Московский Государственный Университет имени М.В.Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Вычислительных Технологий и Моделирования

Метод ускорения Андерсона: обзор

Закс Роберт
403 учебная группа
Научный руководитель:
Матвеев С.А.

Москва
Весна 2025

Содержание

Предисловие **3**

1 Описание метода **3**
1.1 Вывод метода 4
1.2 Дальнейшие замечания 6
1.3 Способы выбора параметров 6
1.4 Вычислительные аспекты 6

2 Связь с другими методами **8**
2.1 Крыловские методы решения линейных систем 8
2.2 Квазиньютоновские методы 14
2.3 EDIIS или способ глобализации сходимости 18

3 Сходимость **19**
3.1 Понятие скорости сходимости 19
3.2 Глобальная q-линейная сходимость невязок в методе DIIS для линейных задач с сжимающим отображением 20
3.3 Глобальная r-линейная сходимость метода EDIIS(m) для произвольных задач с сжимающим отображением 21
3.4 Локальная r-линейная сходимость методов DIIS и EDIIS на нелинейных задачах 22

4 Заключение **24**

Список литературы **24**

Предисловие

В данной работе перечисляются и воспроизводятся известные результаты, посвященные методу ускорения Андерсона.
В первой части приводится описание метода, описываются способы выбора его параметров и аспекты численной реализации.
Во второй части приведено подробное изложение связи метода с крыловскими методами решения линейных систем (часть доказательств была переработана) и ознакомительное рассмотрение метода ка квазиньютоновское.
Данная работа не содержит численных экспериментов и детального описания теории сходимости метода: мы ограничимся рассмотрением линейного случая и глобальной сходимости модификации, а в теории локальной сходимости лишь осветим известные результаты (на момент написания текста автор не смог упростить и поправить доказательства, поэтому мы отсылаем читателя к прочтению оригинальных работ из списка литературы).

1 Описание метода

Метод впервые был предложен Дональдом Андерсоном [1] в 1965г. Позднее метод был несколько раз переоткрыт, из-за чего он известен под следующими названиями: метод ускорения Андерсона (Anderson acceleration), метод Андерсона с весами (Anderson mixing), Pulay mixing (Pulay 1980), nonlinear GMRES (Washio 1997), DIIS (Direct Inversion on the Iterative Subspace, Rohwedder/Scheneider).

Пусть $U, \|\cdot\|$ — банахово пространство над полем $\mathbb{P}$. Метод Андерсона решает задачу на нахождение неподвижной точки $u^* \in U$ отображения $g : U \to U$:

$$g(u^*) = u^*, \tag{1.0.1}$$

которое, после введения невязки $f(u) = g(u) - u$, принимает эквивалентный вид:

$$f(u^*) = 0. \tag{1.0.2}$$

Простейший метод решения уравнения (1.0.1) носит название *метода простой итерации* (fixed-point iteration) и может быть записан в виде:

Algorithm fixed-point iteration

input $u_0 \in U$
for $k = 0, 1 \ldots,$ **do**
  $u_{k+1} = g(u_k)$
end for

Как известно, метод простой итерации может расходится и в общем случае имеет линейную скорость сходимости. Метод ускорения Андерсона был создан для увеличения скорости и области сходимости метода простой итерации, за счет чего и получил свое название.

1.1 Вывод метода

Предположим далее, что задача (1.0.1) имеет единственное решение $u^* \in U$. Обозначим через $u_0 \in U$ начальное приближение к u^*. На k-м шаге метода используем $m_k + 1 \leqslant k + 1$ ранее найденных приближений $\{u_{k-m_k+j}\}_{j=0}^{m_k}$ и значений $\{g(u_{k-m_k+j})\}_{j=0}^{m_k}$. Поскольку во многих приложениях вычисление g обходится дорого, постараемся не вычислять на k-м шаге других значений g.
Запишем задачу (1.0.2) в эквивалентном виде

$$\min_{u \in U} \|f(u)\|. \tag{1.1.1}$$

Сузим задачу минимизации на афинную оболочку $\operatorname{aff}\{u_{k-m_k}, \ldots, u_k\}$:

$$\min_{u \in \operatorname{aff}\{u_{k-m_k}, \ldots, u_k\}} \|f(u)\|. \tag{1.1.2}$$

Т.к. $f(u) = g(u) - u$, а вычислять g нам больше нельзя, то воспользуемся приближением первого порядка: на время вывода метода предположим, что $f \in C^1$, тогда $f(u) \approx f(u^*) + f'(u^*)(u - u^*) = b - Mu$ и, поскольку из $u \in \operatorname{aff}\{u_{k-m_k}, \ldots, u_k\}$ следует $u = \sum\limits_{j=0}^{m_k} \alpha_j u_{k-m_k+j}$, $\sum\limits_{j=0}^{m_k} \alpha_j = 1$, то

$$\begin{aligned} f\left(\sum_{j=0}^{m_k} \alpha_j u_{k-m_k+j}\right) &\approx b - M \sum_{j=0}^{m_k} \alpha_j u_{k-m_k+j} = \sum_{j=0}^{m_k} \alpha_j b - \sum_{j=0}^{m_k} \alpha_j M u_{k-m_k+j} = \\ &= \sum_{j=0}^{m_k} \alpha_j (b - M u_{k-m_k+j}) \approx \sum_{j=0}^{m_k} \alpha_j f(u_{k-m_k+j}), \end{aligned} \tag{1.1.3}$$

откуда получим следующую задачу минимизации:

$$\min_{\sum\limits_{j=0}^{m_k} \alpha_j = 1} \left\| \sum_{j=0}^{m_k} \alpha_j f(u_{k-m_k+j}) \right\|. \tag{1.1.4}$$

Пусть решение этой задачи минимизации достигается на $\{\alpha_j^k\}$.
Обозначим $\overline{u}_{k+1} = \sum\limits_{j=0}^{m_k} \alpha_j^k u_{k-m_k+j}$. Тогда $\overline{u}_{k+1}$ дает приближенное решение суженной задачи минимизации, а потому является приближением исходной задачи минимизации.

Предположим теперь, что отображение g является сжимающим с константой $c \in [0; 1)$ на всем U (в этом случае по Теореме Банаха о сжимающем отображении метод простой итерации будет сходится к u^*). Тогда

$$\|g(\overline{u}_{k+1}) - u^*\| = \|g(\overline{u}_{k+1}) - g(u^*)\| \leqslant c\|\overline{u}_{k+1} - u^*\| < \|\overline{u}_{k+1} - u^*\|,$$

т.е. $g(\overline{u}_{k+1})$ дает приближение лучше, чем $\overline{u}_{k+1}$. Т.к. мы не можем на этом шаге больше вычислять g, то воспользуемся соотношением (1.1.3) и $\sum\limits_{j=0}^{m_k} \alpha_j = 1$:

$$g(\overline{u}_{k+1}) = f(\overline{u}_{k+1}) + \sum_{j=0}^{m_k} \alpha_j \overline{u}_{k+1} \approx$$

$$\approx \sum_{j=0}^{m_k} \alpha_j f(u_{k-m_k+j}) + \sum_{j=0}^{m_k} \alpha_j \overline{u}_{k+1} = \sum_{j=0}^{m_k} \alpha_j g(u_{k-m_k+j}).$$

Обозначим $\overline{g}_{k+1} = \sum\limits_{j=0}^{m_k} \alpha_j^k g(u_{k-m_k+j})$, $\overline{f}_{k+1} = \sum\limits_{j=0}^{m_k} \alpha_j^k f(u_{k-m_k+j})$.
В итоге $\overline{u}_{k+1}$ дает приближение к u^*, а в случае сжимающего g, $\overline{g}_{k+1}$ дает приближение лучше. Поскольку в общем случае отображение g не является сжимающим (по крайней мере на всем U), то на практике комбинируют полученные приближения, а именно: введем весовые параметры $\beta_k \in \mathbb{P}$ (mixing parameters; отсюда же иное название метода) и наконец положим:

$$u_{k+1} = (1 - \beta_k)\overline{u}_{k+1} + \beta_k \overline{g}_{k+1}, \tag{1.1.5}$$

а т.к. комбинация афинная, то

$$\|u_{k+1} - u^*\| = \|(1 - \beta_k)\overline{u}_{k+1} + \beta_k \overline{g}_{k+1} - (1 - \beta_k)u^* - \beta_k u^*\| =$$
$$= \|(1 - \beta_k)(\overline{u}_{k+1} - u^*) + \beta_k(\overline{g}_{k+1} - u^*)\| \leqslant |1 - \beta_k|\|\overline{u}_{k+1} - u^*\| + |\beta_k|\|\overline{g}_{k+1} - u^*\|$$

и за счет выбора β_k получается уменьшить ошибку.
Отметим, что (1.1.5) может быть записан в виде:

$$u_{k+1} = \overline{u}_{k+1} + \beta_k \overline{f}_{k+1}. \tag{1.1.6}$$

Таким образом мы готовы дать описание метода Андерсона:

Algorithm Anderson acceleration (mixing)

---

input $g : U \to U$, $u_0 \in U$, $\{m_k : m_k \leqslant k\}$, $\{\beta_k\} \subset \mathbb{P}$
for $k = 0, 1 \ldots,$ **do**
$f_k = g(u_k) - u_k$
$(\alpha_0^k, \ldots, \alpha_{m_k}^k) = \underset{\sum\limits_{j=0}^{m_k} \alpha_j = 1}{\operatorname{argmin}} \left\| \sum\limits_{j=0}^{m_k} \alpha_j f_{k-m_k+j} \right\|$
$u_{k+1} = \sum\limits_{j=0}^{m_k} \alpha_j^k u_{k-m_k+j} + \beta_k \sum\limits_{j=0}^{m_k} \alpha_j^k f_{k-m_k+j}$
end for

1.2 Дальнейшие замечания

Пусть $m \in \mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$. Положим $m_k = \min\{m, k\}$. Метод андерсона с такими параметрами в дальнейшем будем обозначать через Anderson(m). Отметим, что Anderson(0), также известный как simple mixing, при $\beta_k = 1$ дает метод простой итерации.

