Chebychev acceleration of iterative refinement

Jennifer Scott
STFC Rutherford Appleton Laboratory
in collaboration with Mario Arioli

SIAM Conference on Applied Linear Algebra, 21 June 2012

Motivation

- Gaussian elimination with partial pivoting followed by few steps of iterative refinement (IR) can compute solution of sparse linear system that is backward stable.
- But if weaker stability test used (threshold partial pivoting or static pivoting) large number of steps may be needed.
- Or if mixed precision used, number of steps may be very large.
- The solve phase of a direct solver can be bottleneck on multicore machine.
- Thus want to accelerate the refinement (reduce number of solves).

Let $\mathbf{A} \in \mathrm{R}^{n \times n}$ and $b \in \mathrm{R}^n$, with $\mathrm{rank}\,(\mathbf{A}) = n$. Then

$$\mathbf{Ax} = \mathbf{b}$$

has a unique solution $\hat{\mathbf{x}}$.
Assume Gaussian elimination is performed using floating-point arithmetic with relative precision ϵ.

The computed factors satisfy

$$\mathbf{A} + \mathbf{F} = \widehat{\mathbf{L}}\widehat{\mathbf{U}} = \mathbf{M},$$

where

$$|\mathbf{F}| \leq c(n)\epsilon|\widehat{\mathbf{L}}|\,|\widehat{\mathbf{U}}|.$$

It follows that

$$\mathbf{x} = \mathbf{M}^{-1}\big(\mathbf{b} + \mathbf{F}\mathbf{x}\big) = \mathbf{M}^{-1}\big(\mathbf{b} - \mathbf{A}\mathbf{x} + \mathbf{M}\mathbf{x}\big) = \mathbf{M}^{-1}\big(\mathbf{r}(\mathbf{x}) + \mathbf{M}\mathbf{x}\big),$$

where $\mathbf{r}(\mathbf{x})$ is the residual $\mathbf{b} - \mathbf{A}\mathbf{x}$.

Thus, $\mathbf{x}$ is the fixed point of $\mathfrak{F}(\mathbf{x})$, where

$$\mathfrak{F}(\mathbf{x}) = \mathbf{x} + \mathbf{M}^{-1}\mathbf{r}(\mathbf{x}).$$

Iterative refinement algorithm

Let $\mathbf{x}^{(0)} = \mathbf{M}^{-1}\mathbf{b}$ and $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$.
Given a convergence tolerance $\eta > 0$.

Initialise $k = 0$.
while $\beta^{(k)} > \eta$ **do**

$\delta\mathbf{x} = \mathbf{M}^{-1}\mathbf{r}^{(k)}$;

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}$;
$\mathbf{r}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)}$;
$\beta^{(k+1)} = \max_i |\mathbf{r}_i^{(k+1)}|/(|\mathbf{A}|\,|\mathbf{x}^{(k+1)}| + |\mathbf{b}|)_i$;
$k = k + 1$.

end while

Convergence of iterative refinement

If the spectral radius of $\mathbf{M}^{-1}\mathbf{F}$ satisfies

$$\sigma(\mathbf{M}^{-1}\mathbf{F}) < 1$$

in exact arithmetic, the sequence $\mathbf{x}^{(k)}$ converges to $\hat{\mathbf{x}}$.
Furthermore,

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \mathbf{A}\mathbf{M}^{-1}\mathbf{r}^{(k)} = \mathbf{F}\mathbf{M}^{-1}\mathbf{r}^{(k)}.$$

Therefore, if $\sigma(\mathbf{F}\mathbf{M}^{-1}) < 1$, the residuals converge to zero.

Convergence example

Assume $\sigma(\mathbf{M}^{-1}\mathbf{F}) = 0.5$.

To achieve reduction of three orders of magnitude in initial residual, number of iterations needed is

$$iter = \lceil \frac{log_{10}(10^{-3})}{log_{10}(0.5)} \rceil,$$

which is approximatively 10.

Cost may be unacceptably high (eg. factors held out-of-core).

Our objectives

We seek alternatives to IR that require fewer iterations and preserve two key properties:

- component-wise stability, and
- the absence of scalar products.

Why no scalar products?

They are communication intensive and hence not suited to parallel computers with high communication costs.

In this talk, we explore possibility of using Chebychev acceleration of IR.

