SCALING FOR ORTHOGONALITY

By

Alan Edelman

and

G.W. Stewart

IMA Preprint Series # 984

May 1992
Scaling for Orthogonality*
Alan Edelman† and G. W. Stewart‡
April 1992

ABSTRACT
In updating algorithms where orthogonal transformations are accumulated, it is important to preserve the orthogonality of the product in the presence of rounding error. Moonen, Van Dooren, and Vandewalle have pointed out that simply normalizing the columns of the product tends to preserve orthogonality — though not, as DeGroat points out, to working precision. In this note we give an analysis of the phenomenon.

---

*This report is available by anonymous ftp from thales.cs.umd.edu in the directory pub/reports.
†Department of Mathematics and Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720, edelman@math.berkeley.edu. Supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.
‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742. stewart@cs.umd.edu. This work was performed while the author was visiting the Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN 55455.
Scaling for Orthogonality

Alan Edelman
G. W. Stewart

In many updating algorithms it is required to accumulate a product of the form

\[ X_k = Q_1 \cdots Q_{k-1} Q_k, \]

where the matrices \( Q_i \) are orthogonal. Although mathematically speaking \( X_k \) must be orthogonal, in practice rounding error will cause it to drift from orthogonality with increasing \( k \). If we take the deviation of \( X_k^T X_k \) from the identity as a measure of the loss of orthogonality, then typically

\[ \| I - X_k^T X_k \|_F \leq k \theta_n \epsilon_M, \]

where \( \| \cdot \|_F \) is the Frobenius norm, \( \epsilon_M \) is the rounding unit for the arithmetic in question, and \( \theta_n \) is a slowly growing function of the size \( n \) of \( X_k \) (e.g. \( n^{1.5} \)).

As a cure for this problem DeGroat and Roberts [1] have proposed that each \( X_k \) be subjected to a partial reorthogonalization in which the second column is orthogonalized against the first, the third against the second, and so on with all the columns being renormalized after orthogonalization. In a subsequent note on their paper Moonen, Van Dooren, and Vandewalle [2] pointed out that the normalization alone is sufficient to maintain orthogonality and supported their claim with a heuristic argument. In a reply DeGroat pointed out that normalization “does not yield working precision orthogonality.” However, the error remains quite small.

The purpose of this note is to give a more complete analysis of the method, one that explains the phenomena mentioned in the last paragraph. In particular, we show that this method succeeds when the \( Q_i \) manage to transfer off-diagonal error in the matrices \( I - X_i^T X_i \) to the diagonal. We also show that normalizing is the best possible scaling up to to first order. However, it can actually decrease orthogonality in certain unlikely circumstances.

For notational convenience we will drop subscripts and write

\[ \tilde{X} = XQ, \]

where \( X \) is scaled so that its its column norms are one and \( Q \) is orthogonal (for the moment we ignore rounding error). Since \( X \) is normalized, we can write

\[ A \equiv X^T X = I + E, \]
where the diagonals of $E$ are zero. Write

$$\hat{A} \equiv \hat{X}^T \hat{X} = I + \hat{D} + \hat{E},$$

where

$$\hat{D} + \hat{E} = Q^T EQ$$

is a decomposition of $Q^T EQ$ into its diagonal and off-diagonal parts. In this notation, the scaling of $\hat{X}$ amounts to setting

$$\hat{S} = (I + \hat{D})^{-1}$$

and

$$\hat{X} = \hat{X} \hat{S}^{1/2}.$$ 

The deviation from orthogonality of $\hat{X}$ is the Frobenius norm of

$$\hat{E} = \hat{S}^{1/2} \hat{E} \hat{S}^{1/2}.$$ 

The above equations define a recurrence for $E$, $\hat{E}$, etc., which we are going to analyze. But first we will motivate the scaling by comparing it with the optimal scaling, which is characterized in the following theorem.

**Theorem 1.** For any diagonal matrix $D$ let $\text{diag}(D)$ denote the vector consisting of the diagonal element of $D$. Then for all sufficiently small $E$, the optimal scaling matrix $S$ satisfies

$$\hat{A} \circ \hat{A} \text{diag}(S) = \text{diag}(I + D),$$

where $\hat{A} \circ \hat{A}$ is the component-wise product (a.k.a., the Schur or Hadamard product) of $\hat{A}$ with itself.