Для анализа поведения метода полезно заметить, что метод Андерсона при $\beta_k \equiv \beta$ эквивалентен применению метода Андерсона при $\beta_k \equiv 1$ к отображению $g_\beta(u) = (1-\beta)u + \beta g(u) = u + \beta f(u)$.

В работе [2] вводится семейство методов Андерсона. Мы не будем подробно рассматривать эти методы, ограничившись несколькими комментариями в разделе 2.2.

1.3 Способы выбора параметров

Во многих приложениях U — пространство $\mathbb{R}^n$ $(\mathbb{C}^n)$ с нормой $\|\cdot\|_p$, $p \in \{1, 2, \infty\}$, поскольку при $p \in \{1, \infty\}$ задача минимизации (1.1.4) сводится к задаче линейного программирования, а при $p = 2$ сводится к задаче наименьших квадратов:

$$\begin{gathered}(\alpha_0^k, \ldots, \alpha_{m_k-1}^k) = \operatorname{argmin} \left\| f(u_k) + \sum_{j=0}^{m_k-1} \alpha_j \left(f(u_{k-m_k+j}) - f(u_k)\right) \right\|_2, \\ \alpha_{m_k}^k = 1 - \sum_{j=0}^{m_k-1} \alpha_j^k.\end{gathered} \tag{1.3.1}$$

Вопрос выбора весов β_k неоднозначен. В работе [3] предложен и исследован в линейном случае выбор весов, при котором невязка $\|u_{k+1}\|$ была мимимальна среди всех возможных выборов β_k. На практике достаточно часто используют $\beta_k \equiv \beta$, который подбирают эмпирическим путем.

Выбор m_k также неоднозачен: при малых m_k хранимых итераций может не хватить для ускорения сходимости, а при больших m_k решение задачи наименьших квадратов усложняется и она становится плохо обусловлена. По этой причине в работе [4] предложено динамически изменять m_k до достижения приемлимой обусловленности задачи наименьших квадратов.

В качестве критерия останова обычно используют останов по числу итераций и останов по малости нормы невязки $\|f(u_k)\|$.

1.4 Вычислительные аспекты

В этом разделе обсудим организацию вычислений для случая $\mathbb{R}^n(\mathbb{C}^n), \|\cdot\|_2$.

В работе [2] была предложена иная форма записи задачи наименьших квадратов (1.3.1): обозначим $\Delta f_i = f_{i+1} - f_i, \Delta u_i = u_{i+1} - u_i, i \in \overline{k - m_k, k-1}$,

$\mathcal{F}_k = (\Delta f_{k-m_k}, \ldots, \Delta f_{k-1})$, $\mathcal{U}_k = (\Delta u_{k-m_k}, \ldots, \Delta u_{k-1})$ и рассмотрим задачу минимизации

$$\gamma^k = (\gamma_0^k, \ldots, \gamma_{m_k-1}^k) = \underset{\gamma=(\gamma_0,\ldots,\gamma_{m_k-1})}{\operatorname{argmin}} \left\| f(u_k) - \sum_{j=0}^{m_k-1} \gamma_j \left(f(u_{k-m_k+j+1}) - f(u_{k-m_k+j}) \right) \right\|_2 = \\ = \underset{\gamma=(\gamma_0,\ldots,\gamma_{m_k-1})}{\operatorname{argmin}} \|f_k - \mathcal{F}_k \gamma\|_2, \tag{1.4.1}$$

где

$$\alpha_j^k = \begin{cases} \gamma_0^k, & j = 0 \\ \gamma_j^k - \gamma_{j-1}^k, & j \in \overline{1, m_k - 1} \\ 1 - \gamma_{m_k-1}, & j = m_k. \end{cases}$$

Тогда

$$\overline{u}_{k+1} = \sum_{j=0}^{m_k} \alpha_j^k u_{k-m_k+j} = u_k - \sum_{j=0}^{m_k-1} \gamma_j \left(u_{k-m_k+j+1} - u_{k-m_k+j} \right) = u_k - \mathcal{U}_k \gamma,$$

$$\overline{f}_{k+1} = \sum_{j=0}^{m_k} \alpha_j^k f(u_{k-m_k+j}) = f(u_k) - \sum_{j=0}^{m_k-1} \gamma_j \left(f(u_{k-m_k+j+1}) - f(u_{k-m_k+j}) \right) = f_k - \mathcal{F}_k \gamma.$$

$$u_{k+1} = \overline{u}_{k+1} + \beta_k \overline{f}_{k+1} = u_k + \beta_k f_k - (\mathcal{U}_k + \beta_k \mathcal{F}_k)\gamma_k \tag{1.4.2}$$

Эта форма записи обусловлена гораздо лучше, чем (1.3.1), а также удобна для хранения и обновления информации с предыдущей итерации, о чем пойдет речь в этом разделе.

Пусть $\mathcal{F}_k = \hat{Q}_k \hat{R}_k = \begin{bmatrix} Q_k & \tilde{Q}_k \end{bmatrix} \times \begin{bmatrix} R_k \\ O \end{bmatrix} = Q_k R_k$ — QR-разложение матрицы $\mathcal{F}_k$, где $\hat{Q}_k \in \mathbb{R}^{n \times n}, \hat{R}_k \in \mathbb{R}^{n \times m_k}, Q_k \in \mathbb{R}^{n \times m_k}, R_k \in \mathbb{R}^{m_k \times m_k}$. Тогда из (1.4.1) имеем:

$$\gamma^k = \underset{\gamma}{\operatorname{argmin}} \|f_k - \mathcal{F}_k \gamma\|_2 = \underset{\gamma}{\operatorname{argmin}} \left\| f_k - \hat{Q}_k \hat{R}_k \gamma \right\|_2 = \\ = \underset{\gamma}{\operatorname{argmin}} \left\| \hat{Q}_k^* f_k - \hat{R}_k \gamma \right\|_2 = \underset{\gamma}{\operatorname{argmin}} \left\| \begin{bmatrix} Q_k^* \\ \tilde{Q}_k^* \end{bmatrix} f_k - \begin{bmatrix} R_k \\ O \end{bmatrix} \gamma \right\|_2 = \\ = \underset{\gamma}{\operatorname{argmin}} \left\| \begin{bmatrix} Q_k^* f_k - R_k \gamma \\ \tilde{Q}_k^* f_k \end{bmatrix} \right\|_2 = \underset{\gamma}{\operatorname{argmin}} \|Q_k^* f_k - R_k \gamma\|_2,$$

т.е. задача (1.4.1) сводится к решению системы $R_k \gamma = Q_k^* f_k$ и обусловленность задачи наименьших квадратов есть $\operatorname{cond}_2 R_k$.

Т.о. для решения задачи наименьших квадратов достаточно уметь находить Q_k и R_k, а т.к. матрица $\mathcal{F}_k$ получается из $\mathcal{F}_{k-1}$ добавлением одного столбца справа и удалением левых столбцов, то за счет $O(m_k n)$ арифметических операций Q_k и R_k

могут быть получены из Q_{k-1} и R_{k-1}.
В случае динамического изменения m_k, мы удаляем крайние левые столбцы $\mathcal{F}_k$ до достижения малого $\operatorname{cond}_2 R_k$, что тоже допускает пересчет R_k за $O(m_k n)$ арифметических операций.

Полное описание данного способа организаций вычислений можно найти в работе [6], где, среди прочего, приведены псевдокод и MATLAB-реализация.

2 Связь с другими методами

2.1 Крыловские методы решения линейных систем

В данном разделе мы рассмотрим поведение метода ускорения Андерсона на линейных задачах, впервые рассмотренное в работах [3] и [4]. Мы покажем связь между крыловскими методами решения линейных систем и методом ускорения Андерсона в точной арифметике, т.е. будем анализировать приближения, получаемые методами на каждой итерации, не вдаваясь в подробности различных вариантов численного нахождения этих приближений.

