Anderson Acceleration

C. T. Kelley, Alex Toth, Austin Ellis
NC State University
tim_kelley@ncsu.edu
Supported by NSF, DOE(CASL/ORNL), ARO

Wamplerfest, June 2017

Outline

Collaborators

- My NCSU Students: Alex Toth, Austin Ellis
- Multiphysics coupling
  - ORNL: Steven Hamilton, Stuart Slattery, Kevin Clarno, Mark Berrill, Tom Evans
  - Sandia: Roger Pawlowski, Alex Toth
- Electronic Structure Computations at NCSU
  - Jerry Bernholc, Emil Briggs, Miro Hodak, Elena Jakubikova, Wenchang Lu
- Hong Kong Polytechnic: Xiaojun Chen
- LLNL: Carol Woodward, Jean-Luc Fattebert

Anderson Acceleration Algorithm

Solve fixed point problems

$$\mathbf{u} = \mathbf{G}(\mathbf{u})$$

faster than Picard iteration

$$\mathbf{u}_{k+1} = \mathbf{G}(\mathbf{u}_k).$$

Motivation (Anderson 1965) SCF iteration in electronic structure computations.

Why not Newton?

Newton's method

$$\mathbf{u}_{k+1} = \mathbf{u}_k - (\mathbf{I} - \mathbf{G}'(\mathbf{u}_k))^{-1}(\mathbf{u}_k - \mathbf{G}(\mathbf{u}_k))$$

- converges faster,
- does not require that $\mathbf{G}$ be a contraction,
- needs $\mathbf{G}'(\mathbf{u})$ or $\mathbf{G}'(\mathbf{u})\mathbf{w}$.

Sometimes you will not have $\mathbf{G}'$.

Electronic Structure Computations

Nonlinear eignevalue problem: Kohn-Sham equations

$$\mathbf{H}_{ks}[\psi_i] = -\frac{1}{2}\nabla^2\psi_i + V(\rho)\psi_i = \lambda_i\psi_i \quad i = 1, \ldots, N$$

where the charge density is

$$\rho = \sum_{i=1}^{N} |\psi_i|^2.$$

Write this as

$$\mathbf{H}(\rho)\Psi = \Lambda\Psi$$

Self-Consistent Field iteration (SCF)

Given ρ

- Solve the linear eigenvalue problem

$$\mathbf{H}(\rho)\Psi = \Lambda\Psi$$

  for the N eigenvalues/vectors you want.
- Update the charge density via

$$\rho \leftarrow \sum_{i=1}^{N} |\psi_i|^2.$$

- Terminate if change in ρ is sufficiently small.

This is in the backend of most quantum chemistry/physics codes.

SCF as a fixed-point iteration

SCF is a fixed point iteration

$$\rho \leftarrow \mathbf{G}(\rho)$$

Not clear how to differentiate $\mathbf{G}$

- termination criteria in eigen-solver
- multiplicities of eigenvalues not known at the start

Anderson Acceleration

anderson($\mathbf{u}_0, \mathbf{G}, m$)

$\mathbf{u}_1 = \mathbf{G}(\mathbf{u}_0)$; $\mathbf{F}_0 = \mathbf{G}(\mathbf{u}_0) - \mathbf{u}_0$

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

$m_k \leq \min(m, k)$

$\mathbf{F}_k = \mathbf{G}(\mathbf{u}_k) - \mathbf{u}_k$

Minimize $\| \sum_{j=0}^{m_k} \alpha_j^k \mathbf{F}_{k-m_k+j} \|$ subject to $\sum_{j=0}^{m_k} \alpha_j^k = 1$.