**Proof.** Regarded as a function of the elements of $S$, the function $\|S^{1/2} \hat{E} S^{1/2}\|^2_F$ is a quadratic function that is bounded below by zero. Differentiating this function and setting the results to zero, we obtain (4). It follows that if (4) has a positive solution, then that solution will provide the optimal scaling. Now $\lim_{E \to 0} \hat{A} \circ \hat{A} = (I + D)^2$. Consequently,

$$\lim_{E \to 0} \text{diag}(S) = \lim_{E \to 0} (\hat{A} \circ \hat{A})^{-1} \text{diag}(I + D) = \text{diag}[(I + D)^{-1}] = \text{diag}(\hat{S}) > 0.$$ 

Hence for all sufficiently small $E$, the solution of (4) is positive.
Equation (5) provides a heuristic justification for the method, since it says that to first order in \( E \) our scaling approximates the optimal scaling. However, the matrix

\[
\hat{A} = \begin{pmatrix}
1 - \epsilon^2 & \epsilon \\
\epsilon & 1 - \epsilon^2
\end{pmatrix}
\]

shows that the method is not guaranteed to increase orthogonality for all small \( E \). Nevertheless, this situation is quite unlikely, as we will now demonstrate by an analysis of the recurrence (3).

First note that from (1) and the unitary invariance of the Frobenius norm we have

\[
\|E\|_F^2 = \|\hat{D}\|_F^2 + \|\hat{E}\|_F^2.
\]

Now the square of the \((i,j)\) element of \( \hat{E} \) is

\[
\frac{\hat{e}_{ij}^2}{(1 + \hat{d}_i)(1 + \hat{d}_j)} \leq \frac{\hat{e}_{ij}^2}{(1 - \|\hat{D}\|_F)^2}.
\]

Here \( \hat{d}_i \) is the \( i \)th element of \( \hat{D} \), and we assume that \( \|\hat{D}\|_F < 1 \). Hence

\[
\|\hat{E}\|_F^2 \leq \frac{\|\hat{E}\|_F^2}{(1 - \|\hat{D}\|_F)^2}.
\]

Setting

\[
\epsilon = \|E\|_F \quad \text{and} \quad \hat{\delta} = \|\hat{D}\|_F,
\]

we have from (6) and (7)

\[
\|\hat{E}\|_F^2 \leq \epsilon^2 \equiv \frac{\epsilon^2 - \hat{\delta}^2}{(1 - \hat{\delta})^2}.
\]

A little extra notation will help us decide when the scaling results in an increase of orthogonality. Since from (6) we have \( \hat{\delta} \leq \epsilon \), we can write

\[
\hat{\delta} = \gamma \epsilon, \quad 0 \leq \gamma \leq 1.
\]

In this notation the equality in (8) becomes

\[
\epsilon^2 \equiv \epsilon^2 \left[ \frac{1 - \gamma^2}{(1 - \gamma \epsilon)^2} \right] \equiv \epsilon^2 \varphi(\gamma).
\]

Thus the problem is to ascertain when \( \varphi(\gamma) \) is less than one. The following facts are easily verified.
1. \( \varphi(\gamma) \geq 1 \) in the interval \([0, 2\epsilon/(1 + \epsilon^2)]\). At \( \gamma = \epsilon \) it assumes a maximum of \((1 - \epsilon^2)^{-1}\).

2. \( \varphi(\gamma) \) decreases monotonically from one to zero on the interval \([2\epsilon/(1 + \epsilon^2), 1]\).

In terms of our iteration, if \( \hat{\delta} \) is too small, roughly less than \( 2\epsilon^2 \), then the scaling has the potential to reduce orthogonality—but not by very much if \( \epsilon \) is at all small. For larger \( \hat{\delta} \) the scaling is guaranteed to increase orthogonality. Otherwise put, multiplication by the matrix \( Q \) moves part of \( E \) to the diagonal where it is eliminated by the scaling. The more of \( E \) that is moved to the diagonal the better.

