x}}_k+\boldsymbol{M}_k^{-1}\bar{\boldsymbol{r}}_k,\tag{15}$$

其中$\boldsymbol{M}_k$是对Hessian阵$\nabla^2 f(\boldsymbol{x}_k)$的近似. 将式(13)中的投影步和式(15)结合即为预处理的随机AM. 设整体的迭代式为$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k+\boldsymbol{G}_k\boldsymbol{r}_k$，那么在$\alpha_k=1$，$\delta_k=0$时，$\boldsymbol{G}_k$求解了$\min_{\boldsymbol{G}}\left\|\boldsymbol{G}-\boldsymbol{M}_k^{-1}\right\|_F$ s.t. $\boldsymbol{G}\boldsymbol{R}_k=-\boldsymbol{X}_k$.

方差减少技术起源于SVRG算法(Johnson et al.,2013)，适用于求解经验风险极小化问题. 如果在随机AM中使用方差减少的梯度估计，那么为了确保$\mathbb{E}\left[\left\|\nabla f(\boldsymbol{x})\right\|_2^2\right]\leqslant\epsilon$，所需的梯度采样总次数为$\mathcal{O}(T+(T^{2/3}/\epsilon))$，即算法复杂度得到了改进.

随机AM已经被成功用于深度学习的神经网络训练. 在图像分类和语言模型等任务上，随机AM相比现有的随机优化器有明显更好的收敛性，在大部分任务上节约了总的计算时间，因此继承了AM在确定性问题中的优良效果，在非凸随机优化中有很好的适用性.

4　短递归的Anderson混合

短递归的Anderson混合(Wei et al.,2022a)减少AM的内存开销，适用于高维问题的求解. 与各种类型的拟Newton法类似，AM需要存储历史序列$\boldsymbol{X}_k$，$\boldsymbol{R}_k\in\mathbb{R}^{d\times m_k}$，因此相比定点迭代需要额外存储$2m_k$个维数为$d$的向量. 如果历史序列长度较大，内存开销将成为AM的瓶颈，致使算法无法在存储资源受限的机器上求解高维问题. 短递归的AM将历史序列的长度降到2，同时能够保证良好的收敛性.

首先介绍短递归AM的基础形式. 与AM不同，短递归的AM使用修正的历史序列$\boldsymbol{P}_k$，$\boldsymbol{Q}_k\in\mathbb{R}^{d\times 2}$，在每步迭代之初需要对向量对$\Delta\boldsymbol{x}_{k-1}$，$\Delta\boldsymbol{r}_{k-1}$作修正. 初始化$\boldsymbol{P}_0$，$\boldsymbol{Q}_0=\boldsymbol{0}$，在第$k$步迭代，构造$\boldsymbol{p}_k$，$\boldsymbol{q}_k\in\mathbb{R}^d$：

$$\boldsymbol{p}_k=\Delta\boldsymbol{x}_{k-1}-\boldsymbol{P}_{k-1}\zeta_k,\quad \boldsymbol{q}_k=\Delta\boldsymbol{r}_{k-1}-\boldsymbol{Q}_{k-1}\zeta_k,\tag{16}$$

其中选取$\zeta_k=\arg\min_\zeta\left\|\Delta\boldsymbol{r}_{k-1}-\boldsymbol{Q}_{k-1}\zeta\right\|_2$. 从而得到$\boldsymbol{P}_k=(\boldsymbol{p}_{k-1},\ \boldsymbol{p}_k)$，$\boldsymbol{Q}_k=(\boldsymbol{q}_{k-1},\ \boldsymbol{q}_k)$. 进而迭代为

$$\bar{\boldsymbol{x}}_k=\boldsymbol{x}_k-\boldsymbol{P}_k\boldsymbol{\Gamma}_k\text{(投影步)},\qquad \boldsymbol{x}_{k+1}=\bar{\boldsymbol{x}}_k+\beta_k\bar{\boldsymbol{r}}_k\text{(混合步)},\tag{17}$$

其中$\bar{\boldsymbol{r}}_k=\boldsymbol{r}_k-\boldsymbol{Q}_k\boldsymbol{\Gamma}_k$，$\boldsymbol{\Gamma}_k=\arg\min_\Gamma\left\|\boldsymbol{r}_k-\boldsymbol{Q}_k\Gamma\right\|_2$.