Пусть $g(u) = Au + b, A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$, тогда, если обозначить $M = I - A$, получим $f(u) = b-(I-A)u = b-Mu = (b-Mu_0)-M(u-u_0) = f(u_0)-M(u-u_0)$ и приближенные равенства в (1.1.3) заменяются на точные, откуда следует

$$\overline{f}_{k+1} = f(\overline{u}_{k+1}) \tag{2.1.1}$$

и т.к.

$$\sum_{j=0}^{k} \alpha_j f(u_{k-m_k+j}) = \sum_{j=0}^{m_k} \alpha_j (f(u_0) - M(u_{k-m_k+j} - u_0)) =$$
$$= f(u_0) - M \sum_{j=0}^{m_k} \alpha_j (u_{k-m_k+j} - u_0),$$

то, если обозначить $\mathcal{L}_k = \operatorname{aff}\{u_{k-m_k} - u_0, \ldots, u_k - u_0\}$, то метод Андерсона на линейных задачах может быть записан в виде (в дальнейшем мы считаем, что $\det M \neq 0$ для единственности решения задачи минимизации):

Заметим теперь, что задача $f(u) = 0$ может быть записана в виде линейной системы $Mu = b$. Тогда, в случае невырожденной матрицы M, решение системы может быть найдено крыловскими методами GMRES (Generalized minimal residual) [7] или GCR (Generalized Conjugate Residual) ([7]) следующим образом: пусть $r_0 = b - Mu_0$ — начальная невязка и $\mathcal{K}_k = \mathcal{K}_k(M, r_0) = \operatorname{span}\{r_0, Mr_0, \ldots, M^{k-1}r_0\}$ — крыловское подпространство, тогда решение на k-й итерации ищется в виде

$$u_k^G = u_0 + z_k^G,\ z_k^G = \operatorname*{argmin}_{z\in\mathcal{K}_k} \|b - M(u_0 + z)\|_2 = \operatorname*{argmin}_{z\in\mathcal{K}_k} \|r_0 - Mz\|_2, \tag{2.1.2}$$

Algorithm Anderson acceleration (mixing) on linear problems

input $g(u) = Au + b : \mathbb{R}^n \to \mathbb{R}^n$, $u_0 \in \mathbb{R}^n$, $\{m_k : m_k \leqslant k\}$, $\{\beta_k\} \subset \mathbb{R}$
$M = I - A$, $f(u) = b - Mu$
$f_0 = f(u_0)$
for $k = 0, 1 \ldots,$ **do**
$\mathcal{L}_k = \mathrm{aff}\{u_{k-m_k} - u_0, \ldots, u_k - u_0\}$
$\overline{u}_{k+1} = u_0 + \overline{z}_k$, $\overline{z}_k = \underset{z \in \mathcal{L}_k}{\operatorname{argmin}} \|f_0 - Mz\|_2$
$u_{k+1} = \overline{u}_{k+1} + \beta_k(b - M\overline{u}_{k+1}) = u_0 + \overline{z}_k + \beta_k(f_0 - M\overline{z}_k)$
end for

т.е. методы минимизируют 2-норму невязки на подпространстве Крылова.
Таким образом в точной арифметике методы могут быть записаны в эквивалентном виде:

Algorithm equivalent GMRES (GCR)

input $M \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$, $u_0 \in \mathbb{R}^n$
$r_0 = b - Mu_0$
for $k = 0, 1 \ldots,$ **do**
$\mathcal{K}_k = \mathrm{span}\{r_0, Mr_0, \ldots, M^{k-1}r_0\}$ $(\mathcal{K}_0 = \{0\})$
$u_k^G = u_0 + z_k^G$, $z_k^G = \underset{z \in \mathcal{K}_k}{\operatorname{argmin}} \|r_0 - Mz\|_2$
end for

Заметим, что поскольку $r_0 = f_0$, то, если окажется, что $\mathcal{L}_k = \mathcal{K}_k$, то будет выполнено $\overline{z}_k = z_k^G$, откуда $\overline{u}_{k+1} = u_k^G$, что даст нам требуемую связь методов.

Для получения этого факта нам понадобится вспомнить следующую теорему, описывающую некоторые свойства крыловский подпространств:

Теорема. *Следующие определения числа $\nu = \nu(M, r_0)$ эквивалентны:*

1. *ν — минимальное число, при котором $\mathcal{K}_k = \mathcal{K}_\nu, \forall k \geqslant \nu$.*
2. *ν — минимальное число, при котором $M\mathcal{K}_\nu \subseteq \mathcal{K}_\nu$.*
3. *ν — степень минимального для r_0 полинома от M, т.е. $\nu = \deg p$, где p имеет минимальную степень, среди многочленов со старшим коэффициентом 1 и удовлетворяющих $p(M)r_0 = 0$.*
4. *ν — максимальное число, при котором $\dim \mathcal{K}_\nu = \nu$.*
5. *ν — минимальное число, при котором $M^{-1}r_0 \in \mathcal{K}_\nu$.*

Кроме того, $\nu \leqslant n$.

Доказательство:
Т.к. $\mathcal{K}_{k+1} = \operatorname{span}\{\mathcal{K}_k, M^k r_0\} = \operatorname{span}\{r_0, Mr_0, \ldots, M^k r_0\} = \operatorname{span}\{r_0, M\mathcal{K}_k\}$, то если $\mathcal{K}_k = \mathcal{K}_{k+1}$, то $M\mathcal{K}_k \subseteq \operatorname{span}\{r_0, M\mathcal{K}_k\} = \mathcal{K}_{k+1} = \mathcal{K}_k$,
а если $M\mathcal{K}_k \subseteq \mathcal{K}_k$, то $\mathcal{K}_k \subseteq \mathcal{K}_{k+1} = \operatorname{span}\{r_0, M\mathcal{K}_k\} \subseteq \operatorname{span}\{r_0, \mathcal{K}_k\} = \mathcal{K}_k$, откуда $\mathcal{K}_{k+1} = \mathcal{K}_k$, т.о. $\mathcal{K}_{k+1} = \mathcal{K}_k \Leftrightarrow M\mathcal{K}_k \subseteq \mathcal{K}_k$ и $1 \Leftrightarrow 2$.
Если $\mathcal{K}_{k+1} = \mathcal{K}_k$, то $M^k r_0 \in \operatorname{span}\{\mathcal{K}_k, M^k r_0\} = \mathcal{K}_{k+1} = \mathcal{K}_k$, т.е. найдется набор ξ_j: $0 = M^k r_0 - \sum\limits_{j=0}^{k-1} \xi_j M^j r_0 = p(M) r_0$, где $\deg p = k$ и p имеет старший коэффициент 1. Если же $p(t) = t^k - \sum\limits_{j=0}^{k-1} \xi_j t^j$, то условие $p(M) r_0 = 0$ равносильно $M^k r_0 \in \mathcal{K}_k$, откуда $\mathcal{K}_{k+1} = \operatorname{span}\{\mathcal{K}_k, M^k r_0\} = \mathcal{K}_k$. Из сказанного следует, что $\mathcal{K}_{k+1} = \mathcal{K}_k \Leftrightarrow \exists p : p(A) r_0 = 0, p(t) = t^k + \ldots$, а также $1 \Leftrightarrow 3$.
Т.к. $\mathcal{K}_k \subset \mathbb{C}^n$, то $\dim \mathcal{K}_k \leqslant n$ и т.к. $\mathcal{K}_k \subseteq \mathcal{K}_{k+1}$, то ν из пунтка 1 найдется, причем $\nu \leqslant n$ и $\mathcal{K}_1 \subset \ldots \subset \mathcal{K}_\nu = \mathcal{K}_{\nu+1} = \ldots$, что доказывает 4.
Пусть ν определено пунктом 3, тогда заметим, что в обозначениях выше $\xi_0 \neq 0$, т.к. в противном случае многочлен $q(t) : p(t) = q(t)t$ будет иметь степень меньше, старший коэффициент 1 и $q(A) r_0 = 0$, но это противоречит минимальности p, а т.к. $\xi_0 \neq 0$, то из $0 = M^{-1} p(M) r_0$ получим $M^{-1} r_0 = \dfrac{1}{\xi_0} \left(M^{\nu-1} r_0 - \sum\limits_{j=1}^{\nu-1} \xi_j M^{j-1} r_0 \right) \in \mathcal{K}_\nu$. Т.к. из $M^{-1} r_0 \in \mathcal{K}_k$ следует $r_0 \in M\mathcal{K}_k$ и $\mathcal{K}_{k+1} = \operatorname{span}\{r_0, M\mathcal{K}_k\}$, то $\dim \mathcal{K}_{k+1} = k$, откуда в силу 4 получим, что $k \geqslant \nu$, что, с учетом рассуждений, выше завершает доказательство.

Важность этого числа заключена в следующим факте: т.к. в силу теоремы и (2.1.2) имеем $z_\nu^G = M^{-1} r_0$, $u_\nu^G = u_0 + M^{-1} r_0 = u_0 + M^{-1}(b - Mr_0) = M^{-1} b = u^*$ и GMRES (GCR) найдет неподвижную точку отображения g ровно на $\nu(A, r_0) \leqslant n$ итерации. Как видим, методы GMRES и GCR всегда сходится, за нее более чем n итераций, однако они могут какое-то число итераций стагнировать и нам потребуется обозначение $\eta(M, r_0) = \min\{k : u_k^G = u_{k+1}^G\} = \min\{k : z_k^G = z_{k+1}^G\}$ для номера первой стагнации.

Для дальнейших рассуждений нам потребуется исследовать подпространства $\mathcal{L}_k$ для Anderson(∞), т.е. метода ускорения Андерсона с $m_k = k$ и произвольными весами.

Теорема. *Пусть $M = I - A$ невырождена. Тогда для подпространств $\mathcal{L}_k$ и приближений u_k, получаемых методом Андерсона с $m_k = k$ и любом выборе весов β_k выполнены соотношения:*

1. $\mathcal{L}_0 = \{0\}$, $\mathcal{L}_k = \operatorname{span}\{u_1 - u_0, \ldots, u_k - u_0\}$.
2. $\mathcal{L}_k \subseteq \mathcal{L}_{k+1}, k = 0, 1, \ldots$

3. $\beta_k = 0 \Rightarrow \mathcal{L}_k = \mathcal{L}_{k+1}$.