The amount of \( E \) that is moved will depend on \( Q \), which in turn depends on the application in question. However, it is interesting to note what happens when \( Q \) is chosen at random uniformly from the group of orthogonal matrices. To do so we prove

**Theorem 2.** Let \( Q = (q_1, \ldots, q_n) \) be a random orthogonal matrix, uniformly distributed over the group of orthogonal matrices. Then for any symmetric matrix \( E \)

\[
\mathbb{E}\left( \sum_{i=1}^{n}(q_i^T E q_i)^2 \right) = \frac{1}{n + 2} \left[ \text{trace}(E)^2 + 2\|E\|_F^2 \right],
\]

where \( \mathbb{E} \) is the expectation operator.

**Proof.** Let \( u \) denote a random vector of \( n \) independent standard normals. Let \( r \) denote \( \|u\| \) and \( v = u/r \) (n.b., \( v \) is a typical column of \( Q \)). It is well known that \( v \) is distributed uniformly over the sphere, while \( r^2 \) is independent with \( \chi^2_n \) distribution. Thus using standard results on the moments of the normal and \( \chi^2 \) distributions, we have

\[
\mathbb{E} v_i^4 = \frac{\mathbb{E} u_i^4}{\mathbb{E} r^4} = \frac{3}{n(n + 2)}
\]

and

\[
\mathbb{E}(v_i^2 v_j^2) = \frac{\mathbb{E}(u_i^2 u_j^2)}{\mathbb{E} r^4} = \frac{1}{n(n + 2)}, \quad i \neq j.
\]

It is clearly sufficient to prove the lemma for diagonal matrices, say

\[
E = \text{diag}(\lambda_1, \ldots, \lambda_n).
\]

For this case the result follows easily on expanding \( \sum_{i=1}^{n}(q_i^T E q_i)^2 \) and using the above formulas to take expectations (recall that \( \text{trace}(E)^2 = \|E\|_F^2 + \sum_{i \neq j} \lambda_i \lambda_j \)).
In our application, the trace of $E$ is zero and we have on the average

$$\hat{\delta}^2 = \frac{2}{n+2} \epsilon^2,$$

i.e., $\gamma^2 = 2/(n+2)$. Thus, $\delta$ is of the same order as $\epsilon$, and by the second observations following (9) we can expect to observe an increase of orthogonality. However, this increase decreases as $n$ grows. For if $\epsilon$ is small enough so that the denominator in $\varphi(\gamma)$ can be ignored, an iteration will reduce $\epsilon^2$ on the average by a factor of of only $n/(n+2)$.

Finally, returning to the role of rounding error, its effect is to add errors to $\hat{E}$. The Frobenius norm of this error will be proportional to the rounding unit $\epsilon_M$, say $\theta_n \epsilon_M$. Thus the recurrence (9) must be rewritten in the form

$$\bar{\epsilon} = \varphi(\gamma)^{\frac{1}{2}} \epsilon + \theta_n \epsilon_M.$$

If we assume that $\gamma$ is constant, then this recurrence has the fixed point

$$\epsilon = \frac{\theta_n \epsilon_M}{1 - \varphi(\gamma)^{\frac{1}{2}}} \cong \frac{2 \theta_n \epsilon_M}{\gamma^2},$$

the last approximation holding for small gamma. For example, with random $Q$ we should not expect to reduce the measure of orthogonality much below $(n+2)\theta_n \epsilon_M$. These considerations perhaps explain the lack of orthogonality to working precision noticed by DeGroat.