Chebychev polynomials

Recall Chebyshev polynomials can be defined by the recurrence

$$\begin{cases} T_0(z) = 1 \\ \\ T_1(z) = z \\ \\ T_{k+1}(z) = 2zT_k(z) - T_{k-1}(z) \qquad k \geq 1. \end{cases}$$

Theorem

Let $\mathcal{D}$ be the region enclosed by the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1,$$

where $b < a < 1$. Let $\mathcal{S}_j$ be set of all real polynomials $p_j(z)$ of degree at most j such that $p_j(1) = 1$, then

$$\wp_j(z) = \frac{T_j(z/c)}{T_j(1/c)}, \qquad \text{where } \; c^2 = a^2 - b^2,$$

is the unique polynomial in $\mathcal{S}_j$ such that

$$\max_{z\in\mathcal{D}} |\wp_j(z)| \leq \max_{z\in\mathcal{D}} |p_j(z)|, \qquad p_j(z) \in \mathcal{S}_j.$$

Chebychev acceleration

Recall:

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{F}\mathbf{x} + \mathbf{M}^{-1}\mathbf{b} \qquad (\mathbf{A} + \mathbf{F} = \mathbf{M}).$$

Assume the eigenvalues of $\mathbf{M}^{-1}\mathbf{F}$ lie in the interior of an ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1,$$

with $b < a < 1$ and set $c^2 = a^2 - b^2$.

Chebychev acceleration

Can show Chebychev relations for this are:

$$\begin{cases} \mathbf{x}^{(0)} = \mathbf{M}^{-1}\mathbf{b}, \quad \varrho_1 = 1 \\ \\ \mathbf{x}^{(j+1)} = \varrho_{j+1}\left(\mathbf{M}^{-1}\mathbf{F}\mathbf{x}^{(j)} + \mathbf{M}^{-1}\mathbf{b}\right) + (1 - \varrho_{j+1})\mathbf{x}^{(j-1)}, \quad j = 0, \ldots, \end{cases}$$

where

$$\varrho_{j+1} = \frac{2}{c}\frac{T_j(1/c)}{T_{j+1}(1/c)},$$

and $T_j(z)$ is the Chebychev polynomial of degree j.

Chebychev acceleration of iterative refinement

Simple algebraic manipulation gives:

Let $\mathbf{x}^{(0)} = \mathbf{M}^{-1}\mathbf{b}$, $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$.
Initialise $k = 0$.

while $\beta^{(k)} > \eta$ **do**

$$\delta\mathbf{x}^{(k+1)} = \varrho_{k+1}\mathbf{M}^{-1}\mathbf{r}^{(k)} - (1 - \varrho_{k+1})\delta\mathbf{x}^{(k)};$$

$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta\mathbf{x}^{(k+1)}$;
$\mathbf{r}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)}$;
$\beta^{(k+1)} = \max_i |\mathbf{r}_i^{(k+1)}|/(|\mathbf{A}|\,|\mathbf{x}^{(k+1)}| + |\mathbf{b}|)_i$;
$k = k + 1$.

end while

Chebychev acceleration

Observe computing the ϱ_j is straightforward:

$$\varrho_{j+1} = \begin{cases} 1, & \text{if } j = 0 \\ (1 - \frac{1}{2}c^2)^{-1}, & \text{if } j = 1 \\ (1 - \frac{1}{4}c^2\varrho_j)^{-1}, & \text{if } j \geq 2 \end{cases}$$

Thus computational cost of Chebychev at each iteration is only marginally more than for IR.

Convergence example again

Recall if $\sigma(\mathbf{M}^{-1}\mathbf{F}) = 0.5$, approx. 10 steps of IR to reduce residual by 3 orders of magnitude.

Number of Chebychev iterations to reduce residual by p orders of magnitude is

$$iter = \lceil \frac{log_{10}(10^{-p})}{log_{10}(\frac{a+b}{1+\sqrt{1-c^2}})} \rceil.$$

If $\sigma(\mathbf{M}^{-1}\mathbf{F})$ contained in ellipse

$$\left(\frac{x}{0.5}\right)^2 + \left(\frac{y}{0.05}\right)^2 = 1$$

approx. 6 iterations needed for same reduction.

Choosing the ellipse

For successful convergence must forecast an ellipse that envelops $\sigma(\mathbf{M}^{-1}\mathbf{F})$.

