Anderson Acceleration

C. T. Kelley
NC State University
tim_kelley@ncsu.edu
Supported by ARO, NSF, DOE(CASL/LANL)

Poly U, May 16, 2014

Outline

Anderson Acceleration Algorithm

Solve fixed point problems

$$u = G(u)$$

faster than Picard iteration

$$u_{k+1} = G(u_k).$$

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

Why not Newton?

Newton's method

$$u_{k+1} = u_k - (I - G'(u_k))^{-1}(u_k - G(u_k))$$

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

Sometimes you will not have G'.

Electronic Structure Computations

Nonlinear eignevalue problem: Kohn-Sham equations

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

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.

SCF as a fixed-point iteration

SCF is a fixed point iteration

$$\rho \leftarrow G(\rho)$$

Not clear how to differentiate G

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

Multiphysics Coupling

Given several simulators: $\{S_j\}_{j=1}^{N_S}$

- The simulators depend on a partition $\{X_j\}_{j=1}^{N_S}$ of the primary variables
- S_i computes X_i as a function of $Z_i = \{X_j\}_{j \neq i}$
- The maps S_i could contain
  - Black-box solvers
  - Legacy codes
  - Table lookups
  - Internal stochastics
- Jacobian information very hard to get.

Iteration to self-consistency

Chose one X_i to **expose**. Then

- for $j = 1 : N_S, j \neq i$
  $X_j = S_j(Z_j)$
- $X_i \leftarrow S_i(Z_i)$

This is a fixed point problem

Basic Algorithm

anderson(u_0, G, m)

$u_1 = G(u_0)$; $F_0 = G(u_0) - u_0$

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

$m_k = \min(m, k)$

$F_k = G(u_k) - u_k$

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

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

end for

Terminology

- m, depth. We refer to Anderson(m). Anderson(0) is Picard.
- $F(u) = G(u) - u$, residual
- $\{\alpha_j^k\}$, coefficients
- $\{\beta_k\}$, mixing parameters
- $\|\cdot\|$, ℓ^2, ℓ^1, or ℓ^∞

We set $\beta_k = 1$ in this talk.

Optimization Problem for Coefficients: Version I

This version is useful for analysis. Solve the unconstrained problem

$$\min \|F(u_k) - \sum_{j=1}^{m_k} \alpha_j^k (F(u_{k-m_k+j}) - F(u_k))\|,$$

for $\{\alpha_j^k\}_{j=1}^k$. Then α_0^k by

$$\alpha_0^k = 1 - \sum_{j=1}^{m_k} \alpha_j^k.$$

Not optimal for implementation (more later).

Linear Problems

Here

$$G(u) = Mu + b$$

where M is $N \times N$ and $\|M\| < 1$.
Theorem: (Toth-K, 2013) The residuals for Anderson(m) converges to 0 q-linearly with q-factor c.

$$\|F(u_{k+1})\| \leq c\|F(u_k)\|.$$

No worse that Picard iteration.

Proof: residual convergence

Claim: $\|F(u_{k+1})\| = \|b - (I - M)u_{k+1}\| \leq c\|F(u_k)\|$
proof: Since $\sum \alpha_j = 1$, the new residual is

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

Proof continued

So, by the optimality condition

$$\begin{aligned} \|F(u_{k+1})\| &\leq \|M\| \|\sum_{j=0}^{m_k} \alpha_j F(u_{k-m_k+j})\| \\ &\leq \|M\| \|F(u_k)\| \leq c\|F(u_k)\| \end{aligned}$$

r-linear convergence of $\{u_k\}$

If we set $e = u - u^*$, then $F(u) = -(I - M)e$. So q-linear convergence of residuals implies that

$$(1 - c)\|e_k\| \leq \|F(u_k)\| \leq c^k \|F(u_0)\| \leq c^k(1 + c)\|e_0\|$$

and hence

$$\|e_k\| \leq \left(\frac{1 + c}{1 - c}\right) c^k \|e_0\|.$$

which is r-linear convergence with r-factor c.

