

Lawrence Berkeley National Laboratory

Recent Work

Title

Scaling for Orthogonality

Permalink

<https://escholarship.org/uc/item/5q8135gw>

Authors

Edelman, A.
Stewart, G.W.

Publication Date

1992-04-01



Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

Physics Division

Mathematics Department

To be submitted for publication

Scaling for Orthogonality

A. Edelman and G.W. Stewart

April 1992

For Reference

Not to be taken from this room



DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. Neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California and shall not be used for advertising or product endorsement purposes.

Lawrence Berkeley Laboratory is an equal opportunity employer.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

SCALING FOR ORTHOGONALITY

Alan Edelman¹

Lawrence Berkeley Laboratory
and

Department of Mathematics
University of California
Berkeley, CA 94720

G.W. Stewart²

Department of Computer Sciences
University of Maryland
College Park, MD 20742

April 1992

¹Supported in part by the Applied Mathematical Sciences Subprogram of the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

²This 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 EDELMAN¹
G. W. STEWART²

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, 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 $I - X_i^T X_i$ to the diagonal. We also show that normalizing

¹Department 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.

²Department of Computer Science and Institute for Mathematics and Its Applications, 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.

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 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

$$\hat{D} + \hat{E} = Q^T E Q \quad (1)$$

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

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

and

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

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

$$\tilde{E} = \hat{S}^{\frac{1}{2}} \hat{E} \hat{S}^{\frac{1}{2}}. \quad (3)$$

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

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

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

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

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

$$\|E\|_{\mathbb{F}}^2 = \|\hat{D}\|_{\mathbb{F}}^2 + \|\hat{E}\|_{\mathbb{F}}^2. \quad (6)$$

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

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

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

$$\|\tilde{E}\|_{\mathbb{F}}^2 \leq \frac{\|\hat{E}\|_{\mathbb{F}}^2}{(1 - \|\hat{D}\|_{\mathbb{F}})^2}. \quad (7)$$

Setting

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

we have from (6) and (7)

$$\|\tilde{E}\|_{\mathbb{F}}^2 \leq \tilde{\epsilon}^2 \equiv \frac{\epsilon^2 - \delta^2}{(1 - \delta)^2}. \quad (8)$$

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

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

In this notation the equality in (8) becomes

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

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

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

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

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

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

$$E = \text{diag}(\lambda_1, \dots, \lambda_n).$$

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, 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 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 $\theta_n \epsilon_M$. Thus the recurrence (9) must be rewritten in the form

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

If we assume that γ 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

- [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.

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
TECHNICAL INFORMATION DEPARTMENT
BERKELEY, CALIFORNIA 94720