Wrocław University of Technology
Institute of Mathematics and Computer Science

Numerical experiments with Higham's scaled method for polar decomposition

Andrzej Kiełbasiński, Paweł Zieliński, Krystyna Ziętak

Report I18/2006/P-013

Wrocław, 2006, May 6

Numerical experiments with Higham's scaled method for polar decomposition

Andrzej Kiełbasiński[1], Paweł Zieliński[2] and Krystyna Ziętak[2]

[1] *University of Warsaw, Institute of Applied Mathematics and Mechanics, 02-097 Warsaw, Poland*
[2] *Wrocław University of Technology, Institute of Mathematics and Computer Science, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland*
E-mail: pawel.zielinski@pwr.wroc.pl krystyna.zietak@pwr.wroc.pl

In the paper we present numerical experiments with Higham's scaled method for computing the polar decomposition of a matrix. We present also a further developed theory explaining the phenomena observed in experiments. Both, the theory and tests show how the numerical properties of algorithms for inversion of a matrix influence the accuracy of the computed polar factorization. We show, in particular, that for standard inversion (via `GEPP`-factorization) the computed polar factors can be unacceptable. Some other problems of practical scaling and switching criteria are discussed and experimentally investigated.

Keywords: roundoff error analysis, polar decomposition of a matrix, Higham's method, numerical matrix inversion
AMS subject classification: 65G50, 65F30

1 Introduction

In the paper we deal with the polar decomposition of a complex nonsingular matrix $A \in \mathcal{C}^{n\times n}$

$$A = UH, \quad U - \text{unitary}, \quad H \in \mathcal{HPD}, \tag{1.1}$$

where $\mathcal{HPD}$ is the class of Hermitian positive-definite matrices. If A is real then U is the orthogonal factor for A.

The factorization (1.1) can be computed from the singular values decomposition of A. However, the iterative methods are alternative ways to compute (1.1) (see for example [3, 5, 6, 7, 9]). From among iterative methods for computing the unitary factor U, the scaled Higham's method [6] is distinguished because of its efficiency and good behaviour, even for ill-conditioned matrices A. This phenomenon, confirmed by extensive numerical experiments, is the subject of our interest in [12].

In Higham's scaled method [6], denoted by HS, one constructs a sequence of matrices

$$X_{k+1} = \frac{1}{2}\Big(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H}\Big), \quad X_0 = A, \quad \gamma_k > 0, \tag{1.2}$$

convergent to U (U is the common unitary factor of all X_k). There are several rules of the choice of scaling parameters γ_k which increase the speed of convergence (see [5, 6, 11]).

If matrix A is badly conditioned then the condition numbers of X_k decrease very quickly for $k = 0, 1, \ldots$ (see [6, 11]). This advantage is burdened by a fear that the roundoff errors produced by an algorithm for computing X_k^{-1} can cause inaccuracy of the computed U. Such a loss of the accuracy in the computed X_{k+1} from (1.2) could appear especially in the few initial iterations when the condition numbers of X_k are large. In all cases when the numerical HS algorithm converges a good unitarity of the computed $\widetilde{U} = X_l$:

$$||\widetilde{U}^H\widetilde{U} - I||_2 \leqslant \varepsilon_0. \tag{1.3}$$

is achieved (all ε_s in this paper are of the size of ν, the computing precision). The problem is whether $\widetilde{U}$ is an acceptable unitary factor for A, that means, whether the following conditions

$$||\widetilde{U}\hat{H} - A||_2 \leqslant \varepsilon_1||A||_2, \quad \hat{H} \stackrel{\text{df}}{=} \frac{1}{2}(\widetilde{U}^H A + A^H\widetilde{U}) \in \mathcal{HPD} \tag{1.4}$$

hold.

Theoretical analysis of the numerical HS-process was presented in [12]. Our initial intention was to present in this paper only the numerical experiments illustrating these theoretical results and experiments yielding some information on problems, which we were unable to solve theoretically. But soon it turned out that we need a further extension of the theory [12] to explain the phenomena we observe in our experiments. We present this further developed theory in sections 2 and 4. In section 3 we present our experimental tool: the HSTEST program. The experimental results (and relevant discussion) are divided in three groups of problems:

(i) the problem of the *quality of matrix inversion* in the numerical HS-process. In section 4 we present the theoretical background (an extension of the theory to the case of the matrix-inversion of worse quality) and the results of corresponding experiments.

(ii) the problem of *too small scaling parameters*. Our analysis in [12] indicates the possibility of considerable losses of the accuracy resulting

from this source. May be the experiments could answer whether such a danger really exists? In section 5 we present the corresponding experimental research.

(iii) the problem of the *switching criteria*. For the convenience of the theoretical research we assume in [12] that the sequence $\{\mathrm{cond}_2(X_k)\}_{k=0}^{l}$ is strictly decreasing. To put this assumption on a safe ground we introduced there two *switching criteria*: (1) the switch from $(1,\infty)$-scaling to the unscaled process, (2) the switch to the *last-step mode*. In section 6 we try to answer the question whether these "new criteria" have any practical advantage.

Section 7 contains final conclusions from both, the theoretical and the experimental research.

In the next sections `HS` means: *the numerical* `HS`*-process* (to distinguish from (1.2), where *the theoretical* `HS` is defined).

To simplify the reference to concrete formulae in [12] we add the $'1'$ before the section number. Thus, for example, ... see (13.9) means: ... see (3.9) in [12]. The reference to formula in appendices (A, B, C, D, E, F) in [12] are left unaltered. Thus ... see (D.7) means: ... see (D.7) in [12].

