University of Maryland College Park

Institute for Advanced Computer Studies TR–2007–09
Department of Computer Science TR–4854

A Residual Inverse Power Method*

G. W. Stewart†

February 2007

ABSTRACT

The inverse power method involves solving shifted equations of the form $(A-\sigma I)v = u$. This paper describes a variant method in which shifted equations may be solved to a fixed reduced accuracy without affecting convergence. The idea is to alter the right-hand side to produce a correction step to be added to the current approximations. The digits of this step divide into two parts: leading digits that correct the solution and trailing garbage. Hence the step can be be evaluated to a reduced accuracy corresponding to the correcting digits. The cost is an additional multiplication by A at each step to generate the right-hand side. Analysis and experiments show that the method is suitable for normal and mildly nonnormal problems.

*This report is available by at `ftp://ftp.cs.umd.edu/pub/stewart/reports/Contents.html`.

†Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (`stewart@cs.umd.edu`). This work was supported in part by the National Science Foundation under grant CCR0204084.

A Residual Inverse Power Method

G. W. Stewart

ABSTRACT

The inverse power method involves solving shifted equations of the form $(A-\sigma I)v = u$. This paper describes a variant method in which shifted equations may be solved to a fixed reduced accuracy without affecting convergence. The idea is to alter the right-hand side to produce a correction step to be added to the current approximations. The digits of this step divide into two parts: leading digits that correct the solution and trailing garbage. Hence the step can be be evaluated to a reduced accuracy corresponding to the correcting digits. The cost is an additional multiplication by A at each step to generate the right-hand side. Analysis and experiments show that the method is suitable for normal and mildly nonnormal problems.

In many applications it is necessary to solve linear systems of the form

$$(A - \sigma I)v = u, \tag{1}$$

where A is of order n. If n is so large that direct methods are impractical, one must use iterative methods such as GMRES to solve the system. Unfortunately, σ often lies within the spectrum of A, and experience shows that in such cases iterative methods converge slowly, if at all. Moreover, good preconditioners are hard to find. For this reason, recent efforts have been directed toward getting by with solutions (1) in which v is compute to restricted accuracy. Of course, the definition of "getting by" will depend on the application. For a useful survey see [1].

This note is concerned with the inverse power method. To motivate the method, assume that A has a complete set of eigenpairs (λ_i, x_i) $(i = 1, \ldots, n)$. Let u (which we assume to have 2-norm one) be expanded in the form

$$u = \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_n x_n.$$

Then the solution v of (1) is

$$v = \frac{\gamma_1}{\lambda_1 - \sigma} x_1 + \frac{\gamma_2}{\lambda_2 - \sigma} x_2 + \cdots + \frac{\gamma_n}{\lambda_n - \sigma} x_2. \tag{2}$$

Now suppose that the σ is near λ_1, say $|\lambda_1 - \sigma| = 10^{-3}$, while $|\gamma_i - \sigma| = O(1)$ $(i = 2, \ldots, n)$. Then in passing from u to v, the component along x_1 is enhanced by a factor of 1,000, while the sizes of the components along the other vectors remain essentially unchanged. When this process is iterated, the resulting sequence of vectors, suitably

normalized, converge to x_1 — each successive vector adding three more significant digits to the approximation of x_1.

Since the method is self-correcting, it is only necessary to solve (1) to as many digits as the approximation can be expected to be accurate — in the example of the last paragraph to three additional significant digits with each iteration. However, it would be better if we could solve the equation to a constant number of significant digits and still converge to a fully accurate approximation. The purpose of this note is to describe and analyze a method that realizes this desideratum, and least for normal and mildly nonnormal matrices.

The key idea is to not compute v itself but to compute a correction to u that produces v. Specifically, let

$$s = \alpha v - u, \tag{3}$$

where α is a scalar to be defined later. Then the next approximation will be

$$\hat{u} = \frac{u + s}{\|u + s\|}. \tag{4}$$

The rationale is that, provided α is suitably chosen, we only have to compute s to the accuracy necessary to make an effective correction — three digits in the example above.

To derive an equation for s, substitute (1) into (3) to get

$$s = \alpha(A - \sigma I)^{-1}u - u.$$

Multiplying by $A - \sigma I$, we get

$$(A - \sigma I)s = \alpha u - (A - \sigma I)u = (\sigma + \alpha)u - Au. \tag{5}$$

Thus if we know α, we can solve (5) for s.

The normalizing factor should be chosen so that u and αv are as near as possible. For suppose that αv is, say, three orders of magnitude greater than u. Since $u = \alpha v - s$ has norm one, roughly three of the leading digits of s will have to be identical to the corresponding digits of αv. This means that s must be computed to three more significant digits than if u and v were approximately the same size.

