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 algorthms 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 $X_k^{\mathrm{T}} X_k$ from the identity as a measure of the loss of orthogonality, then typically

$$\|I - X_k^{\mathrm{T}} X_k\|_{\mathrm{F}} \leq k\theta_n \epsilon_{\mathrm{M}},$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, ϵ_{M} is the rounding unit for the arithmetic in question, and θ_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 $I - X_i^{\mathrm{T}} X_i$ 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

$$\hat{X} = XQ,$$

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

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

where the diagonals of E are zero. Write

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

where

$$\hat{D} + \hat{E} = Q^{\mathrm{T}}EQ \tag{1}$$

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

$$\hat{S} = (I + \hat{D})^{-1} \tag{2}$$

and

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

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

$$\tilde{E} = \hat{S}^{\frac{1}{2}}\hat{E}\hat{S}^{\frac{1}{2}}. \tag{3}$$

The above equations define a recurrence for E, $\tilde{E}$, 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* $\mathrm{diag}(D)$ *denote the vector consisting of the diagonal element of D. Then for all sufficiently small E, the optimal scaling matrix S satisfies*

$$\hat{A} \circ \hat{A}\, \mathrm{diag}(S) = \mathrm{diag}(I + D), \tag{4}$$

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 $\|S^{\frac{1}{2}}\hat{E}S^{\frac{1}{2}}\|_{\mathrm{F}}^2$ 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 $\lim_{\hat{E}\to 0} \hat{A} \circ \hat{A} = (I + D)^2$. Consequently,

$$\lim_{\hat{E}\to 0} \mathrm{diag}(S) = \lim_{\hat{E}\to 0} (\hat{A} \circ \hat{A})^{-1}\mathrm{diag}(I + D) = \mathrm{diag}[(I + D)^{-1}] = \mathrm{diag}(\hat{S}) > 0. \tag{5}$$

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

$$\|E\|_{\mathrm{F}}^2 = \|\hat{D}\|_{\mathrm{F}}^2 + \|\hat{E}\|_{\mathrm{F}}^2. \tag{6}$$

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

$$\frac{\hat{e}_{ij}^2}{(1+\hat{d}_i)(1+\hat{d}_j)} \le \frac{\hat{e}_{ij}^2}{(1-\|\hat{D}\|_{\mathrm{F}})^2}.$$

Here $\hat{d}_i$ is the ith element of $\hat{D}$, and we assume that $\|\hat{D}\|_{\mathrm{F}} < 1$. Hence

$$\|\tilde{E}\|_{\mathrm{F}}^2 \le \frac{\|\hat{E}\|_{\mathrm{F}}^2}{(1-\|\hat{D}\|_{\mathrm{F}})^2}. \tag{7}$$

Setting

$$\epsilon = \|E\|_{\mathrm{F}} \quad \text{and} \quad \hat{\delta} = \|\hat{D}\|_{\mathrm{F}},$$

we have from (6) and (7)

$$\|\tilde{E}\|_{\mathrm{F}}^2 \le \tilde{\epsilon}^2 \equiv \frac{\epsilon^2 - \hat{\delta}^2}{(1-\hat{\delta})^2}. \tag{8}$$

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

$$\hat{\delta} = \gamma\epsilon, \qquad 0 \le \gamma \le 1.$$

In this notation the equality in (8) becomes

$$\tilde{\epsilon}^2 = \epsilon^2 \left[\frac{1-\gamma^2}{(1-\gamma\epsilon)^2}\right] \equiv \epsilon^2\varphi(\gamma). \tag{9}$$

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 ϵ 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*

$$\mathbf{E}\left(\sum_{i=1}^{n}(q_i^{\mathrm{T}} E q_i)^2\right) = \frac{1}{n+2}[\mathrm{trace}(E)^2 + 2\|E\|_{\mathrm{F}}^2],$$

where $\mathbf{E}$ 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 χ_n^2 distribution. Thus using standard results on the moments of the normal and χ^2 distributions, we have

$$\mathbf{E}v_i^4 = \frac{\mathbf{E}u_i^4}{\mathbf{E}r^4} = \frac{3}{n(n+2)}$$

and

$$\mathbf{E}(v_i^2 v_j^2) = \frac{\mathbf{E}(u_i^2 u_j^2)}{\mathbf{E}r^4} = \frac{1}{n(n+2)}, \qquad i \neq j.$$

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

$$E = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

For this case the result follows easily on expanding $\sum_{i=1}^{n}(q_i^{\mathrm{T}} E q_i)^2$ and using the above formulas to take expectations (recall that $\mathrm{trace}(E)^2 = \|E\|_{\mathrm{F}}^2 + \sum_{i\neq j}\lambda_i\lambda_j$).