2 The theory of `HS`, the numerical Higham's method

In subsection 2.1 we recall relevant elements of our numerical analysis in [12], using essentially the same notation. The only major difference is the interchange of the role of the symbols: X_k and $\tilde{X}_k$. Contrary to [12] here $\tilde{X}_k$ means the computed iterate and X_k is the matrix defined by the conditions (2.5) below. (In general neither $\tilde{X}_k$ nor X_k *here* is identical with X_k in (1.2), though both these matrices would tend to X_k from (1.2) when $\nu \to 0$).

Subsection 2.2 contains: theorem 2.1 (a stronger version of theorem 3.1 in [12]) and theorem 2.2, an essential supplement to theorem 2.1. Both theorems constitute our basic tool for explaining some phenomena observed in experiments.

We use both, the spectral, $||\cdot||_2$, and the Frobenius, $||\cdot||_F$, norms of matrices. An eventual transfer from one norm to another will be expressed by the function $p : \mathbb{C}^{n\times n} \to [n^{-1/2}, 1]$:

$$p(\mathbf{\Psi}) \stackrel{\mathrm{df}}{=} \begin{cases} 1 & \text{when } \mathbf{\Psi} = 0, \\ \frac{||\mathbf{\Psi}||_2}{||\mathbf{\Psi}||_F} & \text{otherwise.} \end{cases} \tag{2.1}$$

We are using this transfer-function explicitly only in a few points, where it seems to have some importance.

The next reserved notation is connected with the function $f : (0, \infty) \to [1, \infty)$:

$$f(t) \stackrel{\text{df}}{=} \frac{1}{2}(t + t^{-1}), \tag{2.2}$$

playing a special role in the description of the HS-process. These *reserved functions* "produce" a series of derivate symbols $(f_k, p_k, p_+, ...)$, the values of f or p on concrete arguments.

Another reserved notation is connected with the following presentation of the norm of the sum of matrices: $||B+D|| = ||B||+\theta||D||$, where $\theta = \theta(B, D)$ is the mean-value parameter for matrices B, D (defined uniquely when $D \neq 0$) satisfying the bounds: $\max\{-1, 1 - 2||B||/||D||\} \leqslant \theta(B, D) \leqslant 1$. We will introduce such "θ"-parameters $(\theta_\kappa, \theta', \theta_3, \ldots)$ not always defining explicitly the corresponding matrices B, D, but always assuming the following bounds on $b + \theta d$ (when $b \geqslant 0, d > 0$):

$$\max\{b - d, d - b\} \leqslant b + \theta d \leqslant b + d. \tag{2.3}$$

Any "θ" appearing in another context satisfies only the bound: $|\theta| \leqslant 1$.

We assume that the computations are performed in the floating-point arithmetic with the *precision* ν and that neither under-flow nor over-flow occurs.

The epsilons $(\varepsilon_0, \varepsilon_x, \ldots)$ are modest multiples of ν. Not all of them must be positive. We signal it writing, for example, $|\varepsilon'_k| \leqslant \varepsilon$. The only exception (see section 4) are "false epsilons": $\check{\varepsilon}_x, \check{\varepsilon}_k, \ldots$, the quantities which ought to be the *true epsilons* (and sometimes are) but (due to breaking in experiments of the basic assumption (2.5)) can be much larger than could be normally accepted as a modest multiple of ν. Usually these false epsilons satisfy $|\check{\varepsilon}| \ll 1$.

Let us formulate already now the following *general assumptions* (natural for a process with effective numerical matrix inversion):

$$n \leqslant 100, \qquad \hat{\varepsilon}\ \mathrm{cond}_2(A) < 1, \qquad \hat{\varepsilon} < \nu^{2/3} \leqslant 10^{-4} \tag{2.4}$$

for $\hat{\varepsilon}$ specified in (2.7), (2.5). Our analysis is hence valid also for the weakest contemporary arithmetic with $\nu \approx 10^{-6}$; but in our experiments we use the *standard-double arithmetic* with $\nu = \nu_d \approx 2.2 \times 10^{-16}$.

2.1 Main definitions and relations

Let us consider a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ and the sequence $\{\tilde{X}_k\}_{k=0}^{l}$ of matrices (1.2) computed in HS, $\tilde{X}_0 := A$.

Let γ_k be the chosen scaling parameter in HS and G_k the computed inverse of $\tilde{X}_k$. We assume that there exists a nonsingular matrix X_k, such that the following relations hold:

$$G_k = X_k^{-1} + \mathbf{\Delta}'_k,\ \tilde{X}_k = X_k + \mathbf{\Delta}_k,\ ||\mathbf{\Delta}'_k||_F \leqslant \varepsilon_g ||X_k^{-1}||_2,\ ||\mathbf{\Delta}_k||_F \leqslant \varepsilon_x ||X_k||_2. \tag{2.5}$$

This defines (not uniquely) X_k for $k < l$. Let us extend it to $k = l$, assuming $X_l \stackrel{\text{df}}{=} \tilde{X}_l$. $\{X_k\}$ and $\{\tilde{X}_k\}$ are neighbour sequences and all important properties of $\tilde{X}_k$ are close to those of X_k. It is convenient to describe the HS in terms of the sequence $\{X_k\}$, since this sequence imitates well the relation (1.2), see (2.6), (2.7).