The ideal value for α would minimize $\|u - \alpha v\|$. This is a least squares problem in the single variable α, and its solution is

$$\hat{\alpha} = \frac{v^*u}{v^*v} = \frac{u^*(A - \sigma I)^{-*}u}{u^*(A - \sigma I)^{-*}(A - \sigma I)^{-1}u}.$$

Now u is an approximate eigenvector of A corresponding to λ. Hence

$$\hat{\alpha} \cong \lambda_1 - \sigma.$$

Since we don't know λ_1, we will approximate it by its Rayleigh quotient $\tilde{\lambda}_1 = u^*Au$. Hence we take $\alpha = \tilde{\lambda}_1 - \sigma$, which along with (5) gives

$$(A - \sigma I)s = \tilde{\lambda}_1 u - Au. \tag{6}$$

Equations (6) and (4) define our new algorithm. The vector $\tilde{\lambda}_1 u - Au$ is a residual — in fact, given u it is the smallest possible residual. We will therefore call the algorithm the residual inverse power method.

We turn now to an asymptotic analysis of this method. For brevity, we will drop the subscript one from the eigenpair (λ_1, x_1). By an orthogonal change of coordinates we may assume that

$$A = \begin{pmatrix} \lambda & h^* \\ 0 & B \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Let

$$u = \begin{pmatrix} 1 \\ e \end{pmatrix}.$$

The vector u has norm one up to terms of order $\|e\|^2$; and if e is sufficiently small, the normalization has negligible effect. In what follows we will ignore such second order terms.

The Rayleigh quotient is

$$\hat{\lambda} = (1 \;\; e^*) \begin{pmatrix} \lambda & h^* \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 \\ e \end{pmatrix} = \lambda + h^*e + O(\|e\|^2).$$

Hence the right hand side of (6) is

$$(\lambda + h^*e) \begin{pmatrix} 1 \\ e \end{pmatrix} - \begin{pmatrix} \lambda & h^* \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ -(B - \lambda I)e \end{pmatrix} + O(\|e\|^2),$$

and $s^* = (s_1^* \;\; s_2^*)$ is the solution of

$$\begin{pmatrix} \lambda - \sigma & h^* \\ 0 & B - \sigma I \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -(B - \lambda I)e \end{pmatrix} + O(\|e\|^2).$$

Thus

$$\begin{aligned} s_2 &= -(B - \sigma I)^{-1}(B - \lambda I)e + O(\|e\|^2) \\ &= -(B - \sigma I)^{-1}[(B - \sigma I) - (\lambda - \sigma)I]e + O(\|e\|^2) \\ &= -e + (\lambda - \sigma)(B - \sigma I)^{-1}e + O(\|e\|^2), \end{aligned} \tag{7}$$

and

$$s_1 = -(\lambda - \sigma)^{-1}h^*s_2 + O(\|e\|^2) = [(\lambda - \sigma)^{-1}h^* - h^*(B - \sigma I)^{-1}]e + O(\|e\|^2). \tag{8}$$

Hence

$$u + s = \begin{pmatrix} 1 + [(\lambda - \sigma)^{-1} h^* - h^*(B - \sigma I)^{-1}]e \\ (\lambda - \sigma)(B - \sigma I)^{-1} e \end{pmatrix} + O(\|e\|^2),$$

and on normalizing we get

$$\hat{u} = \begin{pmatrix} 1 \\ (\lambda - \sigma)(B - \sigma I)^{-1} e \end{pmatrix} + O(\|e\|^2) \equiv \begin{pmatrix} 1 \\ \hat{e} \end{pmatrix} + O(\|e\|^2),$$

where

$$\hat{e} = (\lambda - \sigma)(B - \sigma I)^{-1} e. \tag{9}$$

To see when the method can be expected to converge, take norms in (9) to get.

$$\|\hat{e}\| \leq |\lambda - \sigma| \cdot \|(B - \sigma I)^{-1}\| \|e\|. \tag{10}$$

The quantity $\|(B - \sigma I)^{-1}\|^{-1}$ is often written as $\operatorname{sep}(\sigma, B)$. It satisfies the inequality

$$\operatorname{sep}(\sigma, B) \leq \max\{\mu \in \text{spectrum}(B) \colon |\sigma - \mu|\},$$

so that in some sense it measures the separation of σ from the spectrum of B. Thus we can rewrite (10) as

$$\|\hat{e}\| \leq \frac{|\lambda - \sigma|}{\operatorname{sep}(\sigma, B)} \|e\|. \tag{11}$$

Thus if we compute s with no errors, our algorithm will work if

$$\rho = \frac{|\lambda - \sigma|}{\operatorname{sep}(\sigma, B)} < 1; \tag{12}$$

and the smaller the ratio, the faster the convergence. Since sep is a continuous function of σ and B, we can make the ρ as small as we like by taking σ sufficiently near λ.[1]

We now turn to the question of how accurately we must compute s to attain the reduction promised in (11). In the following discussion we will drop the $O(\|e\|)$ terms.