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, δ is of the same order as ϵ, 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 ϵ is small enough so that the denominator in $\varphi(\gamma)$ can be ignored, an iteration will reduce ϵ^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 $\tilde{E}$. The Frobenius norm of this error will be proportional to the rounding unit ϵ_{M}, say $\theta_n\epsilon_{\mathrm{M}}$. Thus the recurrence (9) must be rewritten in the form

$$\tilde{\epsilon} = \varphi(\gamma)^{\frac{1}{2}}\epsilon + \theta_n\epsilon_{\mathrm{M}}.$$

If we assume that γ is constant, then this recurrence has the fixed point

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

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

References

[1] R. D. DeGroat and R. A. Roberts. Efficient numerically stabilized rank-one eigenstructure updating. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38:301–316, 1990. Cited in [2].

[2] M. Moonen, P. Van Dooren, and J. Vandewalle. A note on "efficient numerically stabilized rank-one eigenstructure updating. *IEEE Transactions on Signal Processing*, 39:1911–1913, 1991. Reply by DeGroat, pp. 1913–1914.

Recent IMA Preprints

#	Author/s	Title
892	**E.G. Kalnins, Willard Miller, Jr. and Sanchita Mukherjee**, Models of q-algebra representations: the group of plane motions	
893	**T.R. Hoffend Jr. and R.K. Kaul**, Relativistic theory of superpotentials for a nonhomogeneous, spatially isotropic medium	
894	**Reinhold von Schwerin**, Two metal deposition on a microdisk electrode	
895	**Vladimir I. Oliker and Nina N. Uraltseva**, Evolution of nonparametric surfaces with speed depending on curvature, III. Some remarks on mean curvature and anisotropic flows	
896	**Wayne Barrett, Charles R. Johnson, Raphael Loewy and Tamir Shalom**, Rank incrementation via diagonal perturbations	
898	**Mingxiang Chen, Xu-Yan Chen and Jack K. Hale**, Structural stability for time-periodic one-dimensional parabolic equations	
899	**Hong-Ming Yin**, Global solutions of Maxwell's equations in an electromagnetic field with the temperature-dependent electrical conductivity	
900	**Robert Grone, Russell Merris and William Watkins**, Laplacian unimodular equivalence of graphs	
901	**Miroslav Fiedler**, Structure-ranks of matrices	
902	**Miroslav Fiedler**, An estimate for the nonstochastic eigenvalues of doubly stochastic matrices	
903	**Miroslav Fiedler**, Remarks on eigenvalues of Hankel matrices	
904	**Charles R. Johnson, D.D. Olesky, Michael Tsatsomeros and P. van den Driessche**, Spectra with positive elementary symmetric functions	
905	**Pierre-Alain Gremaud**, Thermal contraction as a free boundary problem	
906	**K.L. Cooke, Janos Turi and Gregg Turner**, Stabilization of hybrid systems in the presence of feedback delays	
907	**Robert P. Gilbert and Yongzhi Xu**, A numerical transmutation approach for underwater sound propagation	
908	**LeRoy B. Beasley, Richard A. Brualdi and Bryan L. Shader**, Combinatorial orthogonality	