The assignment-statements $G_k := \tilde{X}_k^{-1}, \tilde{X}_{k+1} := (\tilde{X}_k * \gamma_k + G_k^H/\gamma_k)/2$ and (2.5) imply the equality

$$X_{k+1} = Z_{k+1} + \mathbf{\Phi}_{k+1}, \quad Z_{k+1} \stackrel{\text{df}}{=} \frac{1}{2}\Big(\gamma_k X_k + \frac{1}{\gamma_k} X_k^{-H}\Big) \tag{2.6}$$

and the bound, compare (4.17), (4.18),

$$||\mathbf{\Phi}_{k+1}||_F \leqslant \hat{\varepsilon} f_k, \quad f_k \stackrel{\text{df}}{=} ||Z_{k+1}||_2, \quad \hat{\varepsilon} = 2\varepsilon_x + \varepsilon_g + 3\sqrt{n}\nu + 0(\nu^2). \tag{2.7}$$

Let us consider the SVD of X_k:

$$X_k = P_k \mathrm{diag}(\sigma_1^{(k)}, \ldots, \sigma_n^{(k)}) Q_k^H, \qquad P_k, Q_k - \text{unitary}, \tag{2.8}$$

and define d_k, the distance of X_k from the unitarity:

$$d_k \stackrel{\text{df}}{=} \max_i |\sigma_i^{(k)} - 1| = \max\{\sigma_{\max}^{(k)} - 1, 1 - \sigma_{\min}^{(k)}\}, \tag{2.9}$$

where

$$\sigma_{\max}^{(k)} \stackrel{\text{df}}{=} \max_i \sigma_i^{(k)}, \qquad \sigma_{\min}^{(k)} \stackrel{\text{df}}{=} \min_i \sigma_i^{(k)}. \tag{2.10}$$

The efficiency of HS depends on how quickly the "errors" $\{d_k\}_{k=1}^l$ decrease. The near-unitarity of the computed $\tilde{U} \stackrel{\text{df}}{=} X_l$ depends on the limiting accuracy:

$$d \stackrel{\text{df}}{=} \lim\ \sup\ d_k \tag{2.11}$$

of the conceptional infinite sequence $\{d_k\}_{k=0}^{\infty}$. The last iterate $\tilde{X}_l = X_l$, constructed in HS, should be the first one reaching the level $d_l \lesssim d$.

To describe the behaviour of the sequence $\{d_k\}$, let us define further quantities:

$$c_k \stackrel{\text{df}}{=} \mathrm{cond}_2(X_k) = \frac{\sigma_{\max}^{(k)}}{\sigma_{\min}^{(k)}}, \qquad \gamma_k^{(\text{opt})} \stackrel{\text{df}}{=} (\sigma_{\max}^{(k)} \sigma_{\min}^{(k)})^{-1/2}, \tag{2.12}$$

$$\rho_k \stackrel{\mathrm{df}}{=} \Big(\frac{\gamma_k}{\gamma_k^{(\mathrm{opt})}}\Big)^2, \qquad \tau_k \stackrel{\mathrm{df}}{=} \max\Big\{\rho_k, \frac{1}{\rho_k}\Big\}. \tag{2.13}$$

The parameters ρ_k, τ_k "measure the distance" of γ_k from $\gamma_k^{(\mathrm{opt})}$, the *optimal scaling parameter*, see [6]. From (2.6) and (2.8) it follows

$$Z_{k+1} = P_k \mathrm{diag}(\hat{\sigma}_1^{(k)}, \ldots, \hat{\sigma}_n^{(k)}) Q_k^H, \qquad \hat{\sigma}_i^{(k)} \stackrel{\mathrm{df}}{=} f(\sigma_i^{(k)}\gamma_k). \tag{2.14}$$

Hence the bounds (see (2.7), (2.12), (2.13)):

$$1 \leqslant \hat{\sigma}_i^{(k)} \leqslant f_k = f(\hat{\sigma}^{(k)}), \qquad \hat{\sigma}^{(k)} \stackrel{\mathrm{df}}{=} \sqrt{c_k \tau_k} \tag{2.15}$$

hold. This implies further bounds (see (2.6), (2.7), (2.9)):

$$1 - \hat{\varepsilon} f_k \leqslant \sigma_{\min}^{(k+1)}, \quad (1-\hat{\varepsilon}) f_k \leqslant \sigma_{\max}^{(k+1)} \leqslant (1+\hat{\varepsilon}) f_k, \tag{2.16}$$

$$(1-\hat{\varepsilon}) f_k - 1 \leqslant d_{k+1} \leqslant (1+\hat{\varepsilon}) f_k - 1, \tag{2.17}$$

$$\hat{\varepsilon} f_k < 1 \Longrightarrow c_{k+1} \leqslant \frac{f_k(1+\hat{\varepsilon})}{1-\hat{\varepsilon} f_k}. \tag{2.18}$$

For a given matrix X_k (hence for fixed c_k and $\gamma_k^{(\mathrm{opt})}$) τ_k depends only on the choice of γ_k. When γ_k is close to $\gamma_k^{(\mathrm{opt})}$ then τ_k and f_k decrease, hence the bounds on d_{k+1}, c_{k+1} improve (and d_{k+1}, c_{k+1} tend to decrease). The best case ($\tau_k = 1$) is, when $\gamma_k = \gamma_k^{(\mathrm{opt})}$, what justifies the used terminology. The *minimal reasonable condition* on the choice of $\{\gamma_k\}$ (hence on $\{\tau_k\}$) can be formulated in terms of the sequence $\{\hat{\sigma}_k\}$, see (2.15): $\hat{\varepsilon} f(\hat{\sigma}_0) < 1$ and sequence $\{\hat{\sigma}^{(k)}\}_{k=0}^{l}$ decreasing and approaching 1. This yields in particular upper bounds (2.18) on $\{c_k\}$. But for *fast reduction of errors* $\{d_k\}_{k=1}^{l}$ and for good *near-unitarity* of $\tilde{U} = X_l$ some stronger upper bounds on $\{\tau_k\}$ are necessary. Computing of $\gamma_k^{(\mathrm{opt})}$ is expensive. *Practical scaling* guarantees, see (12.26) and (12.27):