4. $\mathcal{L}_k = \mathcal{L}_{k+1} \Rightarrow \begin{bmatrix} \beta_k = 0 \\ \mathcal{L}_{k+1} = \mathcal{L}_{k+2}, M\overline{z}_k \in \mathcal{L}_k \end{bmatrix}$.

Доказательство:
Т.к. $m_k = k$, то при $k > 0$ $\mathcal{L}_k = \mathrm{aff}\{u_{k-m_k} - u_0, \ldots, u_k - u_0\} =$
$= \mathrm{aff}\{u_0 - u_0, \ldots, u_k - u_0\} = \mathrm{span}\{u_1 - u_0, \ldots, u_k - u_0\}$, что доказывает первые два утверждения.
Если $\beta_k = 0$, то $u_{k+1} - u_0 = \overline{z}_k + \beta_k(f_0 - M\overline{z}_k) = \overline{z}_k \in \mathcal{L}_k$, откуда получим $\mathcal{L}_{k+1} = \mathrm{span}\{\mathcal{L}_k, u_{k+1} - u_0\} = \mathcal{L}_k$.

Пусть $\mathcal{L}_k = \mathcal{L}_{k+1}$. Т.к. $\mathcal{L}_{k+1} = \mathrm{span}\{\mathcal{L}_k, u_{k+1} - u_0\}$, то $\overline{z}_k + \beta_k(f_0 - M\overline{z}_k) =$
$= u_{k+1} - u_0 \in \mathcal{L}_k$ и т.к. $\overline{z}_k \in \mathcal{L}_k$, то либо $\beta_k = 0$, либо $M\overline{z}_k \in \mathcal{L}_k$. Т.к. M — невырожденная, то решение задачи минимизации единственно и в нашем случае $\overline{z}_{k+1} = \overline{z}_k$, но тогда $u_{k+2} - u_0 = \overline{z}_{k+1} + \beta_{k+1}(f_0 - M\overline{z}_{k+1}) =$
$= \overline{z}_k + \beta_{k+1}(f_0 - M\overline{z}_k) \in \mathcal{L}_k$ и $\mathcal{L}_{k+2} = \mathrm{span}\{\mathcal{L}_{k+1}, u_{k+2} - u_0\} = \mathcal{L}_{k+1}$.

Теорема показывает, что использование нулевых весов с $m_k = k$ в линейном случае способно замедлить сходимость метода: если $\beta_k = 0$, то $\mathcal{L}_k = \mathcal{L}_{k+1}$, а т.к. M невырождена, то $\overline{z}_{k+1} = \overline{z}_k$ и если мы рассмотрим метод Андерсона, в котором этот вес удален, то такой метод будет давать те же приближения, но, начиная с k-й итерации, на 1 итерацию раньше.

Обсудим теперь связь Anderson(∞) с методоми GMRES и GCR. Мы будем рассматривать метод Андерсона с ненулевыми весами, поскольку, как было сказано выше, в общем случае мы можем удалением нулевых весов свести метод к рассматриваемому случаю. Если обозначить $\mu(A, u_0) = \max\{k : \dim \mathcal{L}_k = k\}$, то будет верна следующая теорема:

Теорема. *Пусть $M = I - A$ невырождена (т.е. u^* существует и единственно при всех b). Тогда для чисел $\mu = \mu(A, u_0)$, $\nu = \nu(M, r_0)$, $\eta = \eta(M, r_0)$, а также приближений $u_k, \overline{u}_k$, получаемым методом Андерсона с $m_k = k$ и любом выборе ненулевых весов β_k, и приближений u_k^G, получаемых крыловскими методами GMRES или GCR, выполнены следующие утверждения:*

1. $\mu \leqslant \nu$.

2. $\mu = \nu$ *выполнено тогда и только тогда, когда*

$$\|f(\overline{u}_1)\|_2 > \|f(\overline{u}_2)\|_2 > \ldots > \|f(\overline{u}_\mu)\|_2 > \|f(\overline{u}_{\mu+1})\|_2 = 0$$

более того, в этом случае

$$\overline{u}_{k+1} = u_k^G, k = 0, 1, \ldots$$

и метод Андерсона сходится к точному решению u^* либо за μ, либо за $\mu + 1$ итераций.

3. $\mu < \nu$ *выполнено тогда и только тогда, когда*

$$\|f(\overline{u}_1)\|_2 > \|f(\overline{u}_2)\|_2 > \ldots > \|f(\overline{u}_\mu)\|_2 = \|f(\overline{u}_{\mu+1})\|_2 > 0.$$

более того, в этом случае

$$\overline{u}_{k+1} = \begin{cases} u_k^G, & k \leqslant \mu \\ u_{\mu-1}^G, & k \geqslant \mu \end{cases}.$$

4. $\mu = \begin{cases} \eta + 1, & \eta + 1 < \mu \\ \eta, & \eta = \mu \end{cases}$.

Доказательство:

Заметим, что $\mathcal{L}_0 = \{0\} = \mathcal{K}_0$, $\overline{u}_1 = u_0 = u_0^G$. Покажем по индукции, что $\mathcal{L}_k \subseteq \mathcal{K}_k$: база индукции при $k = 0$ записана выше, а т.к. при $\mathcal{L}_k \subseteq \mathcal{K}_k$ имеем $u_{k+1} - u_0 = \overline{z}_k + \beta_k(f_0 - M\overline{z}_k) \in \mathcal{L}_k + M\mathcal{L}_k \subseteq \mathcal{K}_k + M\mathcal{K}_k = \mathcal{K}_{k+1}$ и $\mathcal{L}_{k+1} = \operatorname{span}\{\mathcal{L}_k, u_{k+1} - u_0\} \subseteq \operatorname{span}\{\mathcal{K}_k, \mathcal{K}_{k+1}\} = \mathcal{K}_{k+1}$, то переход доказан.

Т.к. в силу первой теоремы $\nu(M, u_0) = \max\{k : \dim \mathcal{K}_k = k\}$, и по определению $\mu(A, u_0) = \max\{k : \dim \mathcal{L}_k = k\}$, то, в силу доказанного соотношения $\mathcal{L}_k \subseteq \mathcal{K}_k$ и выражения $\mathcal{L}_k$, $\mathcal{K}_k$ как линейной комбинации k векторов, получим $\mu \leqslant \nu$ и при всех $k = 0, \ldots, \mu$ выполнено $\dim \mathcal{L}_k = k = \dim \mathcal{K}_k$, а, значит, и $\mathcal{L}_k = \mathcal{K}_k$. Как было сказано ранее, последнее влечет $\overline{z}_k = z_k^G$, из которого получается $\overline{u}_{k+1} = u_k^G$ при всех $k = 0, \ldots, \mu$.

Т.к. $\mu = \max\{k : \dim \mathcal{L}_k = k\}$, то $\mathcal{L}_k \subset \mathcal{L}_{k+1}, 0 \leqslant k < \mu$, $\mathcal{L}_\mu = \mathcal{L}_{\mu+1}$, откуда по предыдущей теореме при $k \geqslant \mu$ выполнено $\mathcal{L}_k = \mathcal{L}_\mu$, откуда при $k \geqslant \mu$ получим $\overline{u}_{k+1} = \overline{u}_{\mu+1} = u_\mu^G$.

Заметим, что т.к. $\|f(\overline{u}_{k+1})\|_2 = \|f_0 - M\overline{z}_k\|_2 = \min\limits_{z \in \mathcal{L}_k} \|f_0 - Mz\|_2$, то, в силу $\mathcal{L}_k \subseteq \mathcal{L}_{k+1}$, имеем: $\|f(\overline{u}_{k+1})\|_2 \geqslant \|f(\overline{u}_{k+2})\|_2$ при всех $k = 0, 1, \ldots$ и равенство $\|f(\overline{u}_{k+1})\|_2 = \|f(\overline{u}_{k+2})\|_2$ равносильно $\overline{z}_k, \overline{z}_{k+1} \in \operatorname{argmin}\limits_{z \in \mathcal{L}_{k+1}} \|f_0 - Mz\|_2$, а т.к. решение этой задачи единственно из невырожденности M, то $\overline{z}_{k+1} = \overline{z}_k \in \mathcal{L}_k$. В случае $k + 1 < \mu$ из $\overline{z}_{k+1} \in \mathcal{L}_k = \mathcal{K}_k$ и показанного ранее следует $u_{k+2} - u_0 \in \mathcal{K}_{k+1} = \mathcal{L}_{k+1}$, откуда получим, что $\mathcal{L}_{k+2} = \operatorname{span}\{\mathcal{L}_{k+1}, u_{k+2} - u_0\} = \mathcal{L}_{k+1}$, но это возможно только при $k+1 \geqslant \mu$ — противоречие, значит, $k + 1 \geqslant \mu$. Таким образом, с учетом результата предыдущего абзаца, было доказано, что $\|f(\overline{u}_1)\|_2 > \|f(\overline{u}_2)\|_2 > \ldots > \|f(\overline{u}_\mu)\| \geqslant \|f(\overline{u}_{\mu+1})\|_2 = \|f(\overline{u}_{k+1})\|_2, k > \mu$. Значит, с учетом предыдущего абзаца, приближения $\overline{u}_k$ начинают стагнировать либо на μ-й, либо на $\mu + 1$-й итерации.