If $\sigma(\mathbf{M}^{-1}\mathbf{F})$ lies outside the chosen ellipse and $c^2 \ll 1$ then

$$\varrho_\infty = \lim_{j\to\infty} \varrho_j = 2/(1+\sqrt{1-c^2}) \approx 1$$

Thus asymptotic behaviour same as for IR and no acceleration.

Also, if ellipse degenerates to a circle ($a = b$), $\varrho_j = 1\ \forall j$ and no acceleration.

Choosing the ellipse

- Want to limit use of scalar products and thus we do not use adaptive method of Manteuffel ('78).
- Instead, introduce some strong assumptions on parameters defining the ellipse.
- Because eigenvalues of $\mathbf{M}^{-1}\mathbf{F}$ are either real or complex conjugate, centre of ellipse lies on real axis. We choose centre of ellipse to be zero.

Choosing the ellipse

If spectral radius $\sigma(\mathbf{M}^{-1}\mathbf{F})$ lies between $(0,1)$, can scale so that

$$\left.\begin{array}{rcl} a & = & \sigma(\mathbf{M}^{-1}\mathbf{F}) \\ b & = & t * a \end{array}\right\},$$

with t chosen such that spectrum is contained within the ellipse.

This choice is based on our empirical experience: Gaussian elimination frequently produces $\mathbf{M}^{-1}\mathbf{F}$ with a spectrum characterised by large cluster of very small eigenvalues and for remaining eigenvalues $\lambda + i\eta$, $|\lambda| \gg |\eta|$.

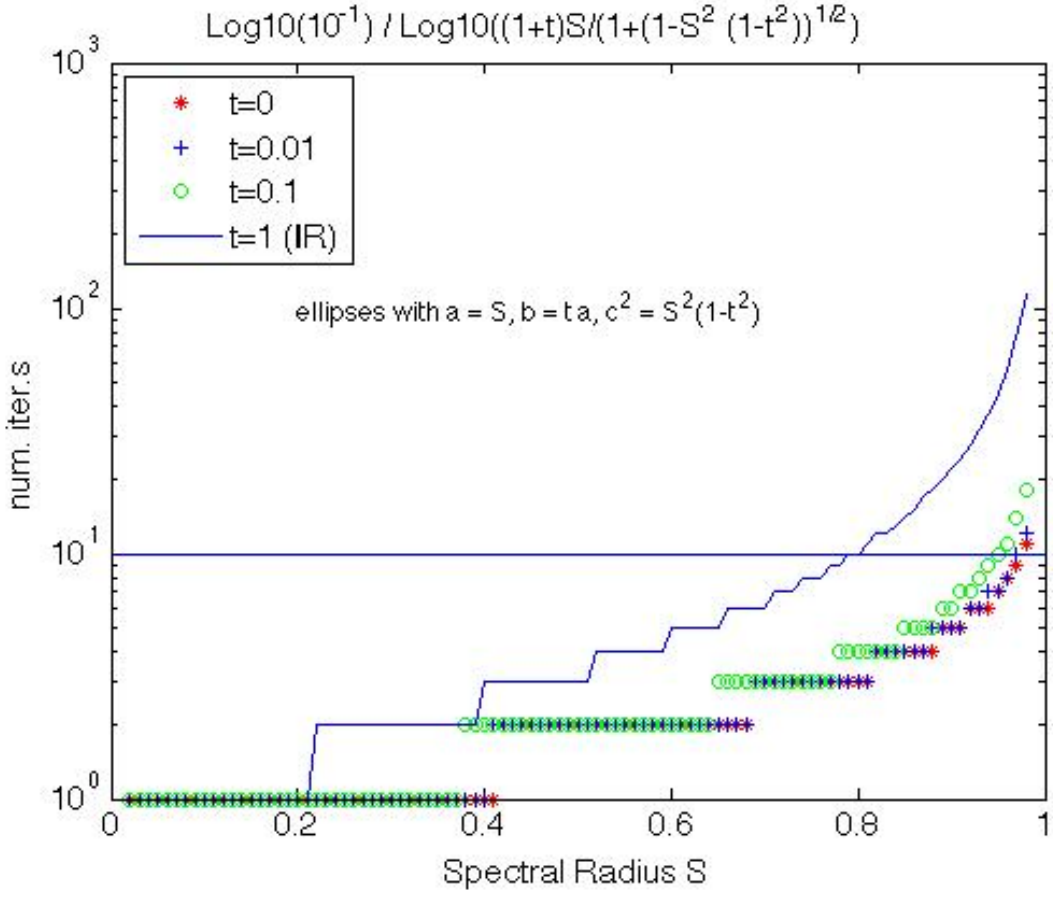
Choosing the ellipse

But this is not always the case.