$$\tau_k^{(1,\infty)} \leqslant \min\{\sqrt{n}, c_k + (c_k-1)^2\sqrt{n}\}(1+\varepsilon_\tau), \qquad \text{for } (1,\infty)-\text{scaling} \tag{2.19}$$

$$\tau_k^{(F)} \leqslant \min\{\sqrt{n-1}, c_k\}(1+\varepsilon_\tau), \qquad \text{for } (F)-\text{scaling}. \tag{2.20}$$

With the assumptions (2.4), (2.5) both techniques guarantee fast reduction of large d_k, c_k to the level, say: $d_k \leqslant 2, c_k \leqslant 3$. But only (2.20) guarantees further fast, *quadratic convergence* of $\{d_k\}$ to the acceptable *limiting accuracy level*: $d = \limsup d_k \lesssim \hat{\varepsilon}$, see appendix C in [12].

For $(1,\infty)$-scaling the situation is not clear. In [6] Higham suggests the switch to the *unscaled iterations* ($\gamma_k \equiv 1$) with a *switch criterion* corresponding roughly to the error level $d_k \lesssim 10^{-2}$. We do not know whether such a level will be always achieved with $(1,\infty)$-scaling. In [12] we consider hence another, *safer switch-criterion*, corresponding roughly to the error level $d_k \leqslant 2$ ($d_k \leqslant 1$ when $n \leqslant 50$). Unscaled iterations guarantee quadratic convergence of $\{d_k\}$ to the limiting accuracy level: $d \lesssim \hat{\varepsilon}$. Practical scaling will now mean: either (F)-scaling or $(1,\infty)$-scaling with a switch to unscaled iterations in a right moment to guarantee monotonic decrease of $\{\hat{\sigma}_k\}_{k=0}^{l}$.

In some experiments presented in sections 4 and 5 we modify the normal `HS`-process introducing (in a few initial steps only) *either* matrices G_k not satisfying (2.5) *or* scaling parameters γ_k retarding the convergence. But these modifications *neither* destroy the monotonic decrease of $\{\hat{\sigma}_k\}$ *nor* influence the final convergence.

Assuming that also in (F)-scaling the last step is unscaled (and using $X_l = \tilde{X}_l$) we obtain a better bound on d_l than $\hat{\varepsilon}$:

$$d_l \lesssim \varepsilon_l \stackrel{\mathrm{df}}{=} \frac{1}{2}p'(\varepsilon_x + \varepsilon_g + 2\sqrt{n}\nu), \qquad p' \stackrel{\mathrm{df}}{=} p(\vartheta X_l), \tag{2.21}$$

where the matrix ϑX_l is defined in (12.8). Hence (1.3) is valid with $\varepsilon_0 \approx p'(\varepsilon_x + \varepsilon_g + 2\sqrt{n}\nu)$. We achieve a good *near unitarity* of $\tilde{U} = \tilde{X}_l$!

Let the abbreviations: `AUF`, `APF` mean: *approximate unitary factor, approximate polar factors*, respectively. We turn now to the crucial problem (1.4):

$$\textit{how good is } \tilde{U} \textit{ as an } \texttt{AUF} \textit{ of } A? \tag{2.22}$$

Our way to answer this question is long. It consists in the following: replacing $\tilde{U}$ with the unitary factor U of $\tilde{U}$, answering a sequence of questions:

$$\textit{How good is } U \textit{ as an } \texttt{AUF} \textit{ of } X? \tag{2.23}$$

for $X = X_k$ ($k = l, l-1, \ldots, 0$) and for $X = A$. and returning to the question (2.22). For any $H \in \mathcal{HPD}$ the equality

$$UH = X + \mathbf{\Delta} \tag{2.24}$$

means that the matrices $\{U, H\}$ are `APF` of X with *accuracy*:

$$\vartheta(H) \stackrel{\mathrm{df}}{=} \frac{||\mathbf{\Delta}||_F}{||X||_2}, \qquad \mathbf{\Delta} \stackrel{\mathrm{df}}{=} UH - X. \tag{2.25}$$

The *accuracy* of U as an AUF of X corresponds to matrices $H \in \mathcal{HPD}$ minimizing the error $\vartheta(H)$:

$$\mathtt{acc}(U,X) \stackrel{\mathrm{df}}{=} \inf_{H\in\mathcal{HPD}} \frac{||UH-X||_F}{||X||_2}. \tag{2.26}$$

Hence U is AUF of X with the error $\mathtt{acc}(U,X)$, but this must not mean that in $\mathcal{HPD}$ exists such a H that the matrices $\{U,H\}$ are APF of X with the error (2.26). Namely, it is known, see [1, pp.214–215] that defining the following matrices (and a number):

$$B_{ux} \stackrel{\mathrm{df}}{=} U^H X, \quad H_{ux} \stackrel{\mathrm{df}}{=} \frac{1}{2}\Big(B_{ux}+B_{ux}^H\Big), \quad \delta_{ux} \stackrel{\mathrm{df}}{=} \frac{||B_{ux}-B_{ux}^H||_F}{2||X||_2} \tag{2.27}$$

we obtain the *excluding* alternative: **either** $H_{ux} \in \mathcal{HPD}$, then $\mathtt{acc}(U,X) = \vartheta(H_{ux}) = \delta_{ux}$ and H_{ux} is the unique minimizer of $\vartheta(H)$, **or** $H_{ux} \notin \mathcal{HPD}$, then $\mathtt{acc}(U,X) \geqslant \delta_{ux}$ and $\vartheta(H) > \delta_{ux}$ holds for any $H \in \mathcal{HPD}$.

