Staggered Aitken Acceleration for EM

Tim Hesterberg <timh@insightful.com>
Insightful Corp. *

Abstract

The EM algorithm can be very slow to converge. This can be sped up by Aitken acceleration, a "step-lengthening" method in which the direction between successive parameter vectors is chosen by vanilla EM, but the step size is modified. Aitken is particularly effective when convergence is dominated by a single large eigenvalue, with other eigenvalues near zero. For other situations there are multivariate versions of Aitken, but they can be unstable.

We propose a "multiple univariate" version of Aitken, where a sequence of step length factors is used to speed convergence for all eigenvalues, without explicitly identifying the eigenvalues.

Keywords: EM algorithm, Aitken acceleration, acceleration, convergence, relaxation

1 Introduction

Step-lengthening methods apply to certain kinds of linearly convergent algorithms, including the EM method and some iterative procedures in linear algebra. Let θ be a vector-valued parameter of interest of length p, g an iterative operation that produces a new estimate for θ given the current estimate

$$\hat{\theta}_{k+1} = g(\hat{\theta}_k)$$

and let $\theta^* = \hat{\theta}_\infty$ be the optimum value. Let $J = J(\theta)$ be the gradient of g at θ, and $J_k = J(\hat{\theta}_k)$. We have the following two relationships:

$$\hat{\theta}_{k+1} - \theta^* \doteq J_k(\hat{\theta}_k - \theta^*) \tag{1}$$

and

$$\Delta_{k+1} = \hat{\theta}_{k+1} - \hat{\theta}_k \doteq J_k \Delta_k = J_k(\hat{\theta}_k - \hat{\theta}_{k-1}). \tag{2}$$

In a linear problem J is independent of θ and the equalities hold exactly. More generally, J is continuous, and approaches $J(\theta^*)$ as k increases.

*Acknowledgments: This work was supported by NIH 2R44CA65147-02.

The convergence of the sequence (θ_k) is determined by the eigenvalues of J (at θ^*). We have

$$\theta_{k+j} - \theta^* \doteq J^j(\theta_k - \theta^*)$$

with equality in linear problems. I assume that all eigenvalues of J are in $[0, 1)$ (this assumption may be relaxed in some cases). The largest eigenvalue λ determines the convergence rate asymptotically, and corresponds to the *fraction of missing information* in an EM problem. The corresponding eigenvalue is the *least-favorable direction*. Figure 1 gives a small example, when $p = 2$ and the eigenvalues are 0.9 and 0.5.

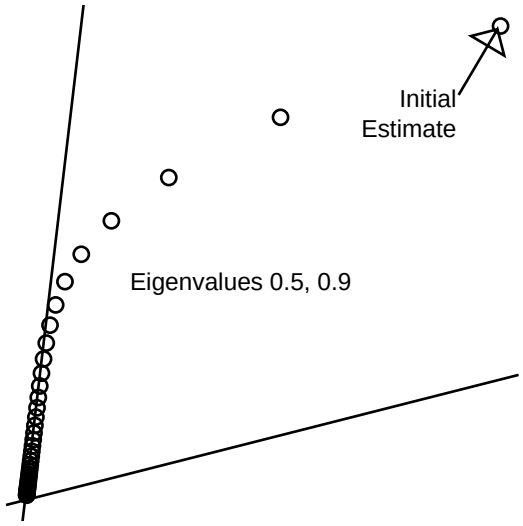

Figure 1: Linear convergence in two dimensions, eigenvalues 0.9 and 0.5

2 Aitken Acceleration

Ignoring the changes in J from one iteration to the next, we have

$$\begin{aligned}\theta^* &= \hat{\theta}_k + \sum_{j=1}^{\infty} \Delta_{k+j} \\ &\doteq \hat{\theta}_k + \sum_{j=1}^{\infty} J^j \Delta_k \\ &= \hat{\theta}_k + J(I-J)^{-1}\Delta_k \\ &= \hat{\theta}_{k-1} + (I-J)^{-1}\Delta_k \qquad (3)\end{aligned}$$

where the sum converges because the eigenvalues are less than 1. If J or a good approximation to it is known this would provide an estimate of θ^*. The idea in Aitken acceleration is to estimate J using the most recent $p+1$ values of Δ. It has been used successfully for small p in (Laird et al. 1987) In high dimensions the full version of Aitken is impractical because of storage requirements, and may fail to converge even in small dimensions, even with some safeguards (Lansky and Casella 1990). The process is also numerically instable. Jamshidian and Jennrich (1997) give references, from the EM and numerical analysis literature.

It's no surprise that Aitken acceleration fails. Let $\mathbf{Q}\Lambda\mathbf{Q}' = J$ be the eigen-decomposition of J, with eigenvalues Λ_j for $j = 1, \ldots, p$ (with $\lambda = \Lambda_1$) and corresponding eigenvectors $\mathbf{Q}_j$ (columns of $\mathbf{Q}$), then $\mathbf{Q}_j \cdot \Delta_k \to c_j \Lambda_j^k$ and $\mathbf{Q}_j \cdot (\hat{\theta}_k - \theta^*) \to C_j \Lambda_j^k$ for some constants c_j, C_j. Unless all eigenvalues are approximately the same, then for reasonably large k both the step sizes and the "errors" $(\hat{\theta}_k - \theta^*)$ will nearly lie in the subspace of R^p generated by the largest eigenvectors (by which we mean the eigenvectors corresponding to the largest eigenvalues). This near singularity makes the matrix inversions performed in Aitken unstable.