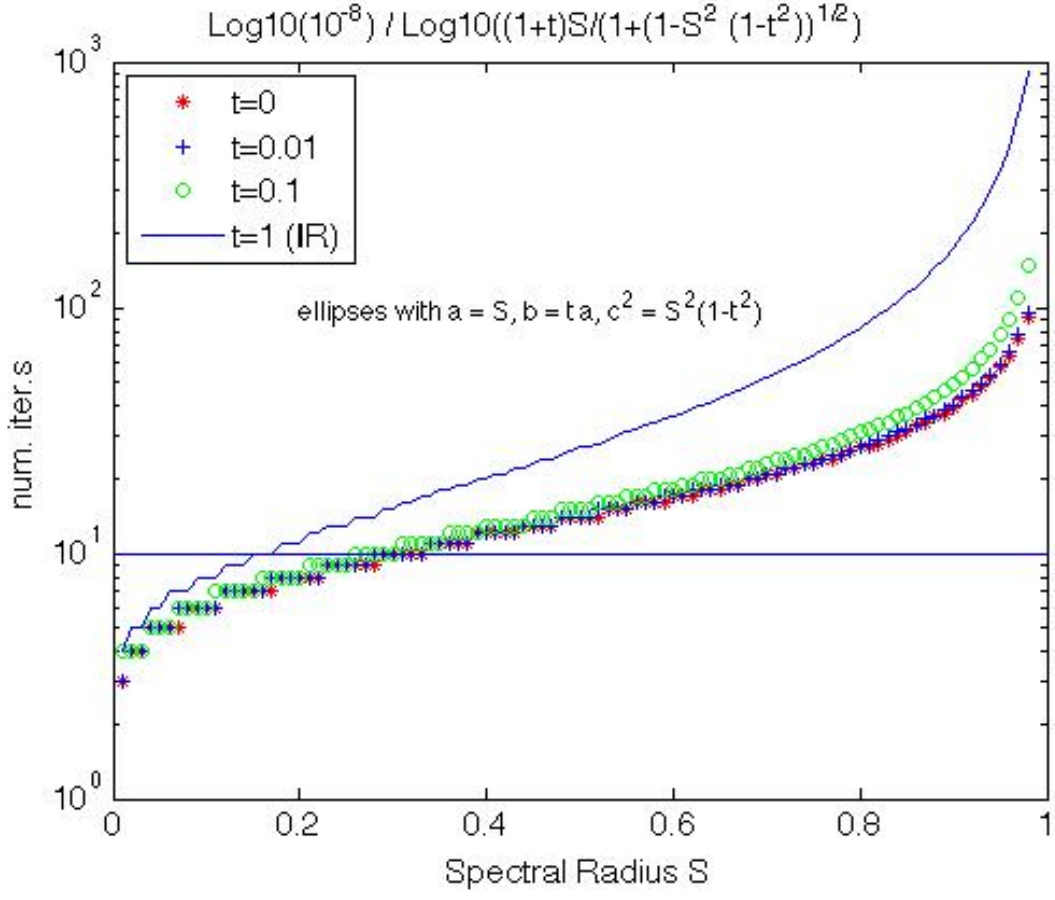
If $|\eta| \gg |\lambda|$, Chebychev algorithm using our choice of ellipse will diverge immediately

So exchange a and b (rotating the ellipse by $\pi/2$) and restart.

Asymptotic rate of convergence for reducing residual by 10^{-1}



Asymptotic rate of convergence for reducing residual by 10^{-8}



Conclude:

If σ sufficiently large, potential saving offered by Chebychev IR is significant.

Estimating σ

In our experiments, we use $a = \sigma(\mathbf{M}^{-1}\mathbf{F})$, $b = t * a$.

So want to estimate $a = \sigma(\mathbf{M}^{-1}\mathbf{F})$.