909	**Richard A. Brualdi and Bryan L. Shader**, Strong hall matrices	
910	**Håkan Wennerström and David M. Anderson**, Difference versus Gaussian curvature energies; monolayer versus bilayer curvature energies applications to vesicle stability	
911	**Shmuel Friedland**, Eigenvalues of almost skew symmetric matrices and tournament matrices	
912	**Avner Friedman, Bei Hu and J.L. Velazquez**, A Free Boundary Problem Modeling Loop Dislocations in Crystals	
913	**Ezio Venturino**, The Influence of Diseases on Lotka-Volterra Systems	
914	**Steve Kirkland and Bryan L. Shader**, On Multipartite Tournament Matrices with Constant Team Size	
915	**Richard A. Brualdi and Jennifer J.Q. Massey**, More on Structure-Ranks of Matrices	
916	**Douglas B. Meade**, Qualitative Analysis of an Epidemic Model with Directed Dispersion	
917	**Kazuo Murota**, Mixed Matrices Irreducibility and Decomposition	
918	**Richard A. Brualdi and Jennifer J.Q. Massey**, Some Applications of Elementary Linear Algebra in Combinations	
919	**Carl D. Meyer**, Sensitivity of Markov Chains	
920	**Hong-Ming Yin**, Weak and Classical Solutions of Some Nonlinear Volterra Integrodifferential Equations	
921	**B. Leinkuhler and A. Ruehli**, Exploiting Symmetry and Regularity in Waveform Relaxation Convergence Estimation	
922	**Xinfu Chen and Charles M. Elliott**, Asymptotics for a Parabolic Double Obstacle Problem	
923	**Yongzhi Xu and Yi Yan**, An Approximate Boundary Integral Method for Acoustic Scattering in Shallow Oceans	
924	**Yongzhi Xu and Yi Yan**, Source Localization Processing in Perturbed Waveguides	
925	**Kenneth L. Cooke and Janos Turi**, Stability, Instability in Delay Equations Modeling Human Respiration	
926	**F. Bethuel, H. Brezis, B.D. Coleman and F. Hélein**, Bifurcation Analysis of Minimizing Harmonic Maps Describing the Equilibrium of Nematic Phases Between Cylinders	
927	**Frank W. Elliott, Jr.**, Signed Random Measures: Stochastic Order and Kolmogorov Consistency Conditions	
928	**D.A. Gregory, S.J. Kirkland and B.L. Shader**, Pick's Inequality and Tournaments	
929	**J.W. Demmel, N.J. Higham and R.S. Schreiber**, Block *LU* Factorization	
930	**Victor A. Galaktionov and Juan L. Vazquez**, Regional Blow-Up in a Semilinear Heat Equation with Convergence to a Hamilton-Jacobi Equation	
931	**Bryan L. Shader**, Convertible, Nearly Decomposable and Nearly Reducible Matrices	
932	**Dianne P. O'Leary**, Iterative Methods for Finding the Stationary Vector for Markov Chains	
933	**Nicholas J. Higham**, Perturbation theory and backward error for $AX - XB = C$	
934	**Z. Strakos and A. Greenbaum**, Open questions in the convergence analysis of the lanczos process for the real symmetric eigenvalue problem	
935	**Zhaojun Bai**, Error analysis of the lanczos algorithm for the nonsymmetric eigenvalue problem	
936	**Pierre-Alain Gremaud**, On an elliptic-parabolic problem related to phase transitions in shape memory alloys	
937	**Bojan Mohar and Neil Robertson**, Disjoint essential circuits in toroidal maps	