Let us consider hence *only the case*: $H_{ux} \in \mathcal{HPD}$. U is AUF of X with the error δ_{ux}. U is the better AUF of X the smaller δ_{ux} is. U is a *good* AUF of X if δ_{ux} is a modest multiple of ν (the highest quality of numerical computation). The same terminology will be used for matrices $\{U,H_{ux}\}$ as APF of X.

Remark 2.1. We chose the Frobenius norm in (2.25), (2.26) since there exists constructive matrix-approximation theory in this norm, yielding the minimizer H_{ux}. If we choose, instead of (2.25), the definition:

$$\vartheta_2(H) \stackrel{\mathrm{df}}{=} \frac{||\mathbf{\Delta}||_2}{||X||_2}, \quad \mathbf{\Delta} \stackrel{\mathrm{df}}{=} UH-X \tag{2.28}$$

of the relative error, then the matrices $\{U,H_{ux}\}$ are APF of X with the error $p_{ux}\delta_{ux}$, where $p_{ux} \stackrel{\mathrm{df}}{=} p(B_{ux}-B_{ux}^H)$, see (2.1), (2.27). But in general H_{ux} is not a minimizer of $\vartheta_2(H)$ in $\mathcal{HPD}$.

Let us define for $k = 0,\ldots,l$ the following matrices and numbers:

$$B_k \stackrel{\mathrm{df}}{=} U^H X_k, \quad H_k \stackrel{\mathrm{df}}{=} \frac{1}{2}(B_k+B_k^H), \quad \delta_k \stackrel{\mathrm{df}}{=} \frac{||X_k-UH_k||_F}{||X_k||_2}. \tag{2.29}$$

For good numerical behaviour of the HS-process U should be a good AUF for all $X_k, 0 \leqslant k \leqslant l (H_k \in \mathcal{HPD}, \delta_k$ a modest multiple of $\nu)$. The same holds

for the neighbour-sequence $\{\tilde{X}_k\}_{k=0}^{l}$. Considering the polar decomposition of $\tilde{U}$ we find:

$$\tilde{U} = UH_u, \quad H_l = H_u \in \mathcal{HPD}, \quad \delta_l = 0. \tag{2.30}$$

This is a good start for `BIT`, the *backward-induction theorem* (see the next section). Assuming $H_{k+1} \in \mathcal{HPD}$ we give there (in terms of: $\delta_{k+1}, \rho_k, c_k, \hat{\varepsilon}$): an explicit formula for δ_k and a condition, sufficient for the positive definiteness of H_k. This opens a way to get some *a priori* idea about the relevant properties of the sequences $\{H_k\}, \{\delta_k\}$.

We must be prepared that U can be a worser `AUF` for X_k than for X_{k+1}: $\delta_k > \delta_{k+1}$, since the rounding errors in computing G_k and $\tilde{X}_{k+1}$ can partly spoil the information on $\tilde{X}_k$ transferred to $\tilde{X}_{k+1}$ (hence also to $\tilde{X}_l = \tilde{U}$). Our research in the next sections depends on indicating the "benign rounding errors", such that δ_k is at most slightly larger than δ_{k+1}, and on revealing "dangerous rounding errors", such that $\delta_k \gg \delta_{k+1}$ can succeed.

Let us close this section with an answer to the question (2.22), given in terms of the pair $\{\delta_0, H_0\}$. We present the following lemma, skipping a banal proof.

Lemma 2.1. Let us assume that (2.5) is satisfied for $k = 0$ and let us consider the computed `APF` $\{\tilde{U}, \tilde{H}\}$ of A:

$$\tilde{U} := \tilde{X}_l, \quad \tilde{B} := \tilde{U}^H * A, \quad \tilde{H} := (\tilde{B} + \tilde{B}^H)/2. \tag{2.31}$$

(i) If $H_0 \in \mathcal{HPD}$ and $\mathrm{cond}_2(A) * (\delta_0 + 2\varepsilon') < 1$ holds, $\varepsilon' \approx \varepsilon_x + \nu\sqrt{n}$, then $H_a \stackrel{\mathrm{df}}{=} (U^H A + A^H U)/2 \in \mathcal{HPD}$ and the bound $|\mathtt{acc}(U, A) - \delta_0| \leqslant \varepsilon'$ holds.

(ii) If $H_0 \in \mathcal{HPD}$ and $\mathrm{cond}_2(A) * (\hat{p}_0\delta_0 + \varepsilon_{\mathrm{I}}) < 1$ holds, see (2.32), then $\tilde{H} \in \mathcal{HPD}$ and the bound

$$\Big|\frac{||\tilde{U}\tilde{H} - A||_2}{||A||_2} - \hat{p}_0\delta_0\Big| \leqslant \varepsilon_{\mathrm{I}} \approx 2\varepsilon_x + \varepsilon_g + \nu m(\sqrt{n}), \hat{p}_0 \stackrel{\mathrm{df}}{=} p(UH_0 - X_0), \tag{2.32}$$

holds where $m(t)$ is a modest polynomial in t, depending on the way of computing $\tilde{B}$ in (2.31).