References


<table>
<thead>
<tr>
<th>#</th>
<th>Author/s</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>892</td>
<td>E.G. Kalins, Willard Miller, Jr. and Sanchita Mukherjee</td>
<td>Models of q-algebra representations: the group of plane motions</td>
</tr>
<tr>
<td>893</td>
<td>T.R. Hoffend Jr. and R.K. Kaul</td>
<td>Relativistic theory of superpotentials for a nonhomogeneous, spatially isotropic medium</td>
</tr>
<tr>
<td>894</td>
<td>Reinhold von Scherwin</td>
<td>Two metal deposition on a microdisk electrode</td>
</tr>
<tr>
<td>895</td>
<td>Vladimir I. Oliker and Nina N. Ural'tseva</td>
<td>Evolution of nonparametric surfaces with speed depending on curvature, III. Some remarks on mean curvature and anisotropic flows</td>
</tr>
<tr>
<td>896</td>
<td>Wayne Barrett, Charles R. Johnson, Raphael Loewy and Tamir Shalom</td>
<td>Rank incrementation via diagonal perturbations</td>
</tr>
<tr>
<td>898</td>
<td>Mingxiang Chen, Xu-Yan Chen and Jack K. Hale</td>
<td>Structural stability for time-periodic one-dimensional parabolic equations</td>
</tr>
<tr>
<td>899</td>
<td>Hong-Ming Yin</td>
<td>Global solutions of Maxwell’s equations in an electromagnetic field with the temperature-dependent electrical conductivity</td>
</tr>
<tr>
<td>900</td>
<td>Robert Grone, Russell Merris and William Watkins</td>
<td>Laplacian unimodular equivalence of graphs</td>
</tr>
<tr>
<td>901</td>
<td>Miroslav Fiedler</td>
<td>Structure-ranks of matrices</td>
</tr>
<tr>
<td>902</td>
<td>Miroslav Fiedler</td>
<td>An estimate for the nonstochastic eigenvalues of doubly stochastic matrices</td>
</tr>
<tr>
<td>903</td>
<td>Miroslav Fiedler</td>
<td>Remarks on eigenvalues of Hankel matrices</td>
</tr>
<tr>
<td>904</td>
<td>Charles R. Johnson, D.D. Olesky, Michael Tsatsomeros and P. van den Driessche</td>
<td>Spectra with positive elementary symmetric function</td>
</tr>
<tr>
<td>905</td>
<td>Pierre-Alain Gremaud</td>
<td>Thermal contraction as a free boundary problem</td>
</tr>
<tr>
<td>906</td>
<td>K.L. Cooke, Janos Turi and Gregg Turner</td>
<td>Stabilization of hybrid systems in the presence of feedback delays</td>
</tr>
<tr>
<td>907</td>
<td>Robert P. Gilbert and Yongzhi Xu</td>
<td>A numerical transmutation approach for underwater sound propagation</td>
</tr>
<tr>
<td>908</td>
<td>LeRoy B. Beasley, Richard A. Brualdi and Bryan L. Shader</td>
<td>Combinatorial orthogonality</td>
</tr>
<tr>
<td>909</td>
<td>Richard A. Brualdi and Bryan L. Shader</td>
<td>Strong hall matrices</td>
</tr>
<tr>
<td>910</td>
<td>Håkan Wennerström and David M. Anderson</td>
<td>Difference versus Gaussian curvature energies; monolayer versus bilayer curvature energies applications to vesicle stability</td>
</tr>
<tr>
<td>911</td>
<td>Shmuel Friedland</td>
<td>Eigenvalues of almost skew symmetric matrices and tournament matrices</td>
</tr>
<tr>
<td>912</td>
<td>Avner Friedman, Bei Hu and J.L. Velazquez</td>
<td>A Free Boundary Problem Modeling Loop Dislocations in Crystals</td>
</tr>
<tr>
<td>913</td>
<td>Ezio Venturino</td>
<td>The Influence of Diseases on Lotka-Volterra Systems</td>
</tr>
<tr>
<td>914</td>
<td>Steve Kirkland and Bryan L. Shader</td>
<td>On Multipartite Tournament Matrices with Constant Team Size</td>
</tr>
<tr>
<td>915</td>
<td>Richard A. Brualdi and Jennifer J.Q. Massey</td>
<td>More on Structure-Ranks of Matrices</td>
</tr>
<tr>
<td>916</td>
<td>Douglas B. Meade</td>
<td>Qualitative Analysis of an Epidemic Model with Directed Dispersion</td>
</tr>
<tr>
<td>917</td>
<td>Kazuo Murota</td>
<td>Mixed Matrices Irreducibility and Decomposition</td>
</tr>
<tr>
<td>918</td>