Рассмотрим случай $\mu = \nu$. Тогда при $k \geqslant \mu = \nu$, в силу сходимости GMRES (GCR) на ν-й итерации, получим $\overline{u}_{k+1} = u_\mu^G = u_\nu^G = u^*$, $f(\overline{u}_{k+1}) = f(u^*) = 0$, $u_{k+1} = \overline{u}_{k+1} + \beta_k f(\overline{u}_{k+1}) = u^*$. Заметим также, что

$\|f(\overline{u}_\mu)\| > 0$, т.к. в противном случае $u^G_{\nu-1} = u^G_{\mu-1} = \overline{u}_\mu = u^*$, что противоречит тому, что GMRES (GCR) останавливается ровно на ν-й итерации. Значит, метод Андерсона при $\mu = \nu$ сходится к точному решению не более чем за $\mu + 1$ итерацию. Ранее было показано, что $u_{k+1} - u_0 \in \mathcal{K}_{k+1}$, откуда $\|b - Mu^G_{k+1}\|_2 = \|b - M(u_0 + z^G_{k+1})\|_2 = \min\limits_{z \in \mathcal{K}_{k+1}} \|r_0 - M(u_0 + z)\|_2 \leqslant$
$\leqslant \|b - Mu_{k+1}\|$, а значит, если $u_{k+1} = u^*$, то $u^G_{k+1} = u^*$, откуда получим $k + 1 \geqslant \nu$. Таким образом, при $\mu = \nu$ метод Андерсона сойдется к точному решению либо за ν, либо $\nu + 1$ итерацию. Заметим также, что поскольку $u^G_k = \overline{u}_{k+1}, k \in \overline{0, \mu} = \overline{0, \nu}$ из неравенства на нормы невязок $\overline{u}_k$ получим $\|f(u^G_0)\|_2 > \|f(u^G_1)\|_2 > \ldots > \|f(u^G_{\nu-1})\| > \|f(u^G_\nu)\|_2 = \|f(u^G_k)\|_2 = 0, k > \nu$, откуда $\eta = \nu = \mu$.

Рассмотрим случай $\mu < \nu$. Тогда, в силу первой теоремы и сказанного ранее, $\mathcal{L}_{\mu+1} = \mathcal{L}_\mu = \mathcal{K}_\mu \subset \mathcal{K}_{\mu+1}$, откуда из предыдущей имеем $M\overline{z}_\mu \in \mathcal{L}_\mu$. Т.к. $\overline{z}_\mu \in \mathcal{L}_\mu = \mathcal{K}_\mu$, то $\overline{z}_\mu = \sum\limits_{j=0}^{\mu-1} \xi_j M^j r_0$, откуда $\sum\limits_{j=0}^{\mu-1} \xi_j M^{j+1} r_0 = M\overline{z}_\mu \in \mathcal{L}_\mu =$
$= \mathcal{K}_\mu$ и если $\xi_{\mu-1} \neq 0$ получим $M^\mu r_0 \in \mathcal{K}_\mu$, что влечет $\mathcal{K}_\mu = \mathcal{K}_{\mu+1}$ — противоречие. Значит, $\xi_{\mu-1} = 0$, откуда $\overline{z}_\mu \in \mathcal{K}_{\mu-1} = \mathcal{L}_{\mu-1}$, откуда немедленно следует $\overline{z}_\mu = \overline{z}_{\mu-1}$, а значит $u^G_\mu = \overline{u}_{\mu+1} = \overline{u}_\mu = u^G_{\mu-1}$. Т.о. в случае $\mu < \nu$ при $k \geqslant \mu$ имеем: $\overline{u}_k = u^G_\mu = u^G_{\mu-1}$, а т.к. выше было показано, что $u^G_k = \overline{u}_{k+1} \neq \overline{u}_{k+2} = u^G_{k+1}, k = 0, \ldots, \mu - 2$, то $\eta = \mu - 1$.

Т.к. $\|f(\overline{u}_1)\|_2 > \|f(\overline{u}_2)\|_2 > \ldots > \|f(\overline{u}_\mu)\| \geqslant \|f(\overline{u}_{\mu+1})\|_2 = \|f(\overline{u}_{k+1})\|_2, k > \mu$, $\mu \leqslant \nu$ и в случае $\mu = \nu$ было получено $\|f(\overline{u}_\mu)\| > \|f(\overline{u}_{\mu+1})\|_2 = 0$, а в случае $\mu < \nu$ — $\|f(\overline{u}_\mu)\| = \|f(\overline{u}_{\mu+1})\|_2 = \|f(u^G_{\mu-1})\|_2 > 0$, то возможность обратного перехода от поведения норм невязок $\overline{u}_k$ к связи между μ и ν очевидна. Более того, было показано, что μ однозначно определеяется значением η: при $\eta = \nu$ у нас $\mu = \eta$ и при $\eta < \nu - 1$ у нас $\mu = \eta + 1$. Здесь стоит ометить, что ситуации $\eta = \nu - 1$ быть не может, т.к. в этом случае $u^G_{\nu-1} = u^G_\eta = u^G_{\eta+1} = u^G_\nu = u^*$ и GMRES (GCR) сошелся на $\nu - 1$ итерации, что невозможно.

Таким образом метод Андерсона либо за μ, либо за $\mu + 1$ итерацию сходится к некоторому приближению, не обязательно являющемуся решением. При этом параметр μ однозначно определяется индексом стагнации η метода GMRES (GCR).

Обсудим теперь кратко более общие случаи метода ускорения Андерсона. Большинство результатов про эти случаи получаются из доказанной теоремы.

- Случай Anderson(∞) с произвольными весами ранее был сведен к случаю с ненулевыми весами.

- Можно сформулировать метод Anderson(∞) с рестартами: произведем m итераций Anderson(∞), а затем запустим новый Anderson(∞), использовав в качестве начального приближения $\overline{u}_{m+1}$, тогда, в случае повторения этой

процедуры, в силу доказанной теоремы, получится метод существенно связанный с методом GMRES(m) — обобщенным методом минимальных невязок с рестартами ([7]).

- Метод Anderson(m) (и более общий случай произвольного выбора m_k), также известный как Anderson acceleration with truncation, существенно связан с методом truncated GMRES ([7]).

Подводя итог, отметим, что в большинстве линейных задач предпочтительнее использовать GMRES, поскольку, согласно теореме, метод Андерсона может остановиться сильно раньше требуемой точности в случае стагнации метода GMRES, а также поскольку методы Андерсона и GCR уступает в вычислительной устойчивости методу GMRES.

2.2 Квазиньютоновские методы

В этом разделе мы покажем, что метод ускорения Андерсона можно рассматривать в качестве квазиньютоновского метода. Впервые это было показано в работе [8], а затем существенно расширено в работе [2], которой мы и будем следовать в данном разделе. Вначале мы опишем идеи, лежащие в основе квазиньютоновксих методов, приведем конкрентные примеры и покажем, что метод Андерсона в сущности является квазиньютоновксим методом.

Напомним одну из идей, лежащих в основе метода Ньютона для нахождения решения u^* уравнения $f(u) = 0$ для $f : \mathbb{R}^n \to \mathbb{R}^n$, $f \in C^1$: если обозначить через J якобиан отображения f, то при достаточно малых Δu имеем

$$f(u + \Delta u) \approx f(u) + J(u)\Delta u \tag{2.2.1}$$

Значит, если, используя приближение u_k, на k-й итерации метода найти шаг Δu_k из уравнения

$$J(u_k)\Delta u_k = -f(u_k), \tag{2.2.2}$$

то, в случае малости Δu_k, если положить $u_{k+1} = u_k + \Delta u_k$, получим

$$f(u_{k+1}) \approx f(u_k) + J(u_k)\Delta u_k = 0$$

и в итоге шаг метода Ньютона принимает вид

$$u_{k+1} = u_k - J(u_k)^{-1} f(u_k).$$

Можно показать, что при некоторых дополнительных предположениях на функцию f, приближение u_{k+1} к решению u^* окажется лучше, чем u_k: (достаточно потребовать липшицевость J и невырожденность J в точке u^*, тогда, если u_k было достаточно близко к u^*, то u_{k+1} будет еще ближе, причем, при некоторых условиях, может наблюдаться квадратичная сходимость). Отметим также, что в случае линейного f метод сходится к решению за одну итерацию.

Как можно заметить, метод Ньютона требует вычислять $J(u_k)$ на каждой итерации, что не всегда практично. Поэтому рассматриваются квазиньютоновские методы, которые некоторым образом аппроксимируют $J(u_k)$ матрицей J_k и стараются получать J_{k+1} из J_k за счет малоранговых поправок.