Let g be the error we introduce when we solve equation (6) and let

$$\gamma = \frac{\|g\|}{\|s\|} \tag{13}$$

be the normwise relative error. Since $\|g\|$ must be smaller that $\|\hat{e}\|$ for full reduction, we must have.

$$\gamma \leq \frac{\|\hat{e}\|}{\|s\|}.$$

[1] An analysis of the method based on (2) gives a convergence ratio of $|\lambda_1 - \sigma|/|\lambda_2 - \sigma|$, where λ_2 is the eigenvalue nearest σ after λ_1. The replacement of the denominator with $\operatorname{sep}(\sigma, B)$ is the price we pay for not having to assume that A has a complete system of eigenvectors. See [2, §2.1].

Now

$$\|\hat{e}\| = |\lambda - \sigma| \cdot \|(B - \sigma I)^{-1} e\| = \tau\rho\|e\|.$$

where

$$\tau = \frac{\|(B - \sigma I)^{-1} e\|}{\|(B - \sigma I)^{-1}\|\|e\|}$$

and ρ is defined by (12). On the other hand, from (7)

$$\|s_2\| \leq (1 + |\lambda - \sigma| \cdot \|(B - \sigma I)^{-1}\|)\|e\| = (1 + \rho)\|e\|,$$

and from (8)

$$\|s_1\| \leq |\lambda - \sigma|^{-1}\|h\|\|s_2\| = \rho^{-1}\frac{\|h\|}{\mathrm{sep}(\sigma, B)}(1 + \rho)\|e\|.$$

Hence

$$\|s\| \leq \left(1 + \rho^{-1}\frac{\|h\|}{\mathrm{sep}(\sigma, B)}\right)(1 + \rho)\|e\|.$$

It follows that

$$\frac{\|\hat{e}\|}{\|s\|} \geq \frac{\tau\rho^2}{\left(\rho + \frac{\|h\|}{\mathrm{sep}(\sigma,B)}\right)(1 + \rho)}.$$

Letting

$$\eta = \frac{\|h\|}{\mathrm{sep}(\sigma, B)},$$

we require that the relative error γ in s satisfy

$$\gamma \leq \frac{\tau\rho^2}{(\rho + \eta)(1 + \rho)} \equiv \bar{\gamma}(\rho). \tag{14}$$

Note that for the values of ρ we are interested in, the factor $1 + \rho$ does not play an important role, and we shall ignore it.

Let us first consider the case $\eta = 0$, which holds for any normal matrix and specifically for symmetric or Hermitian matrices. The inequality (14) becomes

$$\gamma \leq \tau\rho.$$

This is a nice result: we need only solve the system (6) to a relative accuracy corresponding to the decrease in the error, modified by the factor τ (more on this later).

On the other hand if $\eta > 0$, then when $\rho < 0.5\eta$, the function $\bar{\gamma}(\rho)$ becomes effectively ρ^2/η, and we must have

$$\gamma \leq \frac{\rho^2}{\eta}. \tag{15}$$

$\gamma = 0.001$		$\gamma = 0.01$		$\gamma = 0.1$	
$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$	$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$	$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$
9.8e−03	1.6e−02	1.3e−02	2.1e−02	7.3e−02	1.2e−01
7.0e−05	7.2e−03	2.6e−04	2.0e−02	9.2e−02	1.3e+00
6.7e−07	9.5e−03	2.5e−06	9.9e−03	3.7e−02	4.0e−01
6.5e−09	9.8e−03	2.5e−08	9.7e−03	8.3e−02	2.2e+00
6.6e−11	1.0e−02	2.4e−10	9.7e−03	1.8e−02	2.1e−01
6.7e−13	1.0e−02	2.7e−12	1.1e−02	3.4e−03	1.9e−01
6.6e−15	1.0e−02	2.4e−14	8.9e−03	7.2e−04	2.1e−01

Figure 1: Example with $\eta = 0$

Thus if $\eta = 1$ and $\rho = 10^{-3}$ we have to solve the system to a relative accuracy of 10^{-6} to get full reduction at each step.

The number $1 + \eta$ is a bound on the condition number of the eigenvalue λ [2, p. 48]. But we cannot really argue that the unfavorable requirement (15) is an artifact of ill-conditioning, since we have just seen that an eigenvalue with a condition number of 2 can cause the ρ^2 problem. However, two factors mitigate the problem. First, we never expect to have very small values of ρ. If we are working in IEEE double precision, we can reduce the error by a factor of at most 10^{-16}. Thus a $\rho = 10^{-8}$ would give convergence in two iterations, and little would be gained by using the residual inverse power method. The second factor is that we do not have to opt for a full reduction in the error. If we increase γ by a factor of 10, we simple handicap the reduction by the same factor. But eventually, the residual inverse power method will converge.