当求解对称正定线性方程组(或强凸二次优化问题)时，由式(16)~(17)定义的短递归AM和全记忆的第二类AM完全等价，这意味着短递归的AM虽然仅使用长度为2的历史序列，但是却与AM-II(∞)有相同的收敛性，因此不存在历史信息的遗忘.

对于求解一般的非线性定点问题，通过引入周期性重启动和对$\left\|\boldsymbol{P}_{k-1}\zeta_k\right\|_2$，$\left\|\boldsymbol{Q}_{k-1}\zeta_k\right\|_2$的有界性检查，在对$g$的标准假设(Toth et al.,2015；Evans et al.,2020)下，短递归的AM具有局部线性收敛性：

$$\left\|\boldsymbol{r}_{k+1}\right\|_2\leqslant s_k\left(\left|1-\beta_k\right|+\kappa\beta_k\right)\left\|\boldsymbol{r}_k\right\|_2+\hat{\kappa}\sum_{j=0}^{m_k}\mathcal{O}\left(\left\|\boldsymbol{r}_{k-j}\right\|_2^2\right),\tag{18}$$

其中$s_k=\left\|\bar{\boldsymbol{r}}_k\right\|_2/\left\|\boldsymbol{r}_k\right\|_2\leqslant 1$，$\kappa$和$\hat{\kappa}$分别为$g$和$g$的导数的Lipschitz常数. 因此短递归AM在理论上没有减弱AM的收敛性. 对于求解非凸优化问题，通过引入阻尼投影和正则化，短递归的AM具有全局收敛性，并且在非凸随机优化中有收敛性保证. 因此，如果应用对迭代算法的内存占用有限制，那么相较于有限内存的AM和随机AM，短递归的AM更有优势.

5　具有极小内存开销的Anderson混合

具有极小内存开销的Anderson混合(Wei et al.,2022b)(Min-AM)是一种基于AM的高效优化算法，适用于大规模优化问题的求解. Min-AM的历史序列长度为1，因此具有极小的内存开销. 同时，Min-AM在优化问题中仍有不输于拟Newton法的收敛性.

对于求解优化问题，因为光滑函数在最优解的局部邻域能用一个二次函数近似，所以如果优化器能

快速地优化该二次函数，那么在优化原目标函数时也有望有良好的收敛效果. 因此，考虑强凸二次优化

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \coloneqq \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}, \tag{19}$$

其中 $\boldsymbol{A}\in\mathbb{R}^{d\times d}$ 对称正定. 令 $\boldsymbol{s}_k \coloneqq \Delta\boldsymbol{x}_{k-1}$，$\boldsymbol{t}_k \coloneqq \Delta\boldsymbol{r}_{k-1}$，那么 AM-II(1)的迭代公式为 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{G}_k\boldsymbol{r}_k$，其中由式(6)~(7)得到 $\boldsymbol{G}_k = \beta_k I - (\boldsymbol{s}_k + \beta_k\boldsymbol{t}_k)(\boldsymbol{t}_k^{\mathrm{T}}\boldsymbol{t}_k)^{-1}\boldsymbol{t}_k^{\mathrm{T}}$. 从拟 Newton 法的观点来看，$\boldsymbol{G}_k$ 不是对称的，因此不是一个对 Hessian 矩阵的逆矩阵的良好近似. Min-AM 使用修正的历史向量对 $\boldsymbol{p}_k$，$\boldsymbol{q}_k\in\mathbb{R}^d$，并且在 AM 的迭代格式基础上引入了附加的投影步. 具体描述如下：

初始化 $\boldsymbol{p}_1 = \boldsymbol{s}_0$，$\boldsymbol{q}_1 = \boldsymbol{t}_0$. 对 $k\geqslant 2$，假设 $\boldsymbol{p}_{k-1}^{\mathrm{T}}\boldsymbol{q}_{k-1}\neq 0$，那么在第 k 步迭代之初构造 $\boldsymbol{p}_k$，$\boldsymbol{q}_k$ 为

$$\boldsymbol{p}_k = \boldsymbol{s}_k - \boldsymbol{p}_{k-1}\zeta_k, \qquad \boldsymbol{q}_k = \boldsymbol{t}_k - \boldsymbol{q}_{k-1}\zeta_k,$$