В классических квазиньютоновские методах шаг Δu_k находят из уравнения (2.2.2), в котором значение якобиана заменено на его приближение:

$$J_k \Delta u_k = -f(u_k), \tag{2.2.3}$$

затем для нового приближения J_{k+1} требуют точного выполнения соотношения (2.2.1):

$$J_{k+1} \Delta u_k = \Delta f_k, \tag{2.2.4}$$

где $\Delta f_k = f(u_{k+1}) - f(u_k)$. Это условие в русской литературе называется квазиньютоновским, а в английской носит название secant condition.
Кроме того, на J_{k+1} обычно накладывают следующее ограничение:

$$\forall q : q \perp \Delta u_k \Rightarrow J_{k+1} q = J_k q, \tag{2.2.5}$$

которое называют nochange condition и означающее, что мы не получаем никакой новой информации, переходя от J_k к J_{k+1} вдоль ортогональных Δu_k направлений. Бройден [9] показал, что существует единственный метод, одновременно удовлетворяющий условиям (2.2.4) и (2.2.5):

$$J_{k+1} = J_k + (\Delta f_k - J_k \Delta u_k) \frac{\Delta u_k^T}{\Delta u_k^T \Delta u_k}, \tag{2.2.6}$$

который теперь носит название первого метода Бройдена (Broyden's first method). В работе [10] было показано, что это приближение J_{k+1} минимизирует

$$E(J_{k+1}) = \|J_{k+1} - J_k\|_F^2, \tag{2.2.7}$$

среди всех J_{k+1}, удовлетворяющих квазиньютоновскому условию (2.2.4).
Поскольку на каждом итерации метода необходимо решать систему (2.2.3) для нахождения следующего шага, удобнее сразу строить приближения $J(u_k)^{-1}$ (отметим, что данный подход требует обращения якобиана на первой итерации). Если обозначить $G_k = J_k^{-1}$ и использовать формулу Шермана-Моррисона (Sherman-Morrison), получим:

$$G_{k+1} = G_k + (\Delta u_k - G_k \Delta f_k) \frac{\Delta u_k^T G_k}{\Delta u_k^T G_k \Delta f_k}. \tag{2.2.8}$$

Поскольку нам нужно решать систему с якобианом, можно попробовать строить приближения G_k к $J(u_k)^{-1}$, а не к $J(u_k)$, заменив условия (2.2.4), (2.2.5) их аналогами:

$$G_{k+1} \Delta f_k = \Delta u_k, \tag{2.2.9}$$

$$\forall q : q \perp \Delta f_k \Rightarrow G_{k+1} q = G_k q. \tag{2.2.10}$$

Как и раньше, существует единственный метод, одновременно удовлетворяющий условиям (2.2.9) и (2.2.10):

$$G_{k+1} = G_k + (\Delta u_k - G_k \Delta f_k) \frac{\Delta f_k^T}{\Delta f_k^T \Delta f_k}, \tag{2.2.11}$$

носящий название второго метода Бройдена (Broyden's second method) и минимизируещий

$$E(G_{k+1}) = \|G_{k+1} - G_k\|_F^2, \tag{2.2.12}$$

среди всех G_{k+1}, удовлетворяющих условию (2.2.9).

Перейдем к рассмотрению обобщенного метода Бройдена второго типа (generalized Broyden's second method), в котором аппроксимируем $J(u_k)^{-1}$. В его основе лежит использование на каждой итерации нескольких квазиньютоновских условий (2.2.9): пусть $\Delta f_{k-m_k}, \ldots, \Delta f_{k-1}$ линейно независимы и $m_k \leqslant n$, тогда для приближения G_k к $J(u_k)^{-1}$ потребуем

$$G_k \Delta f_j = \Delta u_j, j = k - m_k, \ldots, k-1.$$

Если обозначить $\mathcal{F}_k = \begin{bmatrix} \Delta f_{k-m_k} & \ldots & \Delta f_{k-1} \end{bmatrix}$, $\mathcal{U}_k = \begin{bmatrix} \Delta u_{k-m_k} & \ldots & \Delta u_{k-1} \end{bmatrix}$, то это условие можно записать в виде

$$G_k \mathcal{F}_k = \mathcal{U}_k. \tag{2.2.13}$$

Аналогом условия (2.2.10) в данном случае является

$$\forall q : q \perp \operatorname{span}\{\Delta f_{k-m_k}, \ldots, \Delta f_{k-1}\} \Rightarrow G_k q = G_{k-m_k} q. \tag{2.2.14}$$

Т.к. $\operatorname{span}\{\Delta f_{k-m_k}, \ldots, \Delta f_{k-1}\} = \operatorname{Im} \mathcal{F}_k$, то условие на q может быть записано в виде $\mathcal{F}_k^T q = 0$, и из (2.2.14) следует $\ker \mathcal{F}_k^T \subseteq \ker (G_k - G_{k-m_k})$, что равносильно существованию матрицы Z, что $G_k - G_{k-m_k} = Z \mathcal{F}_k^T$ (т.к. $\ker B = (\operatorname{Im} B^T)^\perp$, то условие выше влечет $\operatorname{Im} (G_k - G_{k-m_k})^T \subseteq \operatorname{Im} (\mathcal{F}_k^T)^T$, что означает, что строки $G_k - G_{k-m_k}$ линейно выражаются через строки $\mathcal{F}_k^T$). Т.к. $\Delta f_{k-m_k}, \ldots, \Delta f_{k-1}$ линейно независимы и $m_k \leqslant n$, то $\operatorname{rank}(\mathcal{F}_k^T \mathcal{F}_k) = \operatorname{rank} \mathcal{F}_k = m_k$ и т.к. $\mathcal{F}_k^T \mathcal{F}_k \in \mathbb{R}^{m_k \times m_k}$, то матрица $\mathcal{F}_k^T \mathcal{F}_k$ невырождена и из условия (2.2.13) получим

$$\begin{gathered} G_k - G_{k-m_k} = Z \mathcal{F}_k^T \Rightarrow (G_k - G_{k-m_k}) \mathcal{F}_k = Z \mathcal{F}_k^T \mathcal{F}_k \Rightarrow \\ \Rightarrow Z = (G_k - G_{k-m_k}) \mathcal{F}_k (\mathcal{F}_k^T \mathcal{F}_k)^{-1} = (\mathcal{U}_k - G_{k-m_k} \mathcal{F}_k)(\mathcal{F}_k^T \mathcal{F}_k)^{-1}, \end{gathered}$$

откуда получим формулу для обновления G_k:

$$G_k = G_{k-m_k} + Z \mathcal{F}_k^T = G_{k-m_k} + (\mathcal{U}_k - G_{k-m_k} \mathcal{F}_k)(\mathcal{F}_k^T \mathcal{F}_k)^{-1} \mathcal{F}_k^T, \tag{2.2.15}$$

представляющую из себя m-ранговое обновление.
Тогда, если обозначить через $f_k = f(u_k)$, формула для u_{k+1} примет вид:

$$\begin{aligned} u_{k+1} = u_k + \Delta u_k = u_k - G_k f_k = \\ = u_k - G_{k-m_k} f_k - (\mathcal{U}_k - G_{k-m_k}\mathcal{F}_k)(\mathcal{F}_k^T\mathcal{F}_k)^{-1}\mathcal{F}_k^T f_k = \\ = u_k - G_{k-m_k} f_k - (\mathcal{U}_k - G_{k-m_k}\mathcal{F}_k)\gamma_k, \end{aligned} \tag{2.2.16}$$

где введено обозначение γ_k для решения нормального уравнения

$$(\mathcal{F}_k^T\mathcal{F}_k)\gamma_k = \mathcal{F}_k^T f_k,$$

что эквивалентно решению задачи наименьших квадратов

$$\min_{\gamma} \|f_k - \mathcal{F}_k\gamma\|_2. \tag{2.2.17}$$

Заметим, что мы показали, что полученная G_k определяется единственным образом из условий (2.2.13), (2.2.14). Кроме того, можно показать, что полученная G_k минимизирует

$$E(G_k) = \|G_k - G_{k-m_k}\|_F^2, \tag{2.2.18}$$

среди всех матриц, удовлетворяющих (2.2.13).
Отметим также, что в случае $m = n$ формула (2.2.16) принимает вид

$$u_{k+1} = u_k - \mathcal{U}_k\mathcal{F}_k^{-1} f_k,$$

что есть ни что иное как метод секущих (secant method).

Теперь мы готовы обсудить связь метода Андерсона с квазиньютоновскими методами. Напомним, что в разделе 1.4 мы записали метод Андерсона в виде (1.4.1), (1.4.2):

$$\begin{aligned} \gamma_k = \underset{\gamma=(\gamma_0,\ldots,\gamma_{m_k-1})}{\operatorname{argmin}} \|f_k - \mathcal{F}_k\gamma\|_2, \\ u_{k+1} = u_k + \beta_k f_k - (\mathcal{U}_k + \beta_k\mathcal{F}_k)\gamma_k, \end{aligned}$$

но это есть ни что иное, как формулы (2.2.16), (2.2.17) в которых положили $G_{k-m_k} = -\beta_k I$, что в силу сказанного раннее означает, что метод Андерсона — квазиньютоновский метод, который строит G_k, аппроксимирующие $J(u_k)^{-1}$, минимизирующие $E(G_k) = \|G_k + \beta_k I\|_F^2$ среди матриц, удовлетворяющих обобщенному квазиньютоновскому условию второго типа (2.2.13) $G_k\mathcal{F}_k = \mathcal{U}_k$.

Впервые этот результат был получен в работе [8]. Как можно видеть, метод Андерсона — метод второго типа, т.к. аппроксимирует $J(u_k)^{-1}$. В работе [2] аналогично показанному раннее выводится метод Андерсона первого типа (type-I Andeson), который строит приближения J_k к $J(u_k)$, минимизирующие $E(J_k) = \|J_k + \dfrac{1}{\beta_k} I\|_F^2$

среди матриц, удовлетворяющиехобобщенному квазиньютоновскому условию первого типа $J_k\mathcal{U}_k = \mathcal{F}_k$. Отметим, что от классической записи метода Андерсона, метод Андерсона первого типа будет отличаться только выбором α^k. Более того в работе [2] вводится целое семейство методов Андерсона, частью которого являются методы Андерсона первого и второго типа. Описание этих методов выходит за рамки данной работы.

2.3 EDIIS или способ глобализации сходимости

Данный раздел напрямую не относится к методу Андерсона, поскольку описывает другой метод решения задачи на нахождение неподвижной точки, но, поскольку эти методы по своей сути достаточно близки, необходимо осветить эту связь и важность этого метода в приложениях.