After one step of Chebychev IR (which is equal to first step of IR), can estimate σ to be

$$\sigma(\mathbf{M}^{-1}\mathbf{F}) \approx \rho_1 = \frac{||\bar{\mathbf{r}}^{(1)}||}{||\bar{\mathbf{r}}^{(0)}||}.$$

More generally, use ratio between successive computed residuals

$$\rho_k = \frac{||\bar{\mathbf{r}}^{(k)}||}{||\bar{\mathbf{r}}^{(k-1)}||}$$

Tests on sparse symmetric matrices

- Factorize symmetric **A** using single precision version of new sparse multifrontal solver `HSL_MA97`, store computed factors in double precision and then perform refinement using double precision arithmetic. Mixed precision.
- Examples from University of Florida Sparse Matrix Collection chosen that require ≥ 10 steps of IR to achieve component-wise scaled residual $\beta < 5 * 10^{-15}$.
- In Chebychev tests, $b = 0.01 * a$ and experimented with range of values of a.

Chebychev IR versus IR

Problem	*IR*	Chebyshev *IR*				
		$a = \rho_1$		$a = a_{best}$		
	iter	*iter*	ρ_1	*iter*	a_{best}	(k)
HB/nos7	28	15	0.53	13	0.54	
HB/bcsstm27	30	15	0.56	14	0.57	
GHS_indef/bratu3d	25	20	0.34	17	0.25	(2)
Cylshell/s3rmq4m1	14	10	0.25	10	0.25	(1)
GHS_indef/cont-300	218	213	0.08	65	0.87	(9)
Oberwolfach/gyro	25	19	0.28	13	0.46	(3)
GHS_indef/sparsine	38	21	0.51	20	0.50	(2)

Highlighted where iteration count reduced by at least 40 per cent.

Can Chebyshev acceleration improve rate of convergence when $\rho_k \approx 1$?

Problem	$\beta^{(0)}$	$\beta_{IR}^{(200)}$
Boeing/crystk03	$1.42 * 10^{-7}$	$7.03 * 10^{-10}$
Oberwolfach/t2dal	$1.13 * 10^{-7}$	$5.64 * 10^{-10}$
Oberwolfach/t3dh_a	$3.10 * 10^{-7}$	$1.54 * 10^{-9}$

Problem	$\beta_C^{(200)}(a)$		
	$a = 0.99$	0.9999	0.999999
Boeing/crystk03	$9.98 * 10^{-11}$	$1.22 * 10^{-11}$	$7.83 * 10^{-12}$
Oberwolfach/t2dal	$8.01 * 10^{-11}$	$9.74 * 10^{-12}$	$3.18 * 10^{-12}$
Oberwolfach/t3dh_a	$2.18 * 10^{-10}$	$2.63 * 10^{-11}$	$1.55 * 10^{-11}$

Note: with $a \geq 0.9999$, after 15 steps $\beta_C^{(k)}(a)$ is less than $\beta_{IR}^{(200)}$.

What about FGMRES?

- Well-known that for some problems, FGMRES able to compute backward stable solutions when IR fails to converge.
- **But** FGMRES does not preserve either of the desirable properties:
- FGMRES gives rapid convergence but only possible to prove that norm-wise stability, and implementation involves scalar products.

Comparison of the number of steps

Problem	*IR*	Chebyshev *IR*	FGMRES
HB/nos7	28	13	8
HB/bcsstm27	30	14	12
GHS_indef/bratu3d	25	17	12
Cylshell/s3rmq4m1	14	10	8
GHS_indef/cont-300	218	65	28
Oberwolfach/gyro	25	13	12
GHS_indef/sparsine	38	20	12

An efficient restarted version of FGMRES used here that we developed for the mixed precision solver `HSL_MA79`.

Concluding remarks

- Aim of our study was to consider possibility of using Chebychev accelerated IR when IR needs large number of iterations.
- We have developed theory (not presented here) to prove method is component-wise backward stable.
- Experiments on sparse symmetric problems illustrate that using inexpensive estimate of spectral radius obtained by performing small number of steps of IR gives good convergence.
- FGMRES generally requires fewer iterations, but involves scalar products that are inefficient when implemented in parallel.
- When $\sigma(\mathbf{M}^{-1}\mathbf{F}) \approx 1$ or is > 1, FGMRES recommended.

Thank you!

Report available: RAL-TR-2011-10 at

`http://www.stfc.ac.uk/CSE/randd/nag/36276.aspx`