Connection to GMRES

Recall that the k GMRES iteration is the solution of the linear least squares problem

$$\min_{u \in u_0 + \mathcal{K}_k} \|b - (I - M)u\|_2$$

where the Krylov subspace is

$$\mathcal{K}_k = \text{span}(r_0, (I - M)r_0, \ldots, (I - M)^{k-1} r_0)$$

Anderson iteration with the ℓ^2 norm is also an residual minimization.
Connection: (Walker-Ni, 2011)
If $I - M$ is nonsingular and GMRES residuals are strictly decreasing in norm, then

$$u_{k+1} = G(u_k^{GMRES}).$$

- Assume G is a contraction, constant c.
  Objective: do no worse than Picard
- Local theory only; $\|e_0\|$ is small.
- Better results for $\|\cdot\|_2$.

Assumptions: $m = 1$

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

Words: 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

Let $c < \hat{c} < 1$, then Anderson(1) converges and

$$\limsup_{k\to\infty} \frac{\|F(u_{k+1})\|_2}{\|F(u_k)\|_2} \leq c.$$

Assumptions: $m > 1$, any norm

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

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

Convergence for Anderson(m), any norm

Assumptions and $c < \hat{c} < 1$
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

$$\|F(u_k)\| \leq \hat{c}^k \|F(u_0)\| \tag{1}$$

and

$$\|e_k\| \leq \frac{(1+c)}{1-c} \hat{c}^k \|e_0\|. \tag{2}$$

Implementation

At iteration k solve

$$\min_{\theta \in R^{m_k}} \|F(u_k) - \sum_{j=0}^{m_k-1} \theta_j (F(u_{j+1}) - F(u_j))\|$$

for $\theta \in R^{m_k}$. Then

$$u_{k+1} = G(u_k) - \sum_{j=0}^{m_k-1} \theta_j^k (G(u_{j+1}) - G(u_j)).$$

In terms of the original iteration

$\alpha_0 = \theta_0, \alpha_j = \theta_j - \theta_{j-1}$ for $1 \le j \le m_k - 1$ and $\alpha_{m_k} = 1 - \theta_{m_k-1}$.

Why is this better?

- Coefficient matrices only change by a column per iteration.
- Can update QR factorization "fast" to compute θ.
- Somewhat better conditioning.

Only the first point matters since m is typically very small.
We do it this way in the example.

Example from Radiative Transfer

Chandrasekhar H-equation

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

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

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

Numerical Experiments

- Discretize with 500 point composite midpoint rule.
- Compare Newton-GMRES with Anderson(m).
- Terminate when $\|F(u_k)\|_2/\|F(u_0)\|_2 \leq 10^{-8}$
- $\omega = .5, .99, 1.0$
- $0 \leq m \leq 6$
- ℓ^1, ℓ^2, ℓ^∞ optimizations

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

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	4	7	2.64e+09	1.5	6	9.63e+08	1.4	6	9.61e+08	1.4
0.99	4	11	1.85e+09	5.2	11	6.39e+08	5.4	11	1.61e+09	5.9
1.00	4	23	2.32e+08	12.7	21	6.25e+08	6.6	35	1.38e+09	49.0
0.50	5	7	1.80e+13	1.4	6	2.46e+10	1.4	6	2.48e+10	1.4
0.99	5	11	3.07e+10	5.2	12	1.64e+11	5.4	13	3.27e+11	5.9
1.00	5	21	2.56e+09	21.8	27	1.06e+10	14.8	32	4.30e+09	190.8
0.50	6	7	2.65e+14	1.4	6	2.46e+10	1.4	6	2.48e+10	1.4
0.99	6	12	4.63e+11	5.2	12	1.49e+12	5.4	12	2.27e+11	5.9
1.00	6	31	2.61e+10	45.8	35	1.44e+11	180.5	29	3.51e+10	225.7

Observations

- For $m > 0$, Anderson(m) is much better than Picard
- Anderson(m) does better than Newton GMRES
- There is little benefit in $m \geq 3$
- ℓ^∞ optimization seems to be a poor idea
- ℓ^1 optimization appears fine, but the cost is not worth it

Summary

- Anderson acceleration can improve Picard iteration
- Implementation does not require derivatives
  - Good when Newton is not possible
  - Convergence theory (and practice) is local
- Applications to electronic structure computations and multiphysics coupling