$\mathbf{u}_{k+1} = \sum_{j=0}^{m_k} \alpha_j^k \mathbf{G}(\mathbf{u}_{k-m_k+j})$

end for

Other names for Anderson

- Pulay mixing (Pulay 1980)
- Direct iteration on the iterative subspace (DIIS) Rohwedder/Scheneider 2011
- Nonlinear GMRES (Washio 1997)

Terminology

- m, depth. We refer to Anderson(m).
  Anderson(0) is Picard.
- $\mathbf{F}(\mathbf{u}) = \mathbf{G}(\mathbf{u}) - \mathbf{u}$, residual
- $\{\alpha_j^k\}$, coefficients
  Minimize $\| \sum_{j=0}^{m_k} \alpha_j^k \mathbf{F}_{k-m_k+j} \|$ subject to $\sum_{j=0}^{m_k} \alpha_j^k = 1$.
  is the optimization problem.
- $\| \cdot \|$, ℓ^2, ℓ^1, or ℓ^∞

Solving the Optimization Problem

Solve the linear least squares problem:

$$\min \left\| \mathbf{F}_k - \sum_{j=0}^{m_k-1} \alpha_j^k (\mathbf{F}_{k-m_k+j} - \mathbf{F}_k) \right\|_2^2,$$

for $\{\alpha_j^k\}_{j=0}^{m_k-1}$ and then

$$\alpha_{m_k}^k = 1 - \sum_{j=0}^{m_k-1} \alpha_j^k.$$

More or less what's in the codes.
LP solve for $\|\cdot\|_1$ and $\|\cdot\|_\infty$. That's bad for our customers.

Convergence Theory

- Most older work assumed unlimited storage or very special cases.
  - For unlimited storage, Anderson looks like a Krylov method and it is equivalent to GMRES (Walker-Ni 2011).
  - Anderson is also equivalent to a multi-secant quasi-Newton method (Fang-Saad + many others).
- In practice $m \leq 5$ most of the time
  and 5 is generous.
- The first general convergence results for the method
  as implemented in practice are ours.

Convergence Results: Toth-K 2015

Critical idea: prove acceleration instead of convergence.

- Assume $\mathbf{G}$ is a contraction, constant c.
  Objective: do no worse than Picard
- Local nonlinear theory; $\|\mathbf{e}_0\|$ is small.
- Better results for $\|\cdot\|_2$.

Linear Problems, Toth, K 2015

Here

$$\mathbf{G}(\mathbf{u}) = \mathbf{M}\mathbf{u} + \mathbf{b}, \ \|\mathbf{M}\| \leq c < 1, \ \text{and} \ \mathbf{F}(\mathbf{u}) = \mathbf{b} - (\mathbf{I} - \mathbf{M})\mathbf{u}.$$

Theorem: $\|\mathbf{F}(\mathbf{u}_{k+1})\| \leq c\|\mathbf{F}(\mathbf{u}_k)\|$

Proof: I

Since $\sum \alpha_j = 1$, the new residual is

$$\begin{aligned}
\mathbf{F}(\mathbf{u}_{k+1}) &= b - (I - \mathbf{M})\mathbf{u}_{k+1} \\
&= \textstyle\sum_{j=0}^{m_k} \alpha_j \left[b - (I - \mathbf{M})(b + \mathbf{M}\mathbf{u}_{k-m_k+j}) \right] \\
&= \textstyle\sum_{j=0}^{m_k} \alpha_j \mathbf{M} \left[b - (I - \mathbf{M})\mathbf{u}_{k-m_k+j} \right] \\
&= \mathbf{M} \textstyle\sum_{j=0}^{m_k} \alpha_j \mathbf{F}(\mathbf{u}_{k-m_k+j})
\end{aligned}$$

Take norms to get ...

Proof: II

$$\|\mathbf{F}(\mathbf{u}_{k+1})\| \leq c \left\| \sum_{j=0}^{m_k} \alpha_j \mathbf{F}(\mathbf{u}_{k-m_k+j}) \right\|$$

Optimality implies that

$$\left\| \sum_{j=0}^{m_k} \alpha_j \mathbf{F}(\mathbf{u}_{k-m_k+j}) \right\| \leq \|\mathbf{F}(\mathbf{u}_k)\|.$$

That's it.
Use Taylor for the nonlinear case, which means local convergence.