Напомним, что метод ускорения Андерсона призван ускорить и расширить область сходимости метода простой итерации. Как показывает практика, в некоторых приложениях метод ускорения Андерсона может хорошо сходится в случаях, когда отображение g не является сжимающем, однако на данный момент почти все результаты о сходимости метода Андерсона рассматривают случаи сжимающего отображения. Именно эти результаты мы намерены привести в следующей части. Кроме того, как было показано в работе [11], даже в случае сжимающего отображения метод обладает лишь локальной сходимостью: при плохом выборе начального приближения метод может разойтись или проявить крайне низкую скорость сходимости. В той же работе была предложена одна модификация метода ускорения Андерсона, который, напомним, также носит название DIIS (Direct Inversion on the Iterative Subspace), которую назвали EDIIS (Energy Direct Inversion on the Iterative Subspace) и которая оказалась лишена этого недостатка: в работе [12] доказана глобальная сходимоть метода EDIIS в случае сжимающего отображения, что позволяет использовать этот метод для нахождения хороших начальных приближений для метода Андерсона (DIIS).

Идея метода EDIIS совпадает с идеей метода DIIS, за исключением того, что задача минимизации (1.1.1) сужается на на афинную оболочку $\mathrm{aff}\{u_{k-m_k}, \dots, u_k\}$ предыдущих приближений, а на выпуклую: $\mathrm{conv}\{u_{k-m_k}, \dots, u_k\}$. Т.к. выпуклая линейная комбинация является и афинной, то все переходы, сделанные при выводе метода Андерсона, остаются в силе, а потому запись метода EDIIS отличается от DIIS только задачей минимизации.

В следующем разделе мы повторим доказательство [12] глобальной сходимоти метода EDIIS в случае сжимающего отображения. Мы также покажем оценки локальной сходимости этого метода, которые окажутся хуже, чем соответствующие у DIIS. На практике, если приближение выбрано достаточно хорошо, метод EDIIS уступает методу DIIS, что неудивительно, ведь в этом методе мы сильнее сузили исходную задачу минизации.

Отметим, что, как и в случае с методом Андерсона, если метод EDIIS с

Algorithm EDIIS (Energy Direct Inversion on the Iterative Subspace)

input $g : U \to U$, $u_0 \in U$, $\{m_k : m_k \leqslant k\}$, $\{\beta_k\} \subset \mathbb{P}$

for $k = 0, 1 \ldots,$ **do**

$$f_k = g(u_k) - u_k$$

$$(\alpha_0^k, \ldots, \alpha_{m_k}^k) = \underset{\sum_{j=0}^{m_k} \alpha_j = 1,\, \alpha_j \geqslant 0}{\operatorname{argmin}} \left\| \sum_{j=0}^{m_k} \alpha_j f_{k-m_k+j} \right\|$$

$$u_{k+1} = \sum_{j=0}^{m_k} \alpha_j^k u_{k-m_k+j} + \beta_k \sum_{j=0}^{m_k} \alpha_j^k f_{k-m_k+j}$$

end for

$m_k = \min\{m, k\}$ обозначается EDIIS(m).

3 Сходимость

В данном разделе мы обсудим сходимость методов DIIS и EDIIS в случае сжимающего отображения и $\beta_k \equiv 1$. Мы покажем, что в линейном случае метод DIIS сходится глобально не хуже метода простой итерации. После этого мы покажем глобальную сходимость EDIIS в случае сжимающего отображения. Затем мы покажем, что в нелинейном случае сжимающего отображения за счет выбора достаточно хорошего начального приближения возможно сколь угодно близко приблизиться к скорости метода простой итерации при условии равномерной ограниченности $\|\alpha^k\|_1$.

Отметим, что впервые теория сходимостей для метода DIIS была предложена в работе [5], а теория сходимости для метода EDIIS в работе [12].

Отметим, что на практике метод Андерсона часто сходится гораздо быстрее метода простой итреции и может сходится даже в случае несжимающего отображения, однако метод также может расходится в случае сжимающего отображения, для чего его необходимо использовать в связке с EDIIS.

3.1 Понятие скорости сходимости

Для дальнейших рассуждений нам понадобится определить несколько видов сходимости.

Будем говорить, что последовательность $\{u_k\}_{k=0}^{\infty}$ обладает q-линейной сходимостью (linear quotient-convergence) к u^* с множителем $\alpha \in [0; 1)$, если

$$\|u_{k+1} - u^*\| \leqslant \alpha \|u_k - u^*\|. \tag{3.1.1}$$

Будем говорить, что последовательность $\{u_k\}_{k=0}^{\infty}$ обладает r-линейной сходимостью (linear root-convergence) к u^* с множителем $\alpha \in [0; 1)$, если существует

$M > 0$, что

$$\|u_k - u^*\| \leqslant M\alpha^k \|u_0 - u^*\|. \tag{3.1.2}$$

Заметим, что q-линейная сходимость с множителем α влечет r-линейную сходимость с множителем α и $M = 1$.

Будем говорить, что итерационный метод обладает q-линейной (r-линейной) сходимостью, если последовательность приближений сходится к точному решению q-линейно (r-линейно). Например, если $g : D \to D$ — сжимающее отображение с константой $c \in [0; 1)$ в замкнутом подмножестве $D \subseteq U$, то метод простой итерации при выборе начального приближения $u_0 \in D$ обладает q-линейной сходимостью с множителем c. В дальнейшем мы покажем, что метод Андерсона при некоторых предположениях обладает как минимум r-линейной сходимостью.

3.2 Глобальная q-линейная сходимость невязок в методе DIIS для линейных задач с сжимающим отображением

Пусть $g(u) = Au + b, A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$, $M = I - A$, $f(u) = g(u) - u = b - Mu$. Если $\|A\| = c < 1$, то $\|g(u) - g(v)\| = \|A(u - v)\| \leqslant \|A\|\|u - v\| = c\|u - v\|$ — отображение g является сжимающим с константой $c < 1$ и метод простой итерации сходится q-линейно с множителем c. Покажем, что в линейном случае метод Андерсона сходится не хуже метода простой итерации.

Теорема ([5]). *Пусть $g(u) = Au + b$, $\|A\| = c < 1$, тогда метод Андерсона с $\beta_k \equiv 1$ при любом $u_0 \in \mathbb{R}^n$ r-линейно сходится к u^* и невязки $f(u_k)$ сходятся q-линейно к 0 с множителем c.*

Доказательство:
Как было показано ранее, в линейном случае 2.1.1 $\overline{f}_{k+1} = f(\overline{u}_{k+1})$ и из задачи минимизации 1.1.4 имеем $\overline{f}_{k+1} = \min\limits_{\sum\limits_{j=0}^{m_k}\alpha_j=1} \left\|\sum\limits_{j=0}^{m_k} \alpha_j f(u_{k-m_k+j})\right\| \leqslant \|f(u_k)\|$.

Тогда, поскольку $u_{k+1} = \overline{u}_{k+1} + \beta_k\overline{f}_{k+1}$, имеем

$$\begin{gathered} f(u_{k+1}) = b - Mu_{k+1} = b - M(\overline{u}_{k+1} + \beta_k\overline{f}_{k+1}) = b - M\overline{u}_{k+1} - \beta_k M\overline{f}_{k+1} = \\ = f(\overline{u}_{k+1}) - \beta_k M\overline{f}_{k+1} = \overline{f}_{k+1} - \beta_k M\overline{f}_{k+1} = (I - \beta_k(I - A))\overline{f}_{k+1}. \end{gathered}$$

Из сказанного раннее следует, что

$$\begin{gathered} \|f(u_{k+1})\| \leqslant \|I - \beta_k(I - A)\|\|\overline{f}_{k+1}\| \leqslant \|I - \beta_k(I - A)\|\|f(u_k)\| = \\ = \|(1 - \beta_k)I + \beta_k A\|\|f(u_k)\| = \|A\|\|f(u_k)\| \leqslant c\|f(u_k)\|, \end{gathered}$$

т.е. невязки $f(u_k)$ q-линейно сходятся к 0 с множителем c, а значит метод

Андерсона с $\beta_k \equiv 1$ сходится к u^*.

Т.к. $f(u) = b - Mu = Mu^* - Mu = M(u^* - u) = -(I - A)(u - u^*)$, то т.к. $\|(I - A)\| \leqslant \|I\| + \|A\| = 1 + c$, $\|(I - A)^{-1}\| \leqslant \dfrac{1}{1 - \|A\|} = \dfrac{1}{1 - c}$, имеем

$$\|u_k - u^*\| \leqslant \|(I - A)^{-1}\|\|f(u_k)\| \leqslant \frac{1}{1 - c} c^k \|f(u_0)\| \leqslant$$
$$\leqslant \frac{1}{1 - c} c^k (1 + c)\|u_0 - u^*\| = \frac{1 + c}{1 - c} c^k \|u_0 - u^*\|,$$

что означает r-линейную сходимость метода.

3.3 Глобальная r-линейная сходимость метода EDIIS(m) для произвольных задач с сжимающим отображением

В данном разделе мы покажем, что в случае сжимающего отображения, метод EDIIS(m) r-линейно сходится к точному решению из любого начального приближения. Однако, как будет видно позже, множитель в оценке r-линейной сходимости будет достаточно велик, однако позже будет показана локальная сходимость метода EDIIS(m) с меньшим множителем, что говорит о ускорении метода после нескольких итераций.