<td>Richard A. Brualdi and Jennifer J.Q. Massey</td>
<td>Some Applications of Elementary Linear Algebra in Combinations</td>
</tr>
<tr>
<td>919</td>
<td>Carl D. Meyer</td>
<td>Sensitivity of Markov Chains</td>
</tr>
<tr>
<td>920</td>
<td>Hong-Ming Yin</td>
<td>Weak and Classical Solutions of Some Nonlinear Volterra Integrodifferential Equations</td>
</tr>
<tr>
<td>921</td>
<td>B. Leinkuhler and A. Ruehli</td>
<td>Exploiting Symmetry and Regularity in Waveform Relaxation Convergence Estimation</td>
</tr>
<tr>
<td>922</td>
<td>Xinfu Chen and Charles M. Elliott</td>
<td>Asymptotics for a Parabolic Double Obstacle Problem</td>
</tr>
<tr>
<td>923</td>
<td>Yongzhi Xu and Yi Yan</td>
<td>An Approximate Boundary Integral Method for Acoustic Scattering in Shallow Oceans</td>
</tr>
<tr>
<td>924</td>
<td>Yongzhi Xu and Yi Yan</td>
<td>Source Localization Processing in Perturbed Waveguides</td>
</tr>
<tr>
<td>925</td>
<td>Kenneth L. Cooke and Janos Turi</td>
<td>Stability, Instability in Delay Equations Modeling Human Respiration</td>
</tr>
<tr>
<td>926</td>
<td>F. Bethuel, H. Brezis, B.D. Coleman and F. Hélein</td>
<td>Bifurcation Analysis of Minimizing Harmonic Maps Describing the Equilibrium of Nematic Phases Between Cylinders</td>
</tr>
<tr>
<td>927</td>
<td>Frank W. Elliott, Jr., Signed Random Measures: Stochastic Order and Kolmogorov Consistency Conditions</td>
<td></td>
</tr>
<tr>
<td>928</td>
<td>D.A. Gregory, S.J. Kirkland and B.L. Shader</td>
<td>Pick’s Inequality and Tournaments</td>
</tr>
<tr>
<td>929</td>
<td>J.W. Demmel, N.J. Higham and R.S. Schreiber</td>
<td>Block LU Factorization</td>
</tr>
<tr>
<td>930</td>
<td>Victor A. Galaktionov and Juan L. Vazquez</td>
<td>Regional Blow-Up in a Semilinear Heat Equation with Convergence to a Hamilton-Jacobi Equation</td>
</tr>
<tr>
<td>931</td>
<td>Bryan L. Shader</td>
<td>Convertible, Nearly Decomposable and Nearly Reducible Matrices</td>
</tr>
<tr>
<td>932</td>
<td>Dianne P. O’Leary</td>
<td>Iterative Methods for Finding the Stationary Vector for Markov Chains</td>
</tr>
<tr>
<td>933</td>
<td>Nicholas J. Higham</td>
<td>Perturbation theory and backward error for $AX - XB = C$</td>
</tr>
<tr>
<td>934</td>
<td>Z. Strakos and A. Greenbaum</td>
<td>Open questions in the convergence analysis of the lanczos process for the real symmetric eigenvalue problem</td>
</tr>
<tr>
<td>935</td>
<td>Zhaojun Bai</td>
<td>Error analysis of the lanczos algorithm for the nonsymmetric eigenvalue problem</td>
</tr>
<tr>
<td>936</td>
<td>Pierre-Alain Gremaud</td>
<td>On an elliptic-parabolic problem related to phase transitions in shape memory alloys</td>
</tr>
<tr>
<td>937</td>
<td>Bojan Mohar and Neil Robertson</td>
<td>Disjoint essential circuits in toroidal maps</td>
</tr>
</tbody>
</table>
Bojan Mohar and Svatopluk Poljak, Eigenvalues in combinatorial optimization
Richard A. Brualdi, Keith L. Chavey and Bryan L. Shader, Conditional sign-solvability
Roger Fosdick and Ying Zhang, The torsion problem for a nonconvex stored energy function
René Ferland and Gaston Giroux, An unbounded mean-field intensity model:
  Propagation of the convergence of the empirical laws and compactness of the fluctuations
Wei-Ming Ni and Izumi Takagi, Spike-layers in semilinear elliptic singular Perturbation Problems
Henk A. Van der Vorst and Gerard G.L. Sleijpen, The effect of incomplete decomposition preconditioning
  on the convergence of conjugate gradients
S.P. Hastings and L.A. Peletier, On the decay of turbulent bursts
Apostolos Hadjidimos and Robert J. Plemmons, Analysis of p-cyclic iterations for Markov chains