Assumptions: $m = 1$

- There is $\mathbf{u}^* \in R^N$ such that $\mathbf{F}(\mathbf{u}^*) = \mathbf{G}(\mathbf{u}^*) - \mathbf{u}^* = 0$.
- $\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| \leq c\|u - v\|$ for $\mathbf{u}, \mathbf{v}$ near $\mathbf{u}^*$.
- $\mathbf{G}$ is continuously differentiable near $\mathbf{u}^*$

$\mathbf{G}$ *has a fixed point and is a smooth contraction in a neighborhood of that fixed point.*

Convergence for Anderson(1) with ℓ^2 optimization

Anderson(1) converges and

$$\limsup_{k\to\infty} \frac{\|\mathbf{F}(\mathbf{u}_{k+1})\|_2}{\|\mathbf{F}(\mathbf{u}_k)\|_2} \leq c.$$

Very special case:

- Optimization problem is trivial.
- No iteration history to keep track of.

On the other hand . . .

Assumptions: $m > 1$, any norm

- The assumptions for $m = 1$ hold.
- There is M_α such that for all $k \geq 0$

$$\sum_{j=1}^{m_k} |\alpha_j| \leq M_\alpha.$$

Do this by

- Hoping for the best.
- Reduce m_k until it happens.
- Reduce m_k for conditioning(?)

Convergence for Anderson(m), any norm.

Toth-K, Chen-K
If u_0 is sufficiently close to u^* then the Anderson iteration converges to u^* r-linearly with r-factor no greater than $\hat{c}$. In fact

$$\limsup_{k\to\infty} \left(\frac{\|F(u_k)\|}{\|F(u_0)\|} \right)^{1/k} \leq c. \tag{1}$$

Anderson acceleration is not an insane thing to do.

Comments

- The local part is serious and is a problem in the chemistry codes.
- No guarantee the convergence is monotone. See this in practice.
- The conditioning of the least squares problem can be poor.
  But that has only a small effect on the results.
- The results do not completely reflect practice in that...
  - Theory seems to be sharp for some problems. But . . .
    convergence can sometimes be very fast. Why?
  - Convergence can depend on physics.
    The mathematics does not yet reflect that.
  - There are many variations in the chemistry/physics literature,
    which are not well understood.

EDIIS: Kudin, Scuseria, Cancès 2002

EDIIS (Energy DIIS) globalizes Anderson by constraining $\alpha_j^k \geq 0$. The optimization problem is

$$\text{Minimize } \left\| \mathbf{F}_k - \sum_{j=0}^{m_k-1} \alpha_j^k (\mathbf{F}_{k-m_k+j} - \mathbf{F}_k) \right\|_2^2 \equiv \|\mathbf{A}\alpha^k - \mathbf{F}_k\|_2^2.$$

subject to

$$\sum_{j=0}^{m_k-1} \alpha_j^k \leq 1, \alpha_j^k \geq 0.$$

This could be trouble

- This is a QP and we'd have to compute $\mathbf{A}^T\mathbf{A}$.
  $\mathbf{A}$ is often very ill-contitioned.
- We used QR before which exposed the ill-contitioning less badly.
- You're looking for the minimum in a smaller set, can that hurt?