Теорема ([12]). *Пусть $g : D \to D$ — сжимающее отображение с константой c, где D — выпуклое подмножество U. Тогда при $u_0 \in D$ метод EDIIS(m) с $\beta_k \equiv 1$ сходится к неподвижной точке u^* отображения g r-линейно с множителем $\hat{c} = c^{\frac{1}{m+1}}$, т.е.*

$$\|u_k - u^*\| \leqslant \hat{c}^k \|u_0 - u^*\|.$$

Доказательство:
Индукцией по k покажем r-линейную сходимость метода, а также принадлежность $u_k \in D$. База индукции очевидна. Переход основан на том, что в силу $\beta_k \equiv 1$ имеем

$$u_{k+1} = (1 - \beta_k)\overline{u}_{k+1} + \beta_k \overline{g}_{k+1} = \overline{g}_{k+1} = \sum_{j=0}^{m_k} \alpha_j^k g(u_{k-m_k+j}),$$

где $\sum\limits_{j=0}^{m_k} \alpha_j^k = 1, \alpha_j^k \geqslant 0$. Т.к. $u_{k-m_k+j} \in D$ по предположению индукции, то $g(u_{k-m_k+j}) \in D$ в силу $g : D \to D$, откуда в силу выпуклости D получим $u_{k+1} \in D$, как выпуклая комбинация векторов из D.

Кроме того,

$$\|u_{k+1}-u^*\| = \left\|\sum_{j=0}^{m_k}\alpha_j^k g(u_{k-m_k+j})-u^*\right\| = \left\|\sum_{j=0}^{m_k}\alpha_j^k (g(u_{k-m_k+j})-u^*)\right\| =$$
$$= \left\|\sum_{j=0}^{m_k}\alpha_j^k (g(u_{k-m_k+j})-g(u^*))\right\| \leqslant \sum_{j=0}^{m_k}\alpha_j^k \|g(u_{k-m_k+j})-g(u^*)\| \leqslant$$
$$\leqslant \sum_{j=0}^{m_k}\alpha_j^k c\|u_{k-m_k+j}-u^*\| \leqslant c\sum_{j=0}^{m_k}\alpha_j^k \hat{c}^{k-m_k+j}\|u_0-u^*\| \overset{k-m_k+j\geqslant k-m}{\leqslant}$$
$$\leqslant c\sum_{j=0}^{m_k}\alpha_j^k \hat{c}^{k-m}\|u_0-u^*\| = c\hat{c}^{k-m}\|u_0-u^*\| =$$
$$= \hat{c}^{k+1}(c\hat{c}^{-m-1})\|u_0-u^*\| = \hat{c}^{k+1}\|u_0-u^*\|,$$

что доказывает переход индукции.

3.4 Локальная r-линейная сходимость методов DIIS и EDIIS на нелинейных задачах

Рассмотрим класс методов, для которых на k-й итерации переход от $m_k \leqslant k$ известных приближений $\{u_{k-m_k+j}\}_{j=0}^{m_k}$ и значений $\{g(u_{k-m_k+j})\}_{j=0}^{m_k}$ осуществляется по формуле

$$u_{k+1} = \sum_{j=0}^{m_k}\alpha_j^k g(u_{k-m_k+j}), \tag{3.4.1}$$

где коэффициенты α_j^k удовлетворяют следующим двум условиям:

$$\sum_{j=0}^{m_k}\alpha_j^k = 1 \tag{3.4.2}$$

и

$$\|\sum_{j=0}^{m_k}\alpha_j^k f(u_{k-m_k+j})\| \leqslant \|f(u_k)\|. \tag{3.4.3}$$

Заметим, что методы DIIS и EDIIS при $\beta_k \equiv 1$ удовлетворяют этим условиям. Введение данного класса методов позволит нам доказать теоремы, верные, как для DIIS, так и для EDIIS.

В данном разделе мы лишь коротко сформулируем известные результаты.

Теорема ([13]). *Пусть U — гильбертово пространство со скалярным произведением $(\cdot,\cdot)$. Рассмотрим метод из описанного класса с $m_k = \min\{1,k\}$, в котором*

используется норма U. Пусть отображение g осуществляет сжимающее отображение с константой $c < 2 - \sqrt{3}$ в шаре $\mathcal{B}(u^*, \rho) = \{u : \|u - u^*\| \leqslant \rho\}$. Тогда, если $\hat{c} = \dfrac{3c - c^2}{1 - c} < 1$ и u_0 достаточно близко к u^*, то невязки $f(u_k)$ метода q-линейно сходятся к нулю с множителем $\hat{c}$ и u_k сходятся r-линейно к u^* с множителем $\hat{c}$.

Эта теорема была впервые получена в работе [5], однако с более жесткими требованиями на отображение g (требовалась липшицевая непрерывно дифференцируемость), позднее этот результат несколько раз поправляли, пока не был получен описанный вариант в работе [13], в той же работе приведены результаты сходимости при произвольных m_k для нелинейных отображений, распадающихся в сумму гладкой и негладкой липшицевой константой.

Для случая произвольных m_k и произвольных норм необходимо потребовать равномерную ограниченность $\|\alpha^k\|_1$:

$$\sum_{j=0}^{m_k} |\alpha_j^k| \leqslant M_\alpha, k = 0, 1, \ldots$$

В этом случае верна теорема

Теорема ([12]). *Пусть U — банахово пространство и g осуществляет сжимающее отображение с константной $c \in [0; 1)$ на выпуклом замкнутом множестве $D \subset U$, пусть кроме того, g непрерывно дифференцируема в шаре $\mathcal{B}(u^*, \rho) \subset D$. Тогда если $u_0 \in \mathcal{B}(u^*, \rho)$ достаточно близко к u^*, то $u_k \in \mathcal{B}(u^*, \delta)$ и имеет место r-линейная сходимость невязок метода из описанного класса к 0 с множителем c, причем*

$$\limsup_{k\to\infty} \left(\frac{\|f(u_k)\|}{\|f(u_0)\|}\right)^{\frac{1}{k}} \leqslant c, \limsup_{k\to\infty} \left(\frac{\|u_k - u^*\|}{\|u_0 - u^*\|}\right)^{\frac{1}{k}} \leqslant c.$$

Данная теорема впервые была получена в работе [5], но при более жестких предположениях.

Как мы видим, случай произвольного выбора m_k для теоритического анализа требует равномерную ограниченность $\|\alpha^k\|_1$. На практике она обычно наблюдается и ее можно получить за счет динамического изменения m_k аналогично тому, как это предлагалось делать в разделе 1.4, когда обсуждали способы следить за обусловленностью задачи наименьших квадратов. В работе [14] доказано, что в случае динамического изменения m_k для поддержания заданного уровня $\|\alpha^k\|_1$, r-линейная сходимость невязок к нулю становится q-линейной, однако, как показала практика такой метод не всегда работает лучше, чем обычный.

4 Заключение

В данной работе перечисляются и воспроизводятся известные результаты, посвященные методу ускорения Андерсона. Данный метод оказывается полезен в различных приложениях. Мы обсудили описание метода, способы его реализации, связь с другими методами и кратко затронули вопросы сходимости. Мы также обсудили модификацию метода, носящую название EDIIS, которая, обладая глобальной сходимостью, позволяет выбрать хорошее стартовое приближение для запуска метода Андерсона.

В дальнейшем читателю рекомендуется ознакомиться с предложенным списком литературы и самостоятельно поставить численные эксперименты.

Список литературы

[1] D.G. Anderson, *Iterative procedures for nonlinear integral equations*, J. Assoc. Comput. Mach. 12 (1965) 547–560.

[2] H. Fang, Y. Saad, *Two classes of multisecant methods for nonlinear acceleration*, Numer. Linear Algebra Appl. 16 (3) (2009) 197–221

[3] A. Potra and H. Engler, *A characterization of the behavior of the Anderson acceleration on linear problems*, Linear Algebra and its Applications, vol. 438, no. 3, (2013) 1002–1011.

[4] H.F. Walker, P. Ni, *Anderson acceleration for fixed-point iterations*, SIAM J. Numer. Anal. 49 (4) (2011) 1715–1735

[5] A. Toth and C. T. Kelley, *Convergence analysis for Anderson acceleration*, SIAM J. Numer. Anal., 53 (2015), pp. 805–819.

[6] H.F. Walker, *Anderson Acceleration: Algorithms and Implementations*, Research Report, MS-6-15-50, Worcester Polytechnic Institute Mathematical Sciences Department (2011).

[7] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[8] V. Eyert, *A comparative study on methods for convergence acceleration of iterative vector sequences*, J. Comput. Phys., 124 (1996), pp. 271–285.

[9] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19:577–593, 1965.

[10] Jr. J. E. Dennis and J. J. More. *Quasi-Newton methods: Motivation and theory.* SIAM Rev., 19(1):46–89, 1977.

[11] K. N. Kudin, G. E. Scuseria, and E. Cances, *A black-box self-consistent field convergence algorithm: One step closer*, J. Chem. Phys., 116 (2002), pp. 8255–8261.

[12] X. Chen, C. T. Kelley, *Convergence of the EDIIS Algorithm for Nonlinear Equations*, SIAM J. Sci. Comput. 41(1), A365–A379 (2019)

[13] W.Bian, X. Chen, C. T. Kelley, *Anderson Acceleration for a Class of Nonsmooth Fixed-Point Problems*, SIAM J. Sci. Comput. 43(5), S1-S20 (2021)

[14] A. Toth *A theoretical analysis of Anderson acceleration and its application in multiphysics simulation for light-water reactors*, PhD thesis, North Carolina State University (2016).