其中 $\zeta_k = \left(\boldsymbol{p}_{k-1}^{\mathrm{T}}\boldsymbol{q}_{k-1}\right)^{-1}\boldsymbol{p}_{k-1}^{\mathrm{T}}\boldsymbol{t}_k$. 该操作使得 $\boldsymbol{q}_k\perp\boldsymbol{p}_{k-1}$. 之后是对 $\boldsymbol{x}_k$ 的迭代更新，在式(6)之后引入附加的投影步，即

$$\boldsymbol{x}_k^{(1)} = \boldsymbol{x}_k - \boldsymbol{p}_k\boldsymbol{\Gamma}_k^{(1)}\text{(投影步)}, \qquad \boldsymbol{x}_k^{(2)} = \boldsymbol{x}_k^{(1)} + \beta_k\boldsymbol{r}_k^{(1)}\text{ (混合步)}, \qquad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k^{(2)} - \boldsymbol{p}_k\boldsymbol{\Gamma}_k^{(2)}\text{ (投影步)}, \tag{20}$$

其中 $\boldsymbol{r}_k^{(1)} \coloneqq \boldsymbol{r}_k - \boldsymbol{q}_k\boldsymbol{\Gamma}_k^{(1)}$，$\beta_k > 0$. $\boldsymbol{x}_k^{(1)}$，$\boldsymbol{x}_k^{(2)}$ 都是中间步，为了确定 $\boldsymbol{\Gamma}_k^{(1)}$，$\boldsymbol{\Gamma}_k^{(2)}$，定义 $\boldsymbol{r}_k^{(2)} = \boldsymbol{r}_k^{(1)} - \beta_k\boldsymbol{A}\boldsymbol{r}_k^{(1)}$，可见 $\boldsymbol{r}_k^{(1)}$，$\boldsymbol{r}_k^{(2)}$ 分别是 $\boldsymbol{x}_k^{(1)}$，$\boldsymbol{x}_k^{(2)}$ 处的残差. $\boldsymbol{\Gamma}_k^{(1)}$，$\boldsymbol{\Gamma}_k^{(2)}$ 由第一类 AM 的投影条件确定：

$$\boldsymbol{r}_k^{(1)} = \boldsymbol{r}_k - \boldsymbol{q}_k\boldsymbol{\Gamma}_k^{(1)}\perp\boldsymbol{p}_k, \qquad \boldsymbol{r}_{k+1} = \boldsymbol{r}_k^{(2)} - \boldsymbol{q}_k\boldsymbol{\Gamma}_k^{(2)}\perp\boldsymbol{p}_k. \tag{21}$$

假设 $\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{q}_k\neq 0$，由式(20)给出的迭代为

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{H}_k\boldsymbol{r}_k, \qquad \boldsymbol{H}_k = -\frac{\boldsymbol{p}_k\boldsymbol{p}_k^{\mathrm{T}}}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{q}_k} + \beta_k\left(I - \frac{\boldsymbol{p}_k\boldsymbol{q}_k^{\mathrm{T}}}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{q}_k}\right)\left(I - \frac{\boldsymbol{q}_k\boldsymbol{p}_k^{\mathrm{T}}}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{q}_k}\right). \tag{22}$$

可以看到 $\boldsymbol{H}_k$ 是对称的. 迭代公式(22)即为 Min-AM 的基础形式.

在求解强凸二次优化问题(19)中，Wei et al.(2022b)揭示了 Min-AM 与共轭梯度法、Newton 法、BFGS(Nocedal et al.,2006)的本质联系. 以下介绍有关结论.