There is potential for Aitken acceleration limited to a lower dimensional subspace generated by the most recent s values of Δ has promise, where s is a small integer, see (Smith et al. 1987; Sidi et al. 1986). But these procedures are more complicated, and less robust in nonlinear problems like EM.

3 Step-lengthening methods

Consider the case of one dimension, $p = 1$. Then λ may be estimated by $\hat{\lambda}_k = \Delta_k / \Delta_{k-1}$, and (3) reduces to $\hat{\theta}^* = \hat{\theta}_{k-1} + (1-\hat{\lambda})^{-1}\Delta_k$. In linear problems this gives an exact approximation. This is an example of a step-lengthening method; the new estimate is the old value, θ_{k-1}, plus a step-length factor $(1-\hat{\lambda})^{-1}$ times the step Δ_k that would be taken by the EM or other linearly convergent algorithm.

In multivariate problems we may consider estimates of the form

$$\tilde{\theta}_k = \hat{\theta}_{k-1} + r_k \Delta_k \qquad (4)$$

where r_k is a constant that may be determined a-priori or adaptively, and $\Delta_k = \hat{\theta}_k - \hat{\theta}_{k-1}$ is the step that would be taken by vanilla EM. Methods of this form are discussed in a number of articles, reviewed below. But first note the effect of step lengthening in terms of the eigen-decomposition of J. We have

$$\mathbf{Q}'(\hat{\theta}_k - \theta^*) = \Lambda\mathbf{Q}'(\hat{\theta}_{k-1} - \theta^*) = \sum_j \Lambda_j \mathbf{Q}_j \cdot (\hat{\theta}_{k-1} - \theta^*)$$

and

$$\mathbf{Q}'(\tilde{\theta}_k - \theta^*) = \sum_j (1 - r_k(1-\Lambda_j))\mathbf{Q}_j \cdot (\hat{\theta}_{k-1} - \theta^*).$$

Instead of a convergence factor of Λ_j in the direction of eigenvector $\mathbf{Q}_j$, the factor is now $1 - r_k(1-\Lambda_j)$. Since the eigenvalues are in the range $[0, 1)$, the convergence factors are in the range $[1 - r_k, 1 - r_k(1-\lambda))$.

Consider first the case of constant multiplier $r_k = r$. This is termed the "relaxation method," or "successive over-relaxation," for accelerating convergence for some linear algebra problems, e.g. (Golub and Loan 1996). Hämmerlin and Hoffmann (1991) give the optimum value $r = 2/(2-\lambda-\Lambda_p)$ in terms of the largest eigenvalue λ and the smallest (or most negative) eigenvalue Λ_p. Lange (1995) indicate that if $r < 2$ (note the strict inequality) then the method converges, and indicate that "For problems with a high proportion of missing data, the value of $r = 2$ often works well." However, we note that this is not true if any eigenvalues are near zero, e.g. if the number of missing values for one variable is small. If there is a zero eigenvalue, the corresponding convergence factor is $1 - 2(1-0) = -1$, so that the algorithm oscillates without converging. Otherwise, if there is a sufficiently small eigenvalue the convergence is slower than with no multiplier.

Jamshidian and Jennrich (1997) indicate that "Step-lengthening seems to give only small gains over EM compared to $\cdots$", and cite three references. The evidence cited is not sufficient to reject step-lengthening methods. We discussed (Lange 1995) above. The step length calculations in (Jamshidian and Jennrich 1994) involve approximate optimization of an objective function given a direction, a classical technique in nonlinear optimization that bears little relation to the procedures we discuss below. Finally, Laird et al. (1987) do not seem to have applied step-lengthening in any examples. They considered it, using $r_k = (1-\hat{\lambda}_k)^{-1}$, with $\hat{\lambda}_k = p^{-1}\sum_{j=1}^{p} \Delta_{k,j}/\Delta_{k-1,j}$, but quit because it was apparent that something was wrong—asymptotically the terms in the summation should be approximately equal, but they varied substantially in their example. We note that their $\hat{\lambda}_k$ is

very unstable because some of the values in the denominator could be near zero.