Remarks 2.2. Lemma 2.1 is valid also when δ_0 is not a modest multiple of ν. If G_0 is not satisfying (2.5) then the quantities $\varepsilon_x, \varepsilon_g$ in lemma 2.1 should be replaced with $\hat{\varphi}_0$, see theorem 4.1. Note that $\tilde{H} = \hat{H}$ holds, see (1.4), if arithmetic operations in (2.31) are performed exactly.

Conclusion 2.1. U (or $\tilde{U}$) is a good `AUF` of A iff $\{U, H_0\}$ are good `APF` of X_0 and the matrix A is sufficiently well conditioned. The same holds for $\{\tilde{U}, \tilde{H}\}$, the computed `APF` of A.

2.2 BIT4, the backward-induction theorem

Let U be the unitary polar factor of $\tilde{U} = \tilde{X}_l = X_l$ and let us define the quantities

$$\xi_k^* \stackrel{\text{df}}{=} ||\mathbf{\Psi}_{k+1}||_F, \quad \vartheta_k' \stackrel{\text{df}}{=} \xi_k^* f_k^{-1}, \quad \xi_k' \stackrel{\text{df}}{=} ||\mathbf{\Psi}_{k+1}||_2, \quad r_k \stackrel{\text{df}}{=} \frac{f_k}{f(||Y_k||_2)} \tag{2.33}$$

where (see (2.6), (2.7), (2.29))

$$\mathbf{\Psi}_{k+1} \stackrel{\text{df}}{=} UH_{k+1} - Z_{k+1}, \quad Y_k \stackrel{\text{df}}{=} X_k\gamma_k. \tag{2.34}$$

Theorem 2.1. (`BIT4`) If the relations

$$\xi_k' < 1, \quad H_{k+1} \in \mathcal{HPD} \tag{2.35}$$

are satisfied then there exist non-negative numbers: $\chi_k, \mu_k, \kappa_k, \lambda_k$ *either* all equal to zero *or* fulfilling the inequalities:

$$\mu_k < \chi_k \leqslant 1, \quad \kappa_k < 1, \quad \lambda_k < 1, \tag{2.36}$$

and such that the following two relations:

$$\delta_k = \vartheta_k'(\chi_k + \theta_k\kappa_k\zeta_k)r_k, \tag{2.37}$$

$$\vartheta_k'|\mu_k + \theta_k'\lambda_k\zeta_k|r_kc_k < 1 \quad \text{implies} \quad H_k \in \mathcal{HPD}, \tag{2.38}$$

hold where

$$c_k \stackrel{\text{df}}{=} \text{cond}_2(X_k), \quad \zeta_k \stackrel{\text{df}}{=} \frac{(3\sqrt{2}+2)\xi_k'}{2-\xi_k'}. \tag{2.39}$$

Proof. We present here only a main idea of the proof. We are using the index-free notation as in appendix D in [12]. Let us introduce the matrices:

$$Y = X\gamma = P\Sigma Q^H, \quad \Sigma = \text{diag}(\sigma_i), \quad \Sigma^* = \text{diag}(\sigma_i^*), \quad (\sigma_i^* = f(\sigma_i)),$$

$D_z, D_u, D_h, L = \Sigma^* D_u^H D_u = D_z^H D_u + D_h D_u^H$, see (D.19)–(D.23) and (D.29). Let us introduce also the matrix $\mathbf{\Xi} = [\mathbf{\Xi}_{ij}] \stackrel{\text{df}}{=} D_z - D_z^H$.

From (D.16), (D.17) follows that the relations (2.37), (2.38) are equivalent to: $\delta_y = ||S_-||_F(2\sigma_{\min})^{-1}$ **and** $||S_+||_2 < 2\sigma_{\min}$ implies $H_y \in \mathcal{HPD}$, where $\sigma_{\max} \stackrel{\text{df}}{=} \max\{\sigma_i\}, \sigma_{\min} \stackrel{\text{df}}{=} \min\{\sigma_i\}$ and S_-, S_+ are defined in (D.25). The case: $\chi = \mu = \kappa = \lambda = 0$ succeeds iff $\mathbf{\Xi} = 0$ holds. Otherwise we obtain explicit expressions for these parameters applying the *mean-values presentation* to the F-norms of the matrices

$$S_\pm = [\omega_{ij}^{(\pm)}\Xi_{ij}] + \mathbf{\Phi}_\pm, \quad \mathbf{\Phi}_\pm \stackrel{\text{df}}{=} [\omega_{ij}^{(\pm)}\Pi_{ij}] - [\omega_{ij}^{(-)}F_{ij}^{(\pm)}],$$

where $\Pi = [\Pi_{ij}] \stackrel{\text{df}}{=} D_h D_u^H - D_u D_h, F^{(\pm)} = [F_{ij}^{(\pm)}] \stackrel{\text{df}}{=} [\rho_{ji}(L^H)_{ij} \pm \rho_{ij}L_{ij}]$,

$$\omega_{ij}^{(\pm)} \stackrel{\text{df}}{=} \frac{\pm\sigma_i - \sigma_j}{\sigma_i^* + \sigma_j^*}, \quad \rho_{ji} \stackrel{\text{df}}{=} \frac{\sigma_j}{\sigma_i + \sigma_j}.$$

In particular, with $\omega \stackrel{\text{df}}{=} \max\{|\omega_{ij}^{(-)}|\}$, we obtain: $\chi = \tilde{\chi}(\mathbf{\Xi})d_\Xi$, where $\tilde{\chi}(\mathbf{\Xi}) \stackrel{\text{df}}{=} |||[\omega_{ij}^{(-)}\Xi_{ij}]||_F/(\omega||\mathbf{\Xi}||_F), d_\Xi \stackrel{\text{df}}{=} ||\mathbf{\Xi}||_F/(2||D_z||_F)$. □

Remarks 2.3.