关于 Min-AM 的一个重要结论是 Min-AM 与第一类 AM 和共轭梯度法基本等价. 考虑求解问题(19)，设 $\{\boldsymbol{x}_k\}$ 是 Min-AM 生成的迭代序列，$\boldsymbol{x}_k^{(1)}$ 是第 k 步迭代中的第一个中间步(见迭代格式(20))；设 $\{\boldsymbol{x}_k^{\mathrm{I}}\}$ 是第一类 AM 生成的迭代序列，$\bar{\boldsymbol{x}}_k^{\mathrm{I}}$ 是第 k 步迭代中的中间步(见迭代格式(6))；$\{\boldsymbol{x}_k^{\mathrm{CG}}\}$ 是共轭梯度法生成的迭代序列. 基本等价性指如果迭代初值相同，那么

$$\boldsymbol{x}_k^{(1)} = \bar{\boldsymbol{x}}_k^{\mathrm{I}} = \boldsymbol{x}_k^{\mathrm{CG}}. \tag{23}$$

这意味着 3 个算法的收敛性基本相同.

定义 $\boldsymbol{P}_k = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k)$，$\boldsymbol{Q}_k = (\boldsymbol{q}_1, \cdots, \boldsymbol{q}_k)$，并定义 $\boldsymbol{V}_k\in\mathbb{R}^{d\times(d-k)}$ 使得 $\boldsymbol{V}_k^{\mathrm{T}}\boldsymbol{P}_k = 0$. 对优化问题(19)，在第 k 步迭代，将 $\boldsymbol{x}\in\mathbb{R}^d$ 写为 $\boldsymbol{x} = \boldsymbol{x}_k - \boldsymbol{P}_k\gamma - \boldsymbol{V}_k\eta$，其中 $\gamma\in\mathbb{R}^k$，$\eta\in\mathbb{R}^{d-k}$. 先在子空间 $\mathrm{range}(\boldsymbol{P}_k)$ 上运用 Newton 法，接着在 $\mathrm{range}(\boldsymbol{P}_k)^{\perp}$ 上以步长 β_k 梯度下降，最后再在 $\mathrm{range}(\boldsymbol{P}_k)$ 上运用 Newton 法，得到

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{H}_k^{\mathrm{B}}\boldsymbol{r}_k, \qquad \boldsymbol{H}_k^{\mathrm{B}} = -\boldsymbol{P}_k(\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{Q}_k)^{-1}\boldsymbol{P}_k^{\mathrm{T}} + \beta_k(I - \boldsymbol{P}_k(\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{Q}_k)^{-1}\boldsymbol{Q}_k^{\mathrm{T}})(I - \boldsymbol{Q}_k(\boldsymbol{P}_k^{\mathrm{T}}\boldsymbol{Q}_k)^{-1}\boldsymbol{P}_k^{\mathrm{T}}). \tag{24}$$

称迭代格式(24)为 B 迭代. 通过计算可以验证 $\boldsymbol{H}_k^{\mathrm{B}}$ 求解了 $\min_{\boldsymbol{H}}\left\|\boldsymbol{H} - \beta_k I\right\|_{F(A)}$，s.t. $\boldsymbol{H}\boldsymbol{Q}_k = -\boldsymbol{P}_k$，$\boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}}$. 进一步地，由 Min-AM 的性质可以导出关系式(Wei et al.,2022b)：

$$\boldsymbol{H}_k\boldsymbol{r}_k = \boldsymbol{H}_k^{\mathrm{B}}\boldsymbol{r}_k. \tag{25}$$

这表明在求解强凸二次优化问题时，Min-AM 和 B 迭代完全等价. 而 B 迭代由 Newton 法和梯度下降法导出，并且是一种对称化的多重割线拟 Newton 法，当 $k = d$ 时，B 迭代在全空间上使用 Newton 法. 这就建立了 Min-AM 和 Newton 法的联系，可以认为 Min-AM 隐式地构造了 Hessian 阵的近似逆矩阵 $\boldsymbol{H}_k^{B}$.