More stable estimates of λ are available. We run EM twice, first from $\hat{\theta}_{k-2}$ to obtain $\hat{\theta}_{k-1}$, then from $\hat{\theta}_{k-1}$ to obtain $\hat{\theta}_k$. Then three estimates of λ, in order of increasing size, are:

$$\hat{\lambda}^{(1)} = \frac{\Delta_k \cdot \Delta_{k-1}}{|\Delta_{k-1}|^2} \quad (5)$$

$$\hat{\lambda}^{(2)} = \frac{|\Delta_k|}{|\Delta_{k-1}|} \quad (6)$$

$$\hat{\lambda}^{(3)} = \frac{|\Delta_k^2|}{\Delta_k \cdot \Delta_{k-1}} \quad (7)$$

The three estimates are equivalent if Δ_k and Δ_{k-1} are parallel, e.g. if there is only one non-zero eigenvalue. The third is the most accurate (in linear problems, and asymptotically for other problems), and even this one is conservative. In other words, $\hat{\lambda}^{(1)} \leq \hat{\lambda}^{(2)} \leq \hat{\lambda}^{(3)} \leq \lambda$. The proof follows from writing $\Delta_{k-1} = \sum_j c_j \Lambda_j^{k-1} \mathbf{Q}_j = \sum_j b_j \mathbf{Q}_j$ for some constants c_j and b_j, so that $\Delta_{k-1} \cdot \Delta_{k-1} = \sum b_j^2$, $\Delta_{k-1} \cdot \Delta_k = \sum b_j^2 \Lambda_j$, and $\Delta_k \cdot \Delta_k = \sum b_j^2 \Lambda_j^2$. Then $\hat{\lambda}^{(3)} = \sum b_j^2 \Lambda_j^2 / \sum b_j^2 \Lambda_j$ is a weighted average of the values of Λ_j (with weights $b_j^2 \Lambda_j$), so must be less than or equal to λ. The other inequalities are obtained in a similar fashion.

We have used the most conservative estimate and most liberal estimates successfully, with moderate to substantial speedups for EM estimates of Normal parameters in randomly generated datasets with missing values. To do this, we alternated between unaccelerated and accelerated steps; after one step from $\hat{\theta}_{k-2}$ to $\hat{\theta}_{k-1}$ with with no acceleration, we compute the EM step $\hat{\theta}_k$, estimate λ using the sequence of three parameter estimates, and replace the final parameter vector $\hat{\theta}_k$ with an accelerated version.

This alternating procedure gives some interesting behavior. The estimates of the largest eigenvalue converge to an oscillating sequence, as in Figure 2; the subsequences formed by every other value converge to one of two limits. The upper limit is between the largest and second largest eigenvalues; the lower limit may be above or below the second eigenvalue, or smaller yet, depending on the whole set of eigenvalues and on the initial parameter estimates.

The cause of this is that when a relatively accurate estimate of λ is used for acceleration, the error $\tilde{\theta}_k - \theta^*$ in the direction corresponding to the corresponding eigenvalue $\mathbf{Q}_1$ is substantially reduced. Then the next estimate of λ is much smaller, because most of the movement in the sequence of estimates occurs in the other eigen directions. Acceleration with a smaller $\hat{\lambda}$ reduces the errors in the directions corresponding to smaller eigenvalues, so the next time λ is estimated the changes in direction $\mathbf{Q}_1$ again dominate.

Variability in $\hat{\lambda}$ was unexpected, and at first glance seems undesirable. However, followup experiments with the oscillating estimate replaced by the exact largest eigenvalue turned out even worse. When λ is large the multiplier is large, and if some eigenvalues are small the accelerated sequence diverges in the corresponding directions. We see this in Figure 3, where accelerating by a factor of 10 on alternate steps based on $\lambda = 0.9$ causes divergence. In contrast, accelerating on alternating steps with an oscillating sequence of multipliers converges substantially vaster than vanilla EM.

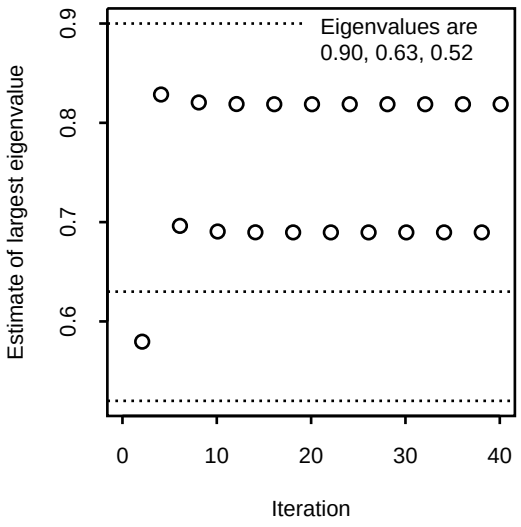

Figure 2: Estimates of $\hat{\lambda}$ in alternate iterations

It turns out that variability in multipliers is desirable, in order to improve convergence in all directions, not just the direction corresponding to one eigenvalue.

3.1 Multi-step step-lengthening

Recall that the effect of a step-length multiplier r_k is to produce a convergence factor $1 - r_k(1 - \Lambda_j)$ in the direction of the jth eigenvalue $\mathbf{Q}_j$. In fact, the monomial

$$f_r(x) = 1 + r(x - 1) \quad (8)$$

describes the convergence factors of all eigenvalues, for a single step. This passes through $(1, 1)$ (so acceleration has no benefit if an eigenvalue is exactly 1) and has slope r (so convergence is approximately r times faster for eigenvalues very near 1).