939 **Bojan Mohar and Svatopluk Poljak** Eigenvalues in combinatorial optimization
940 **Richard A. Brualdi, Keith L. Chavey and Bryan L. Shader**, Conditional sign-solvability
941 **Roger Fosdick and Ying Zhang**, The torsion problem for a nonconvex stored energy function
942 **René Ferland and Gaston Giroux**, An unbounded mean-field intensity model: Propagation of the convergence of the empirical laws and compactness of the fluctuations
943 **Wei-Ming Ni and Izumi Takagi**, Spike-layers in semilinear elliptic singular Perturbation Problems
944 **Henk A. Van der Vorst and Gerard G.L. Sleijpen**, The effect of incomplete decomposition preconditioning on the convergence of conjugate gradients
945 **S.P. Hastings and L.A. Peletier**, On the decay of turbulent bursts
946 **Apostolos Hadjidimos and Robert J. Plemmons**, Analysis of p-cyclic iterations for Markov chains
947 **ÅBjörck, H. Park and L. Eldén**, Accurate downdating of least squares solutions
948 **E.G. Kalnins, Willard Miller, Jr. and G.C. Williams**, Recent advances in the use of separation of variables methods in general relativity
949 **G.W. Stewart**, On the perturbation of LU, Cholesky and QR factorizations
950 **G.W. Stewart**, Gaussian elimination, perturbation theory and Markov chains
951 **G.W. Stewart**, On a new way of solving the linear equations that arise in the method of least squares
952 **G.W. Stewart**, On the early history of the singular value decomposition
953 **G.W. STewart**, On the perturbation of Markov chains with nearly transient states
954 **Umberto Mosco**, Composite media and asymptotic dirichlet forms
955 **Walter F. Mascarenhas**, The structure of the eigenvectors of sparse matrices
956 **Walter F. Mascarenhas**, A note on Jacobi being more accurate than QR
957 **Raymond H. Chan, James G. Nagy and Robert J. Plemmons**, FFT-based preconditioners for Toeplitz-Block least squares problems
958 **Zhaojun Bai**, The CSD, GSVD, their applications and computations
959 **D.A. Gregory, S.J. Kirkland and N.J. Pullman**, A bound on the exponent of a primitive matrix using Boolean rank
960 **Richard A. Brualdi, Shmuel Friedland and Alex Pothen**, Sparse bases, elementary vectors and nonzero minors of compound matrices
961 **J.W. Demmel**, Open problems in numerical linear algebra
962 **James W. Demmel and William Gragg**, On computing accurate singular values and eigenvalues of acyclic matrices
963 **James W. Demmel**, The inherent inaccuracy of implicit tridiagonal QR
964 **J.J.L. Velázquez**, Estimates on the $(N-1)$-dimensional Hausdorff measure of the blow-up set for a semilinear heat equation
965 **David C. Dobson**, Optimal design of periodic antireflective structures for the Helmholtz equation
966 **C.J. van Duijn and Joseph D. Fehribach**, Analysis of planar model for the molten carbonate fuel cell
967 **Yongzhi Xu, T. Craig Poling and Trent Brundage**, Source localization in a waveguide with unknown large inclusions
968 **J.J.L. Velázquez**, Higher dimensional blow up for semilinear parabolic equations
969 **E.G. Kalnins and Willard Miller, Jr.**, Separable coordinates, integrability and the Niven equations
970 **John M. Chadam and Hong-Ming Yin**, A diffusion equation with localized chemical reactions
971 **A. Greenbaum and L. Gurvits**, Max-min properties of matrix factor norms
972 **Bei Hu**, A free boundary problem arising in smoulder combustion
973 **C.M. Elliott and A.M. Stuart**, The global dynamics of discrete semilinear parabolic equations
974 **Avner Friedman and Jianhua Zhang**, Swelling of a rubber ball in the presence of good solvent
975 **Avner Friedman and Juan J.L. Velázquez**, A time-dependence free boundary problem modeling the visual image in electrophotography
976 **Richard A. Brualdi, Hyung Chan Jung and William T. Trotter, Jr.**, On the poset of all posets on n elements
977 **Ricardo D. Fierro and James R. Bunch**, Multicollinearity and total least squares
978 **Adam W. Bojanczyk, James G. Nagy and Robert J. Plemmons**, Row householder transformations for rank-k Cholesky inverse modifications
979 **Chaocheng Huang**, An age-dependent population model with nonlinear diffusion in R^n
980 **Emad Fatemi and Faroukh Odeh**, Upwind finite difference solution of Boltzmann equation applied to electron transport in semiconductor devices
981 **Esmond G. Ng and Barry W. Peyton**, A tight and explicit representation of Q in sparse QR factorization
982 **Robert J. Plemmons**, A proposal for FFT-based fast recursive least-squares
983 **Anne Greenbaum and Zdenek Strakos**, Matrices that generate the same Krylov residual spaces
984 **Alan Edelman and G.W. Stewart**, Scaling for orthogonality
985 **G.W. Stewart**, Note on a generalized sylvester equation
986 **G.W. Stewart**, Updating URV decompositions in parallel