进一步地，如果 Min-AM 和 B 迭代的参数 β_k 为常数 β，那么有 $\boldsymbol{H}_0^{\mathrm{B}} = \beta\boldsymbol{I}$，随后的 $\boldsymbol{H}_k^{\mathrm{B}}$ 求解了

$$\min_{\boldsymbol{H}}\left\|\boldsymbol{H} - \boldsymbol{H}_{k-1}^{\mathrm{B}}\right\|_{F(A)}, \quad \text{s.t. } \boldsymbol{H}\boldsymbol{q}_k = -\boldsymbol{p}_k,\ \boldsymbol{H} = \boldsymbol{H}^{\mathrm{T}}. \tag{26}$$

如果将 $\boldsymbol{p}_k$，$\boldsymbol{q}_k$ 替换为 $\boldsymbol{s}_k$，$\boldsymbol{t}_k$，那么式(26)导出 BFGS 算法. 这个关系表明，B 迭代使用修正的历史序列构造 Hessian 阵的近似逆矩阵，由于 Min-AM 和 B 迭代等价，因此 Min-AM 相较于 BFGS 能够减少大量的内存占用.

对于求解一般的非线性光滑优化乃至随机优化，Min-AM 也有明确的收敛性结论. 在确定性的光滑优

化中，通过在迭代格式(22)上引入重启动和必要的检验，可以证明Min-AM的收敛率最优地依赖于问题的条件数，Min-AM和使用精确线搜索的非线性共轭梯度法有相当的收敛性. 在随机优化中，与随机AM类似，引入阻尼项和正则化，Min-AM有全局收敛性并达到了最优的迭代复杂度. 因此，Min-AM不仅将AM在优化问题中的内存开销降到极小，而且保证了算法的收敛性，在优化问题的求解中具备明显的优势. 此外，对于确定性的光滑优化，Wei et al.(2022b)指出Min-AM可以使用很小的附加计算代价估计Hessian矩阵的特征值信息，从而估计混合参数β_k的最优选取，这对于算法的实际应用也是有益的.

6　总　结

Anderson混合是加速定点迭代的一种强有力的算法，现有的研究揭示了其与Krylov子空间方法和拟Newton法的深刻联系. 目前Anderson混合的收敛性问题还没有得到完全解决，仍需要有更好的理论分析结果来更精确地刻画Anderson混合在非线性定点问题中的收敛行为. 为了改善算法的稳定性、增大算法的使用范围，一些Anderson混合的改进算法被提出并得到了成功应用. 本文介绍了其中一些有代表性的改进算法，结论表明基于Anderson混合的新算法能够被用于随机优化等更困难的问题，并且能够在内存开销上相较于传统的拟Newton法有明显的优势，有望解决科学计算和机器学习等领域中具有挑战性的实际问题，值得被进一步研究.

参考文献：

ANDERSON D G，1965. Iterative procedures for nonlinear integral equations[J]. J ACM，12(4)：547−560.

ANDERSON D G，2019. Comments on "Anderson acceleration，mixing and extrapolation"[J]. Numer Algorithms，80(1)：135−234.

ARORA A，MORSE D C，BATES F S，et al，2017. Accelerating self-consistent field theory of block polymers in a variable unit cell[J]. J Chem Phys，146(24)：244902.

BIAN W，CHEN X J，2022. Anderson acceleration for nonsmooth fixed point problems[J]. SIAM J Numer Anal，60(5)：2565−2591.

BIAN W，CHEN X J，KELLEY C T，2021. Anderson acceleration for a class of nonsmooth fixed-point problems[J]. SIAM J Sci Comput，43(5)：S1−S20.

BREZINSKI C，REDIVO-ZAGLIA M，SAAD Y，2018. Shanks sequence transformations and Anderson acceleration[J]. SIAM Rev，60(3)：646−669.

CALEF M T，FICHTL E D，WARSA J S，et al，2013. Nonlinear Krylov acceleration applied to a discrete ordinates formulation of the *k*-eigenvalue problem[J]. J Comput Phys，238：188−209.

CHEN X J，KELLEY C T，2019. Convergence of the EDIIS algorithm for nonlinear equations[J]. SIAM J Sci Comput，41(1)：A365−A379.

EVANS C，POLLOCK S，REBHOLZ L G，et al，2020. A proof that Anderson acceleration improves the convergence rate in linearly converging fixed-point methods(but not in those converging quadratically)[J]. SIAM J Numer Anal，58(1)：788−810.

FANG H R，SAAD Y，2009. Two classes of multisecant methods for nonlinear acceleration[J]. Numer Linear Algebra Appl，16(3)：197−221.