Furthermore, we can characterize the effect of multiple iterations, possibly with different acceleration factors, by the product of monomials

$$f_{\mathbf{r}}(x) = \prod 1 + r_k(x - 1) \quad (9)$$

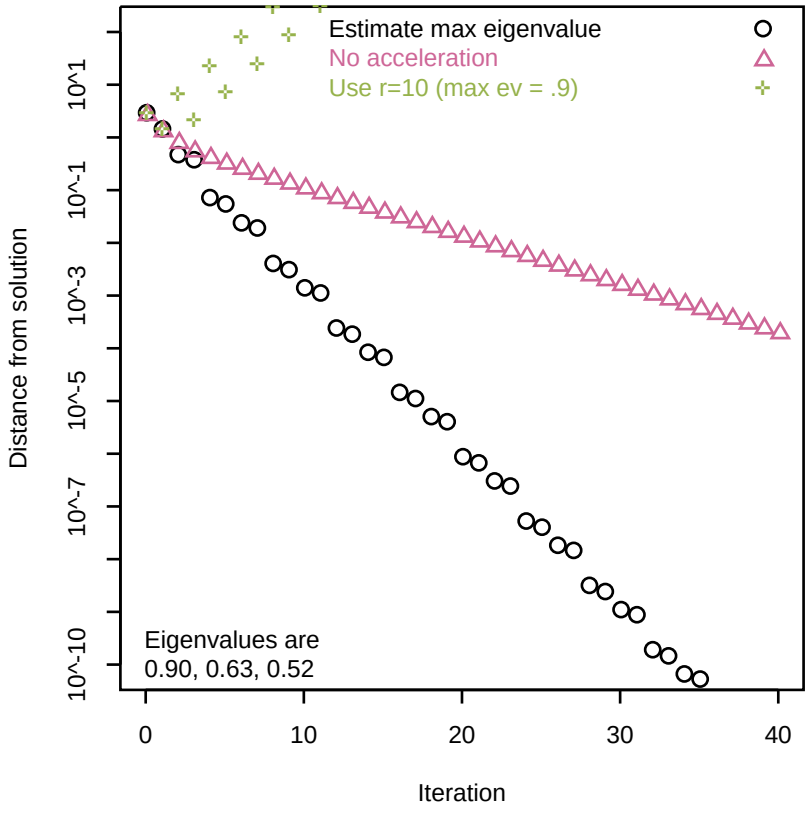

Figure 3: Convergence for vanilla EM, and acceleration on alternating steps with estimated λ, and half-Chebyshev golden ratio procedure

where $\mathbf{r} = (r_1, \ldots, r_k, \ldots, r_K)$. Note that the effect of K iterations of (4) in the direction of eigenvalue $\mathbf{Q}_j$ is given by

$$\mathbf{Q}_j \cdot (\hat{\theta}_K - \theta^*) = f_{\mathbf{r}}(\Lambda_j)\mathbf{Q}_j \cdot (\hat{\theta}_0 - \theta^*).$$

This polynomial allows us to analyze acceleration sequences. A good K-step procedure makes $|f_{\mathbf{r}}(\Lambda_j)|$ as near zero as possible for every j. Vanilla EM corresponds to using $r_k = 1$, with $\max_j |f_{\mathbf{r}}(\Lambda_j)| = \lambda^K$. A constant multiplier of $r_k = 2$ makes $f_{\mathbf{r}}(0) = (-1)^K$, which does not converge to zero. If the eigenvalues were known (and the application were exactly linear) we could obtain perfect estimates after p iterations by choosing $\mathbf{r}$ so that the polynomial has roots at each Λ_j, using $r_k = (1 - \Lambda_k)^{-1}$ for $k = 1, \ldots, p$. Any permutation of this $\mathbf{r}$ would also work.

Steps with $r_k > 2$ can actually increase the distance from the final solution, by increasing the distance in the directions of the small eigenvectors. However such steps should not be viewed in isolation, but rather as part of a sequence of steps. Figure 4 shows six polynomials, some of which have multipliers greater than 2. The first row of panels is for straight EM, either 1 or 4 steps (the latter has vertical lines at $x = 0.5$ and $x = 0.8$ for comparison with later graphs). The left panel in row 2 shows the effect of a single step multiplier of 3; in isolation it is unstable, with f off the page for small eigenvalues. However, when combined with a smaller multiplier, e.g. $r = 1$ (i.e. no acceleration) as in the right middle panel, it results in an effective two-step procedure.

If the eigenvalues are unknown but are known to be in the range $[s, t]$, we might choose $\mathbf{r}$ to minimize the maximum value of $|f_{\mathbf{r}}(x)|$ over the range $s \le x \le t$; the roots of this minimax polynomial are approximately linear translates of the roots of the Chebyshev polynomial,

$$z_k = 1 - 1/r_k = h(\cos(\pi(k - 1/2)/K)) \qquad (10)$$

where z_k is the zero associated with r_k and h is the linear transformation that maps the interval $[-1, 1]$ to $[s, t]$ (any permutation of this $\mathbf{r}$ may be used). ("Approximately", because the minimax problem here is not precisely the same as the usual interpolation minimax problem that leads to Chebyshev's polynomial, because here the monomials have slopes that depend on the roots.)

There is a technique in the numerical analysis literature known as Chebyshev acceleration, which is different than the procedure here. We discuss this in Section 3.2 below.