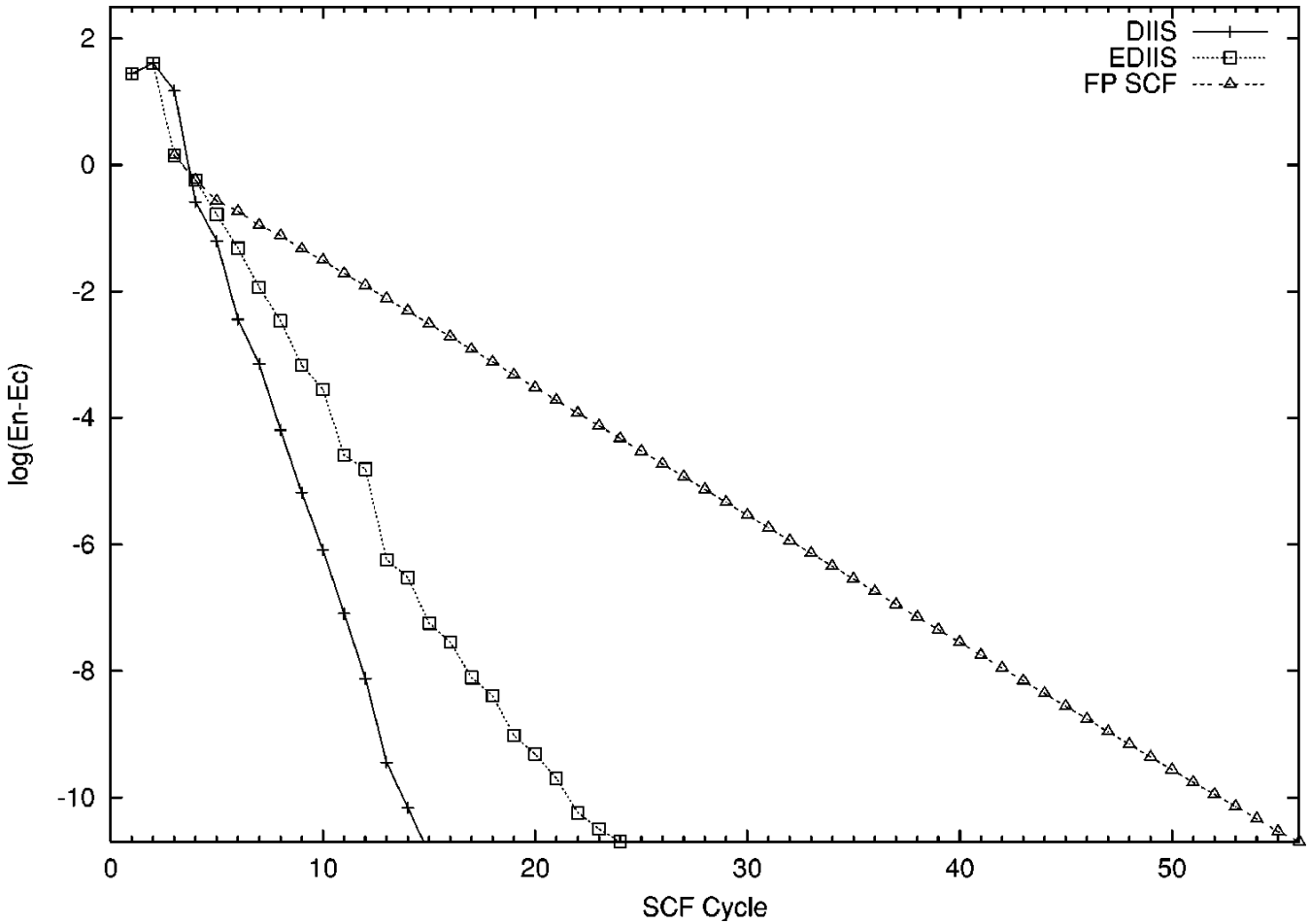
Convergence of EDIIS: Chen-K 2017

If $\mathbf{G}$ is a contraction in convex Ω then

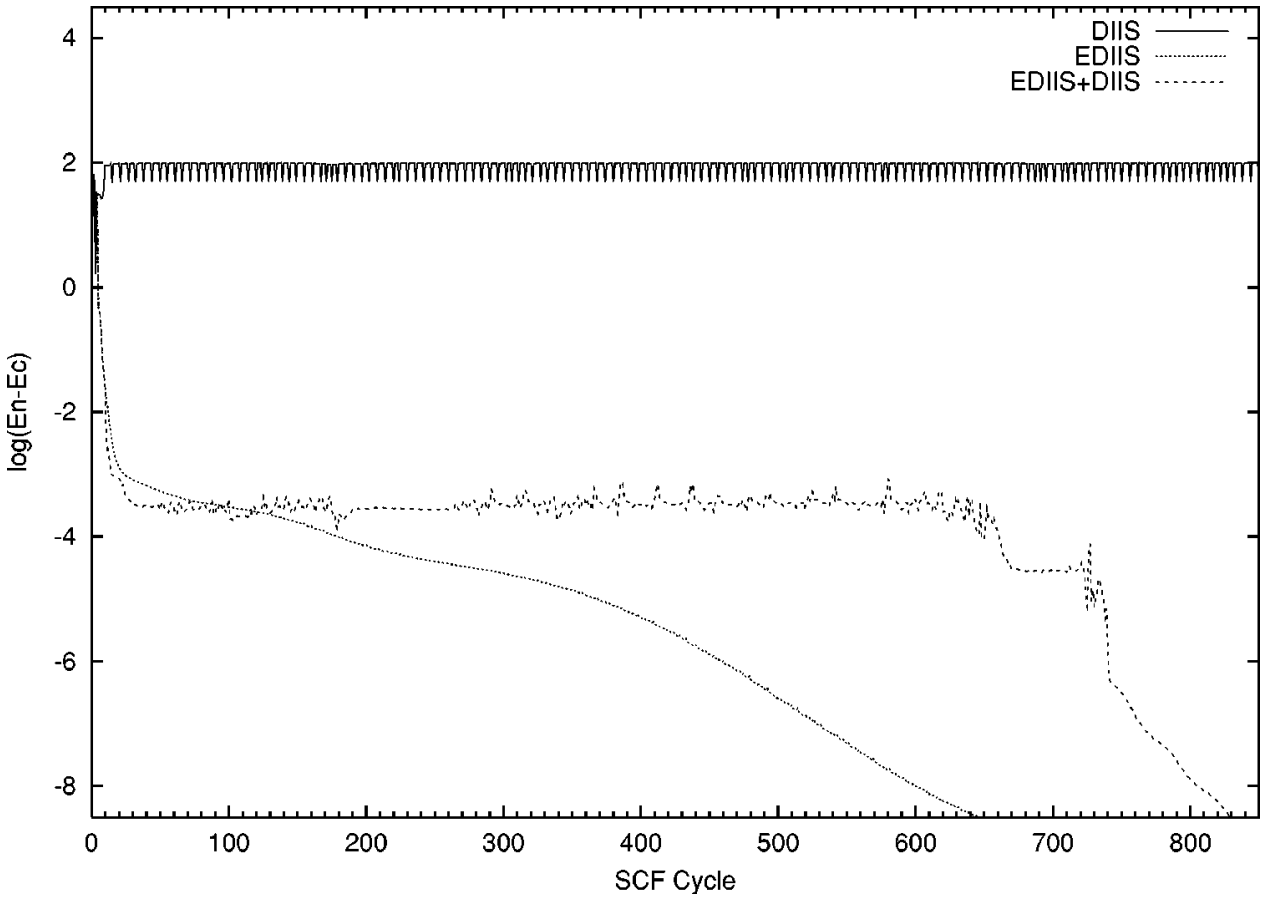
$$\|\mathbf{e}_k - \mathbf{u}^*\| \leq c^{k/(m+1)} \|\mathbf{e}_0 - \mathbf{u}^*\|$$

and the convergence is the same as the local theory when near $\mathbf{u}^*$.
Similar to global results for Newton's method.
Reflects practice reported by Kudin et al.