FU A Q，ZHANG J Z，BOYD S，2020. Anderson accelerated Douglas-Rachford splitting[J]. SIAM J Sci Comput，42(6)：A3560−A3583.

HENDERSON N C，VARADHAN R，2019. Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms[J]. J Comput Graph Stat，28(4)：834−846.

JOHNSON R，ZHANG T，2013. Accelerating stochastic gradient descent using predictive variance reduction[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems：315−323.

KUDIN K N，SCUSERIA G E，CANCÈS E，2002. A black-box self-consistent field convergence algorithm：One step closer[J]. J Chem Phys，116(19)：8255−8261.

LUPO PASINI M，2019. Convergence analysis of Anderson-type acceleration of Richardson's iteration[J]. Numer Linear Algebra

Appl，26(4)：e2241.

MAI V V，JOHANSSON M，2020. Anderson acceleration of proximal gradient methods[C]//Proceedings of the 37th International Conference on Machine Learning：6620−6629.

NOCEDAL J，WRIGHT S J，2006. Numerical optimization[M]. 2nd ed. New York：Springer.

POLLOCK S，REBHOLZ L G，2021. Anderson acceleration for contractive and noncontractive operators[J]. IMA J Numer Anal，41(4)：2841−2872.

POLLOCK S，REBHOLZ L G，XIAO M Y，2019. Anderson-accelerated convergence of Picard iterations for incompressible Navier-Stokes equations[J]. SIAM J Numer Anal，57(2)：615−637.

PRATAPA P P，SURYANARAYANA P，PASK J E，2016. Anderson acceleration of the Jacobi iterative method：An efficient alternative to Krylov methods for large，sparse linear systems[J]. J Comput Phys，306：43−54.

PULAY P，1980. Convergence acceleration of iterative sequences. the case of SCF iteration[J]. Chem Phys Lett，73(2)：393−398.

ROHWEDDER T，SCHNEIDER R，2011. An analysis for the DIIS acceleration method used in quantum chemistry calculations[J]. J Math Chem，49(9)：1889−1914.

SAAD Y，1981. Krylov subspace methods for solving large unsymmetric linear systems[J]. Math Comp，37(155)：105−126.

SAAD Y，2003. Iterative methods for sparse linear systems[M]. 2nd ed. Philadelphia：SIAM.

SAAD Y，SCHULTZ M H，1986. GMRES：A generalized minimal residual algorithm for solving nonsymmetric linear systems[J]. SIAM J Sci Stat Comput，7(3)：856−869.

SCIEUR D，D'ASPREMONT A，BACH F，2020. Regularized nonlinear acceleration[J]. Math Program，179(1/2)：47−83.

SUN K，WANG Y F，LIU Y，et al，2021. Damped Anderson mixing for deep reinforcement learning：Acceleration，convergence，and stabilization[C]//Proceedings of the 35th Conference on Neural Information Processing Systems：3732−3743.

SURYANARAYANA P，PRATAPA P P，PASK J E，2019. Alternating Anderson-Richardson method：An efficient alternative to preconditioned Krylov methods for large，sparse linear systems[J]. Comput Phys Commun，234：278−285.

TOTH A，KELLEY C T，2015. Convergence analysis for Anderson acceleration[J]. SIAM J Numer Anal，53(2)：805−819.

WALKER H F，NI P，2011. Anderson acceleration for fixed-point iterations[J]. SIAM J Numer Anal，49(4)：1715−1735.

WEI F C，BAO C L，LIU Y，2021. Stochastic Anderson mixing for nonconvex stochastic optimization[C]//Proceedings of the 35th Conference on Neural Information Processing Systems：22995−23008.

WEI F C，BAO C L，LIU Y，2022a. A class of short-term recurrence Anderson mixing methods and their applications[C]//The Tenth International Conference on Learning Representations：1565.

WEI F C，BAO C L，LIU Y，et al，2022b. A variant of Anderson mixing with minimal memory size[C]//Proceedings of the 36th Conference on Neural Information Processing Systems：16650−16663.

YANG Y N，2021. Anderson acceleration for seismic inversion[J]. Geophysics，86(1)：R99−R108.

（责任编辑　冯兆永）