(i) The assumption (2.5) implies (2.6), (2.7), hence

$$\vartheta_k' = |\delta_{k+1}(1 + \varepsilon_k') + \varepsilon_k'|, \quad |\varepsilon_k'| \leqslant \hat{\varepsilon}, \tag{2.40}$$

holds since, see (2.29), $\xi_k^* = \delta_{k+1}||X_{k+1}||_2 + \theta_k''||\mathbf{\Phi}_{k+1}||_F$. In section 4 we apply `BIT4` also in cases, when (2.5) is not satisfied (this happens only for large c_k and modifies in (2.40) only the bound on $|\varepsilon_k'|$).

(ii) If (2.5) is satisfied then the following relations:

$$\vartheta_k' \leqslant \vartheta_k, \quad \xi_k' = p_k\xi_k^* \leqslant p_k\xi_k, \quad p_k \stackrel{\text{df}}{=} p(\mathbf{\Psi}_{k+1}), \tag{2.41}$$

hold, where ϑ_k, ξ_k are defined in (13.9). The quantity r_k in (2.33) is identical with r_k in (13.9), hence

$$r_k = \max\{1, (c_k + \rho_k)(c_k\rho_k + 1)^{-1}\}. \tag{2.42}$$

(iii) The bounds (2.36), (2.40) and (2.41) show that `BIT4` implies theorem 3.1 in [12] (the old version of `BIT`).

(iv) The inequalities (2.36) cover all special cases for the parameters $\chi_k, \mu_k, \kappa_k, \lambda_k$ (provided not all are equal to zero). But the construction of these

quantities in the proof shows that in most cases stronger relations can be expected (provided $\chi_k > 0$ holds):

$$\mu_k \ll \chi_k < 1, \quad \kappa_k \ll 1, \quad \lambda_k \ll 1. \tag{2.43}$$

Closer information on χ_k (its dependence on X_k and ρ_k) will be presented in theorem 2.2.

Usually the quantities $\{\xi'_k\}_{k=0}^{l-1}$ are initially fast decreasing until they reach the level ε. In the *normal realization* of `HS` (practical scaling, all matrices G_k satisfying (2.5)) the potentially largest quantity $\xi'_0 \approx p_0|\delta_1 + \varepsilon'_0|f(\sqrt{c_0\tau_0})$ is small, is at most of the order $(K_1+1)\varepsilon^{1/2}, K_1 \stackrel{\text{df}}{=} \delta_1/\hat{\varepsilon}$.

In our *normal* experiments typically $\xi'_0 \leqslant 10^{-7}$ holds. In experiments with *modified scaling* the values ξ'_0 are larger, but never exceed 10^{-3}. Only in experiments with matrices G_k not satisfying (2.5) larger ξ'_0, close to 1 or even larger than 1, appear (but for $k > 0$ all ξ'_k are small, say: $\xi'_k \lesssim 10^{-4}$).

Corollary 2.1. In the case when $\xi'_k \ll 1$ holds, the following simplified version of `BIT4` can be applied:

- If $H_{k+1} \in \mathcal{HPD}$ then there exists real number $\chi_k \in [0,1]$, such that the following *approximate relations* hold:

$$\delta_k \approx \vartheta'_k \chi_k r_k, \tag{2.44}$$

$$\delta_k c_k \lesssim 1 \quad \text{implies} \quad H_k \in \mathcal{HPD}. \tag{2.45}$$

- If matrices G_k, G_{k+1} satisfy (2.5) then we have

$$\delta_k \approx |\delta_{k+1} + \varepsilon'_k|\chi_k r_k, \quad |\varepsilon'_k| \leqslant \hat{\varepsilon}. \tag{2.46}$$

Remarks 2.4.

(i) The implication (2.45) has an informative character. Due to (2.43) $H_k \in \mathcal{HPD}$ is quite probable even when $\delta_k c_k \gg 1$ holds (for example, say: $\delta_k c_k \approx 10$). On the other hand we have not succeeded to prove rigorously that $\delta_k c_k < 1$ implies $H_k \in \mathcal{HPD}$ (though we suppose this might be true in `HS`).

(ii) When c_0 is large the relation $c_k \ll c_0$ for $k > 0$ is typical for all "normal" realizations of `HS`. Hence for $k > 0$ the implication (2.45) yields usually $H_k \in \mathcal{HPD}$ provided δ_k is a modest multiple of $\hat{\varepsilon}$, see (2.4). This means the *continuation* of the *backward induction while* δ_k are *modest multiples* of $\hat{\varepsilon}$. (When c_0 is "small", say: $c_0 < 30$, then all δ_k are modest multiples of $\hat{\varepsilon}$, all $H_k \in \mathcal{HPD}$, see further discussion).

The conclusion 2.1 makes it clear that we can be satisfied with the computed solutions $\{\tilde{U}, \tilde{H}\}$ only if $H_0 \in \mathcal{HPD}$ and δ_0 is a *modest multiple of* ν (a very modest multiple of $\hat{\varepsilon}$). To be on the safe side it is reasonable to expect the same for all $\{H_k, \delta_k\}, k > 0$.