The final two graphs in Figure 4 show the the minimax solutions for the ranges $[0, 0.5]$ and $[0, 0.8]$, respectively. By comparing the maximum values of these polynomials on the corresponding intervals with the x^4 polynomial shown in the right side of the first row over the same ranges, we see that the errors obtained by accelerating based on Chebyshev roots can be much smaller than for vanilla EM.

3.1.1 Practical Issues

We now turn to four practical issues: the number of iterations is not generally determined in advance, order matters in nonlinear problems like EM, reasons to be conservative in nonlinear problems, and the range of eigenvalues is not known.

If K is not determined in advance, then we may replace (10) with

$$z_k = h(\cos(\pi u_k)) \qquad (11)$$

where u_k is a sequence of numbers uniformly distributed between 0 and 1, either randomly or deterministically.

We particularly recommend a "golden ratio sequence"

$$u_k = \mathrm{fp}(\tau k) \qquad (12)$$

where $\mathrm{fp}(x) = x - \lfloor x \rfloor$ is the fractional part of x and $\tau = (3 - \sqrt{5})/2$ is derived from the golden ratio; the result is a sequence such that the gaps between sorted roots are nearly as small as possible. This is shown in Figure 5.

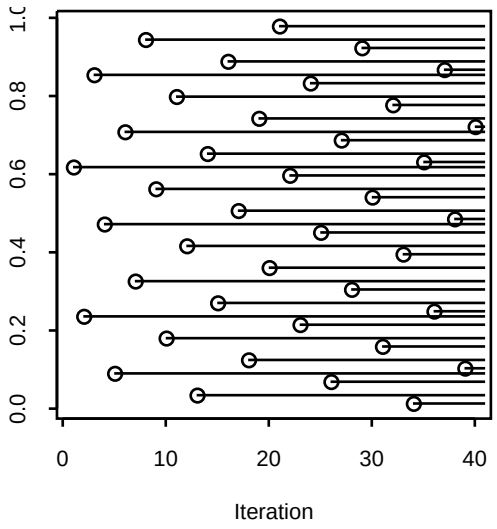

Figure 5: The golden ratio sequence (12)

In a linear problem the order in which roots are used does not matter. However, in EM and other nonlinear applications the order does matter. As $\hat{\theta}$ changes, $J(\hat{\theta})$ and its eigenvalues and eigenvectors change as well. One cannot assume that an eigen direction which has been killed (in that the error in that direction is small) will stay dead. The golden ratio sequence (12) is particularly good when order matters, because it fills in recent gaps.

In the case of a nonlinear procedure such as EM there is some value in choosing values of r_k which are smaller than