Easy problem from Kudin et al



Hard problem from Kudin et al



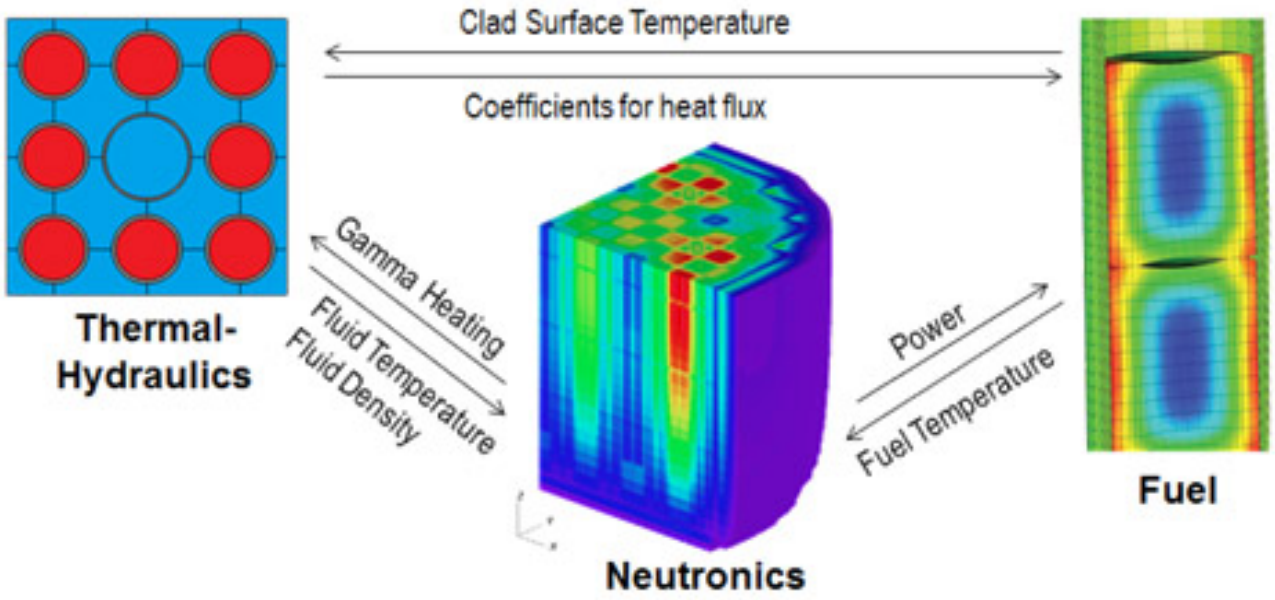
Multiphysics Coupling

Toth, Ellis, Clarno, Hamilton, K, Pawlowski, Slattery 2015-6
Objective: Iterate coupled simulations to consistency.
Problems:

- Black-box solvers
- Legacy codes
- Table lookups
- Internal stochastics

Jacobian information hard to get.

Reactor Physics



Fixed point map has Monte Carlo neutronics.

Results

- Theory and practice for Anderson.
  Extends work for Newton (Willert-K, 2013)
- Technical but reasonable assumptions.
- Asymptotic results as particle count increases.
  Given K, $\hat{c} \in (c, 1)$, and $\omega \in (0, 1)$ there is N_P such that if the number of particles is $\geq N_P$ then, if $\mathbf{e}_0$ is sufficiently small,

$$Prob(\|\mathbf{F}(\mathbf{u}_k)\| \leq \hat{c}^k \|\mathbf{F}(\mathbf{u}_0)\|) > 1 - \omega$$

  for all $0 \leq k \leq K$.

Example from Radiative Transfer

Chandrasekhar H-equation

$$H(\mu) = \mathbf{G}(H) \equiv \left(1 - \frac{\omega}{2}\int_0^1 \frac{\mu}{\mu+\nu} H(\nu)\, d\nu.\right)^{-1}$$

$\omega \in [0,1]$ is a physical parameter.
$\mathbf{F}'(H^*)$ is singular when $\omega = 1$.

$$\rho(\mathbf{G}'(H^*)) \leq 1 - \sqrt{1-\omega} < 1$$

Numerical Experiments

- Discretize with 500 point composite midpoint rule.
- Compare Newton-GMRES with Anderson(m).
- Terminate when $\|\mathbf{F}(H_k)\|_2/\|\mathbf{F}(H_0)\|_2 \leq 10^{-8}$
- $\omega = .5, .99, 1.0$
- $0 \leq m \leq 3$
- ℓ^1, ℓ^2, ℓ^∞ optimizations
- Tabulate
  - κ_{max}: max condition number of least squares problems
  - S_{max}: max absolute sum of coefficients