We give two examples to support our analysis. They are based on a diagonal matrix A of order 51, with equally spaced eigenvalues in the interval $[0, 1]$. We consider the 25th eigenvalue 0.48, and use a shift $\sigma = 0.4802$. Since $\text{sep}(\lambda, B) = 0.02$, we have $\rho = 0.01$.

In the first example, we take $\eta = 0$, so that $\bar{\gamma}(\rho) \cong 0.01\tau$. Figure 1 exhibits the behavior of the algorithm for $\gamma = 0.001, 0.01, 0.1$. The starting vector is random, but the same for all three cases. For each value of γ we display the value $\|\hat{e}\|$ of the error after the current step, and the ratio $\|\hat{e}\|/\|e\|$.

When $\gamma = 0.001$, so that we are solving the system (6) to greater accuracy than is required by our theory, we get steady convergence. The second column shows that the convergence ratios are near 0.01.

For $\gamma = 0.01$, the convergence is reasonable, but a little slower than for $\gamma = 0.001$. This is because the lesser accuracy in the solution throws away some information about the correction.

For $\gamma = 0.1$, the process flounders for a few iterations and then begins to converge. Here it is important to keep in mind that our analysis is only valid up to second order

Figure 2: Example with $\eta = 1$

$\gamma = 0.0001$		$\gamma = 0.001$		$\gamma = 0.01$	
$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$	$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$	$\|\tilde{e}\|$	$\|\tilde{e}\|/\|e\|$
8.8e−03	1.4e−02	9.2e−03	1.5e−02	1.3e−02	2.1e−02
1.3e−04	1.4e−02	1.1e−03	1.2e−01	2.9e−02	2.2e+00
7.5e−07	6.0e−03	4.9e−06	4.5e−03	6.5e−03	2.3e−01
6.0e−09	7.9e−03	7.8e−08	1.6e−02	1.2e−03	1.8e−01
4.8e−11	8.0e−03	4.0e−09	5.1e−02	2.2e−04	1.9e−01
5.0e−13	1.0e−02	1.7e−11	4.4e−03	4.3e−06	1.9e−02
4.2e−15	8.4e−03	6.8e−13	3.9e−02	8.6e−07	2.0e−01

Figure 3: Example with $\eta = 1$.

terms. Initially, these terms dominate and prevent convergence. Only by chance do the second order terms become small enough enable convergence. In repeated runs of this experiment with different starting vectors, more often than not the iteration failed to converge at all.

In performing the experiments we also compared the value of the bound (14) with the actual error, and found them to be reasonably close. Unfortunately, the bound is not computable, and the proper value of γ must be determined empirically. Curiously, τ decreased with increasing γ, being very near one for $\gamma = 0.001$ and about 0.1 for $\gamma = 0.1$. The reason is that for small γ the error e is concentrated in a few components of u. As γ increases, the greater error introduced into s levels the error in u, thus decreasing τ.

As a second example, the elements $A_{24,25}$ and $A_{24,26}$ were set to make $\eta = 1$. The value of $\bar{\gamma}$ drops to about 0.0001. Figure 3 exhibits the results for $\gamma = 0.0001, 0.001, 0.01$. The method converges for each of the three values, but more slowly as γ increases, as predicted by our theory.

To conclude, the residual inverse power method appears to be an effective way of reducing the the accuracy required in the solutions of the linear systems of the inverse power method — certainly for normal matrices. For nonnormal matrices, our analysis shows that a modest degree of deviation from normality can be tolerated, and our limited experiments confirm this assertion. There are two drawbacks to the method. First, the residual must be computed at each step; however, this additional cost is likely to be small compared to the cost of solving (6) for the correction s, even at reduced accuracy. The second drawback is the problem of determining the relative accuracy to which the systems must be solved in the nonnormal case. Unfortunately, we cannot use (14), since quantities η and τ will not usually be available. Thus the user must be guided by experiment and experience. Fortunately, the method is very easy to implement and

try out.

Acknowledgment

The author would like to thank the Mathematical and Computational Sciences Division of the National Institute of Standards and Technology for the use of their facilities during the development of this project.

References

[1] V. Simoncini and D. B. Szyld. Recent computational developments in krylov subspace methods for linear systems. Research Report 05-9-25, Department of Mathematics, Temple University, 2005. Revised May 2006. To Appear in *Numerical Linear Algebra with Applications.*.

[2] G. W. Stewart. *Matrix Algorithms II: Eigensystems.* SIAM, Philadelphia, 2001.