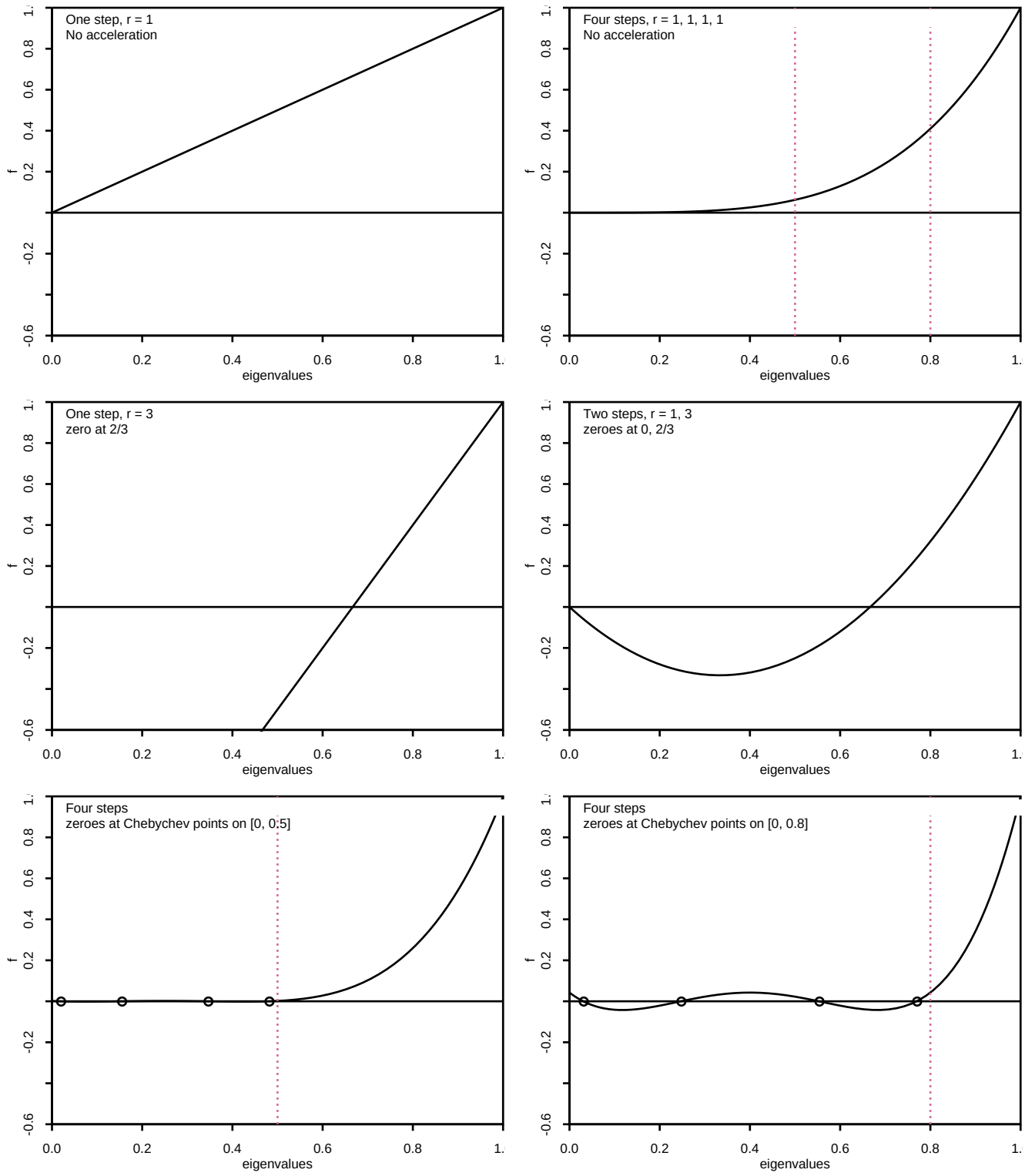

Figure 4: Convergence polynomials. Each polynomial f represents the convergence rate as a function of the eigenvalue — e.g. if f(0.4) = 0.1, then the procedure converges 90% of the way toward the solution, within the subspace spanned by eigenvectors whose eigenvalues are 0.4. For example, the procedure in the bottom left is very effective if all eigenvalues are less than 0.5. The degree of the polynomial matches the number of steps K used.

indicated by Chebyshev roots, particularly in early stages of the EM, because this is more conservative and less likely to cause problems. We have observed divergence in EM applications when too much acceleration is applied too early.

There is an additional reason to prefer using smaller multipliers than suggested by Chebyshev roots. (9) can be rewritten as

$$f_{\mathbf{r}}(x) = \prod 1 + r_k(x-1) = (\prod r_k)(\prod x - z_k). \quad (13)$$

The Chebyshev roots give the approximate minimax solution for the simple polynomial $\prod(x - z_k)$, but $f_{\mathbf{r}}(x)$ has an additional term $\prod r_k$, which is smaller when multipliers are smaller.

The last practical issue is that the range of eigenvalues is unknown. The solution at the lower end is simple; pretend the smallest eigenvalue is zero. It is difficult to estimate the smallest eigenvalue in EM applications, and there is typically little value in doing so; little efficiency is lost by choosing a sequence of roots as if the smallest eigenvalue were 0.

The affect of estimating the largest eigenvalue λ is more interesting, in three ways. First, estimates (5–7) are conservative, weighted averages of all eigenvalues with weights that depend on the magnitudes of current errors in the eigenvalue directions. In the absence of acceleration, a sequence of estimates $\hat{\lambda}_k$ converges to λ from below. This remains true if acceleration is used conservatively, but is not true for more general acceleration sequences. We noted above the "interesting behavior" when acceleration is applied on alternate steps, that estimates $\hat{\lambda}_k$ ultimately oscillate. And that occurs even with a procedure that is conservative in the sense that all acceleration factors were smaller than the optimal factor for the largest eigenvalue. Less conservative acceleration can give less stable λ estimates.

Second, depending how the sequence of values depends on $\hat{\lambda}_k$, there could be undesirable bunching and gaps of the roots. As a toy example, suppose that all roots are of the form $\hat{\lambda}_k j/100$ for integers $0 \le j \le 100$ and that estimates $\hat{\lambda}$ alternate between between 0.25 and 0.5, and that even values of j occur only when $\hat{\lambda}_k = 0.5$. Then the realized sequence of roots are of the form $.25l/100$, where l is an integer from the set $\{0, 1, 1, 2, 3, 3, 4, 5, 5, \ldots, 49, 49, 50, 51, 53, 55, \ldots, 99\}$. In the upper half of the range there are larger gaps, and alternate small integers are chosen with double frequency.

Second, use of (5–7) require that the first of a series of two steps be unaccelerated. Every unaccelerated step corresponds to an extra zero at 0.0. These unaccelerated steps are conservative, allowing other steps to be somewhat liberal.

3.1.2 Asymptotic Convergence

We may combine a sequence of uniform numbers u_k with any monotone transformation to produce a sequence of roots z_k within [0,1] with density g, e.g. by the inverse distribution transformation $G^{-1}(u_k)$. Then the asymptotic convergence rate at eigenvalue x is governed by

$$\lim_{K\to\infty} K \log(|f_{\mathbf{r}}(x)|) = \int_0^1 (\log(|x-z|) - \log(1-z)) g(z) dz. \quad (14)$$

If the sequence of uniform numbers is random, then the convergence holds almost surely; if deterministic then it should be interpreted as referring to the convergence for random points in a sequence of neighborhoods of x, with neighborhood width decreasing to zero at a rate slower than $1/K$.