Newton-GMRES vs Anderson(0)

Function evaluations:

	Newton-GMRES			Fixed Point		
ω	.5	.99	1.0	.5	.99	1.0
Fs	12	18	49	11	75	23970

Anderson(m)

		ℓ^1 Optimization			ℓ^2 Optimization			ℓ^∞ Optimization		
ω	m	Fs	κ_{max}	S_{max}	Fs	κ_{max}	S_{max}	Fs	κ_{max}	S_{max}
0.50	1	7	1.00e+00	1.4	7	1.00e+00	1.4	7	1.00e+00	1.5
0.99	1	11	1.00e+00	3.5	11	1.00e+00	4.0	10	1.00e+00	10.1
1.00	1	21	1.00e+00	3.0	21	1.00e+00	3.0	19	1.00e+00	4.8
0.50	2	6	1.36e+03	1.4	6	2.90e+03	1.4	6	2.24e+04	1.4
0.99	2	10	1.19e+04	5.2	10	9.81e+03	5.4	10	4.34e+02	5.9
1.00	2	18	1.02e+05	43.0	16	2.90e+03	14.3	34	5.90e+05	70.0
0.50	3	6	7.86e+05	1.4	6	6.19e+05	1.4	6	5.91e+05	1.4
0.99	3	10	6.51e+05	5.2	10	2.17e+06	5.4	11	1.69e+06	5.9
1.00	3	22	1.10e+08	18.4	17	2.99e+06	23.4	51	9.55e+07	66.7

Observations

- For $m > 0$, Anderson(m) is much better than Picard
- Anderson(m) does better than Newton GMRES
- There is little benefit in larger m
- ℓ^∞ optimization seems to be a poor idea
- ℓ^1 optimization appears fine, but the cost is not worth it

How well does this REALLY work?

Our experiments and the rest of the world say . . .

- Night and day salvation in electonic structure computations, need a few hacks.
- Varies from a lot better than Picard to only a little better.
- Anderson theory is about residuals.
  Conditioning less important for theory, but maybe in practice.
- Stochastic functions ok.

Summary

- Anderson acceleration can improve Picard iteration
- Implementation does not require derivatives
  - Good when Newton is not possible
  - Convergence theory (and practice) for 1965 version.
  - EDIIS globalizes, but at a cost.
- Applications to electronic structure computations and multiphysics coupling
- In TRILINOS/SUNDIALS for your acceleration pleasure.

References

- D. G. Anderson,
Iterative Procedures for Nonlinear Integral Equations, Journal of the ACM, 12 (1965), pp. 547–560.
- P. Pulay,
Convergence acceleration of iterative sequences. The case of SCF iteration, Chemical Physics Letters, 73 (1980), pp. 393–398.
- K. N. Kudin, G. E. Scuseria, and E. Cancès,
A black-box self-consistent field convergence algorithm: One step closer, Journal of Chemical Physics, 116 (2002), pp. 8255–8261,

References

- A. Toth and C. T. Kelley, Convergence analysis for Anderson acceleration, SIAM J. Numer. Anal., 53 (2015), pp. 805 – 819.
- A. Toth, J. A. Ellis, T. Evans, S. Hamilton, C. T. Kelley, R. Pawlowski, and S. Slattery, Local improvement results for Anderson acceleration with inaccurate function evaluations, 2016. To appear in SISC.
- S. Hamilton, M. Berrill, K. Clarno, R. Pawlowski, A. Toth, C. T. Kelley, T. Evans, and B. Philip, An assessment of coupling algorithms for nuclear reactor core physics simulations, Journal of Computational Physics, 311 (2016), pp. 241–257.
- X. Chen, C. T. Kelley, and players to be named, Analysis and Implementation of EDIIS, in progress.