ÁBjörck, H. Park and L. Eldén, Accurate downdating of least squares solutions
E.G. Kalnins, Willard Miller, Jr. and G.C. Williams, Recent advances in the use of separation of
  variables methods in general relativity
G.W. Stewart, On the perturbation of LU, Cholesky and QR factorizations
G.W. Stewart, Gaussian elimination, perturbation theory and Markov chains
G.W. Stewart, On a new way of solving the linear equations that arise in the method of least squares
G.W. Stewart, On the early history of the singular value decomposition
G.W. Stewart, On the perturbation of Markov chains with nearly transient states
Umberto Mosco, Composite media and asymptotic dirichlet forms
Walter F. Mascarenhas, The structure of the eigenvectors of sparse matrices
Walter F. Mascarenhas, A note on Jacobi being more accurate than QR
Raymond H. Chan, James G. Nagy and Robert J. Plemmons, FFT-based preconditioners for
  Toeplitz-Block least squares problems
Zhaojun Bai, The CSD, GSVD, their applications and computations
D.A. Gregory, S.J. Kirkland and N.J. Pullman, A bound on the exponent of a primitive matrix using
  Boolean rank
Richard A. Brualdi, Shmuel Friedland and Alex Pothen, Sparse bases, elementary vectors and nonzero
  minors of compound matrices
J.W. Demmel, Open problems in numerical linear algebra
James W. Demmel and William Gragg, On computing accurate singular values and eigenvalues of acyclic
  matrices
James W. Demmel, The inherent inaccuracy of implicit tridiagonal QR
J.J.L. Velázquez, Estimates on the \((N-1)\)-dimensional Hausdorff measure of the blow-up set
  for a semilinear heat equation
David C. Dobson, Optimal design of periodic antireflective structures for the Helmholtz equation
C.J. van Duijn and Joseph D. Fehribach, Analysis of planar model for the molten carbonate fuel cell
Yongzhi Xu, T. Craig Poling and Trent Brundage, Source localization in a waveguide with unknown
  large inclusions
J.J.L. Velázquez, Higher dimensional blow up for semilinear parabolic equations
E.G. Kalnins and Willard Miller, Jr., Separable coordinates, integrability and the Niven equations
John M. Chadam and Hong-Ming Yin, A diffusion equation with localized chemical reactions
A. Greenbaum and L. Gurvits, Max-min properties of matrix factor norms
Bei Hu, A free boundary problem arising in smoulder combustion
C.M. Elliott and A.M. Stuart, The global dynamics of discrete semilinear parabolic equations
Avner Friedman and Jianhua Zhang, Swelling of a rubber ball in the presence of good solvent
Avner Friedman and Juan J.L. Velázquez, A time-dependence free boundary problem modeling
  the visual image in electrophotography
Richard A. Brualdi, Hyung Chan Jung and William T. Trotter, Jr., On the poset of all posets on
  \(n\) elements
Ricardo D. Fierro and James R. Bunch, Multicollinearity and total least squares
Adam W. Bojanczyk, James G. Nagy and Robert J. Plemmons, Row householder transformations for
  rank-k Cholesky inverse modifications
Chaocheng Huang, An age-dependent population model with nonlinear diffusion in \(R^n\)
Emad Fatemi and Faroukh Odeh, Upwind finite difference solution of Boltzmann equation applied to
  electron transport in semiconductor devices
Esmond G. Ng and Barry W. Peyton, A tight and explicit representation of \(Q\) in sparse \(QR\)
  factorization
Robert J. Plemmons, A proposal for \(FFT\)-based fast recursive least-squares
Anne Greenbaum and Zdenek Strakos, Matrices that generate the same Krylov residual spaces
Alan Edelman and G.W. Stewart, Scaling for orthogonality
G.W. Stewart, Note on a generalized Sylvester equation
G.W. Stewart, Updating URV decompositions in parallel