The term $-\log(1-z)$ in the integral favors smaller multipliers; this is the analog of the additional term $\prod r_k$ in (13).

3.1.3 Recommendation

As a general rule our bias is toward conservative procedures, to avoid convergence problems, because of the complications caused by needing to estimate the maximum eigenvalue, and because of the extra term $\prod r_k$ in (13). Hence we do not recommend using linear translates of the Chebyshev polynomial; the large number of roots at the right end of the range makes this procedure relatively unstable and blows up the extra factor.

Instead, we begin with a "half-Chebyshev" idea–to use roots which have the high density near zero, but low density near the estimated maximum eigenvalue. Consider the Chebyshev roots over the range $[0, 2]$, which for fixed K are $z_k = 1 + \cos(\pi(k - 1/2)/K))$ for $k = 1, \ldots, K$. The bottom half of these have the desired high density near zero. Hence, we begin with values of the form

$$z_k = 1 + \cos(\pi(1 + u_k)/2) \quad (15)$$

where u_k is a sequence of values in $[0, 1)$.

We generate values u_k from the golden ratio sequence, calculate the corresponding z_k, then accelerate using some and skip others which are too large. What is too large? First, in EM we particularly want to be conservative in early steps, while nonlinear behavior may still be dominating. Hence we skip roots exceeding $j/20$ at step j (the divisor may be adjusted, depending on the degree of nonlinearity of the application). Second, we skip roots that exceed the current estimate $\hat{\lambda}$.

It is not necessary to estimate $\hat{\lambda}$ every other iteration; this would result in more unaccelerated steps than necessary, given that our half-Chebyshev transformation has many roots near 0.0. We estimate every five iterations. The resulting sequence of roots is shown in Figure 6. Note that it tends to be conservative at the beginning.

Note that we use a fixed sequence of roots and skip those that exceed $\hat{\lambda}$, rather than rescaling (e.g. $\hat{\lambda} z_k$), because rescaling can cause odd bunching and large gaps. (This turned out

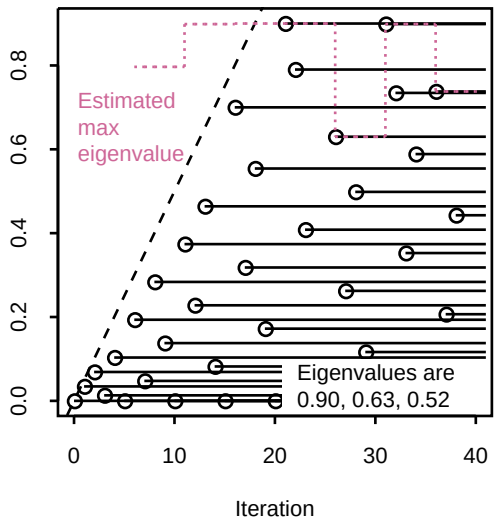

Figure 6: Roots, based on a transformation of the golden ratio sequence, with $\hat{\lambda}$ estimated every 5 iterations, roots exceeding $\hat{\lambda}$ skipped, and roots exceeding $j/20$ skipped at the beginning.

to be the cause of some quite puzzling results in some experiments.)

The estimate $\hat{\lambda}$ is somewhat unstable, generally increasing toward λ but sometimes declining after a zero has been picked that is near λ. It would probably be desirable to keep a working estimate which is "sticky", jumping up whenever a new estimate (5–7) is larger, but decreasing only when two or three are smaller.

The resulting convergence is shown in Figure 7 The new sequence is conservative at the beginning, but eventually makes large improvements in the magnitude of errors.

3.2 Polynomial Acceleration

An alternate approach found in the numerical analysis literature is "polynomial acceleration" (Hageman and Young 1981), generally applied to purely linear problems rather than nonlinear applications such as EM. Given an unmodified sequence of estimates $\hat{\theta}_k$, $k = 1, \ldots$, they create a modified sequence of the form

$$\tilde{\theta}_k = \sum_j = 0^k c_{j,k} \hat{\theta}_j \tag{16}$$

with $\sum_{j=0}^{k} = 1 \forall k$. In effect each modified estimate is a weighted average of the unmodified estimates, with weights that may fall outside the range $[0, 1]$. The convergence of this is also analyzed using polynomial functions of eigenvalues, and coefficients $c_{j,k}$ are chosen to minimize the maximum value of the polynomial over the known or estimated range of eigenvalues.

A special case is Chebyshev acceleration, or the Chebyshev semi-iterative method (Golub and Loan 1996; Hageman and Young 1981). This can be written in a form similar to (4), but based on both the current EM step and the difference between two most recent modified estimates

$$\tilde{\theta}_k = \tilde{\theta}_{k-1} + r_k \Delta_k + s_k (\tilde{\theta}_{k-1} - \tilde{\theta}_{k-2} \tag{17}$$

where Δ_k is the EM step from $\tilde{\theta}_{k-1}$. This eliminates the need to store many previous parameter estimates.

In polynomial acceleration, the zeroes of the polynomial at step $k+1$ need not be a superset of those at step k. In the case of Chebyshev acceleration, the use of nonzero s_k causes previous zeros of the convergence polynomial to shift; if at one step the zeroes are zeroes of the Chebyshev polynomial of degree k, at the next step the zeroes are zeroes of the Chebyshev polynomial of degree $k+1$ (in each case rescaled to the interval given by the range of eigenvalues).

However, such shifting of previous zeroes should be used with caution in nonlinear applications such as EM. The mechanism, perturbation of the current accelerated step by some fraction of the previous step, presumes that the previous step was governed by the same linear behavior as the current location and the solution. Some of the practical issues in Section 3.1.1 also argue against this procedure in EM.

3.3 Other Acceleration Work

Varadhan and Roland (2004) discuss a variety of acceleration procedures, but we became aware of this paper too late to study it carefully.

4 Summary

In summary, if the maximum eigenvalue of a linearly-convergent procedure is known but the other eigenvalues are unknown, then a multi-step procedure is effective where the acceleration constants are chosen so that the roots of the corresponding convergence polynomial are roots of a Chebyshev polynomial, rescaled to the range from 0 to the largest eigenvalue. If the number of steps is not determined in advance, then a sequence based on a nonlinear transformation of the fractional part of multiples of a constant derived from the golden ratio is effective.

For EM, we prefer a conservative approach, with roots derived from the left half of the roots of a Chebyshev polynomial, and skipping roots which are too large.

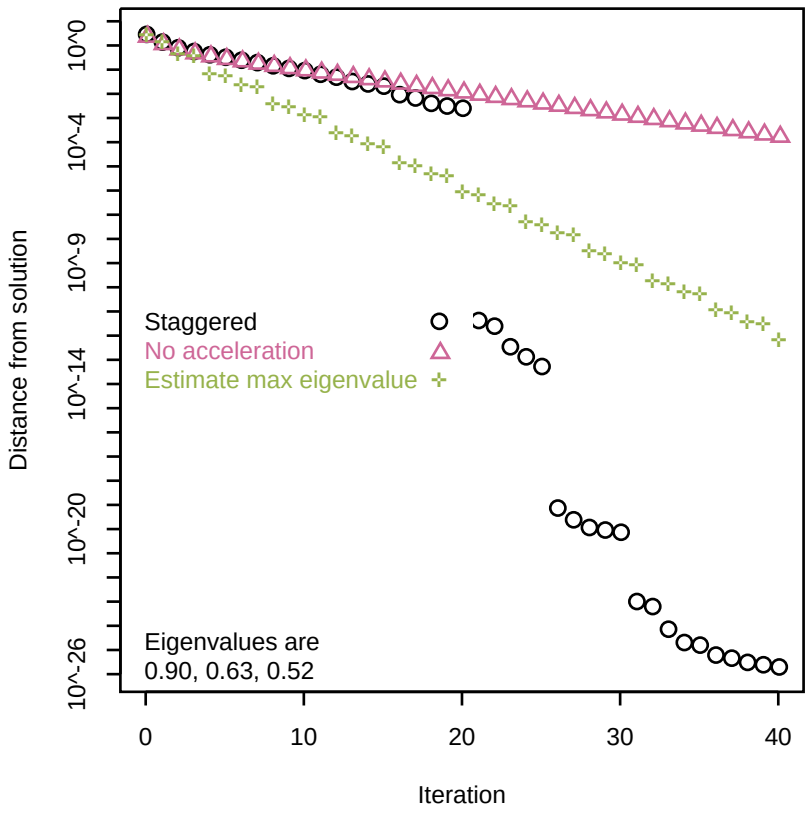

Figure 7: Convergence for vanilla EM, and acceleration on alternating steps with estimated or known λ

References

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edition.

Hageman, L. A. and Young, D. M. (1981). *Applied Iterative Methods*. Academic Press, New York.

Hämmerlin, G. and Hoffmann, K. (1991). *Numerical Mathematics*. Springer-Verlag: New York.

Jamshidian, M. and Jennrich, R. I. (1994). Conjugate Gradient Methods in Confirmatory Factor Analysis. *Computational Statistics and Data Analysis*, 17:247–263.

Jamshidian, M. and Jennrich, R. I. (1997). Acceleration of the EM Algorithm by using Quasi-Newton Methods. *Journal of the Royal Statistical Society, Series B*, 59(3):569–587.

Laird, N., Lange, N., and Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association*, 82:97–105.

Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 52(2):425–437.

Lansky, D. and Casella, G. (1990). Improving the EM Algorithm. In Page, C. and LePage, R., editors, *Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface*, volume 22, pages 420–424. Interface Foundation of North America, Springer-Verlag.

Sidi, A., Ford, W. F., and Smith, D. A. (1986). Acceleration of Convergence of Vector Sequences. *SIAM Journal on Numerical Analysis*, 23(1):178–196.

Smith, D. A., Ford, W. F., and Sidi, A. (1987). Extrapolation Methods for Vector Sequences. *SIAM Review*, 29(2):199–233.

Varadhan, R. and Roland, C. (2004). Squared Extrapolation Methods (SQUAREM): A New Class of Simple and Efficient Numerical Schemes for Accelerating the Convergence of the EM Algorithm. Working Papers 63, Johns Hopkins University, Dept. of Biostatistics.