In the initial steps: $k = l-1, l-2, \ldots, s$ of the backward-induction, when c_k are only slightly larger than 1, the *additive character* of the recursion (2.46) *prevails* and the bounds $\delta_k \lesssim \delta_{k+1} + \hat{\varepsilon} \lesssim (l-k)\hat{\varepsilon}$ $(s \leqslant k < l)$ hold (see (2.30) and note that - due to: $r_k = 1$ **or** r_k only slightly larger than 1 - $\chi_k r_k \lesssim 1$ holds).

For $k < s$ $\{c_k\}$ are already distinctly larger than 1, r_k can be even close to ρ_k^{-1} (when $\rho_k < 1$), hence the *multiplicative character* of (2.46) *can prevail* in some steps of the backward induction, acting *destructively* when $\chi_k r_k > 1$ or *soothingly* when $\chi_k r_k < 1$ holds.

The multiplier χ_k acts always "soothingly". Theorem 2.2 shows that for fixed X_k χ_k tends essentially to decrease with ρ_k, if ρ_k is *sufficiently small*.

Theorem 2.2. Let us assume: $\xi_k' \ll 1, H_{k+1} \in \mathcal{HPD}$ and $\eta_k \stackrel{\text{df}}{=} ||\mathbf{\Phi}_{k+1}||_F f_k^{-1} \ll 1$, see (2.6). If $\chi_k > 0$ then there exist the numbers $\hat{d}_k \in (0, 1]$, depending only on matrices U, X_{k+1}, Z_{k+1}, and $w_k \in (c_k^{-1}, c_k)$, depending only on matrices U, X_k (hence independent of ρ_k), such that χ_k can be presented in the form:

$$\chi_k = \hat{d}_k \hat{\chi}_k + 0(\xi_k'), \tag{2.47}$$

where:

$$\delta_{k+1} > \eta_k \quad \text{implies} \quad \hat{d}_k \gtrsim \sqrt{1 - (\eta_k \delta_{k+1}^{-1})^2}, \tag{2.48}$$

$$\hat{\chi}_k \in \Big[\frac{(c_k^{-1} + \rho_k)w_k}{1 + w_k \rho_k}, \min\{1, (c_k^{-1} + \rho_k)w_k\}\Big]. \tag{2.49}$$

Proof. We present here only a main idea of the proof using the same index-free notation as in appendix D [12] and in the proof of theorem 2.1.

The relation (2.47) follows with (see the proof of theorem 2.1) $\hat{d}_k = d_\Xi = ||\mathbf{\Xi}||_F/(2||D_z||_F), \hat{\chi}_k \stackrel{\text{df}}{=} |||[\omega_{ij}^{(-)} W_{ij}]|||_F/(\omega||W||_F)$, where (see (D.27)) $W = \mathbf{\Xi} - \mathbf{\Omega} = \Sigma^* D_u + D_u \Sigma^* = [(\sigma_i^* + \sigma_j^*)D_{ij}], D_u = [D_{ij}], \mathbf{\Omega} \stackrel{\text{df}}{=} D_u D_h + D_z^H D_u, ||\mathbf{\Omega}||_F = O((\xi^*)^2)$. Let us note that, see (2.8), (2.12), (2.13), $\sigma_i = \gamma_k \sigma_i^{(\text{k})} = \sqrt{\rho_k}\hat{\sigma}_i, \hat{\sigma}_i = \gamma_k^{(\text{opt})} \sigma_i^{(\text{k})}$ and $\hat{\chi}_k$ can be presented in the form

$$\hat{\chi}_k = \frac{(\rho_k + c_k^{-1})||T_k||_F}{||\rho_k T_k + N_k||_F}, \quad T_k \stackrel{\text{df}}{=} [(\hat{\sigma}_i + \hat{\sigma}_j)D_{ij}], \quad N_k \stackrel{\text{df}}{=} [(\hat{\sigma}_i^{-1} + \hat{\sigma}_j^{-1})D_{ij}].$$

The quantities $\{\hat{\sigma}_i\}$, the matrices D_u, T_k, N_k depend only on X_k, U (are not depending on γ_k, or ρ_k). We find further that $\hat{\chi}_k$ decreases with ρ_k (at least for sufficiently small ρ_k) and satisfies the bounds (2.49) with $w_k \stackrel{\text{df}}{=} ||T_k||_F/||N_k||_F$.

The relation (2.48) is a consequence of the orthogonality of the matrix $\boldsymbol{\Xi}$ to the matrix $D_z + D_z^H$ in the space $\mathbb{C}^{n\times n}$ of matrices with the inner product $(D, B) \stackrel{\text{df}}{=} \text{tr}(D^H B)$ (see that $D_z = (I + D_u)\boldsymbol{\Psi} + P^H(X - Z)Q = \boldsymbol{\Psi} + R$ and $\boldsymbol{\Psi} = -\boldsymbol{\Psi}^H, ||\boldsymbol{\Psi}||_F = \delta_x||X||_2 (\approx \delta_{k+1} f_k)$ and $||R||_F = ||X - Z||_F + \Theta||D_u\boldsymbol{\Psi}||_F$). $\square$

Remarks 2.5.

(i) The construction of the quantities $\hat{\chi}_k, w_k$, in the proof, indicates that:

- $\hat{\chi}_k$ is an increasing function of ρ_k (decreases with ρ_k).
- $w_k \ll 1$ happens probably when the *majority* of $\{\sigma_i^{(k)}\}$ is smaller distinctly than $\alpha_k \stackrel{\text{df}}{=} \sqrt{\sigma_{\max}^{(k)}\sigma_{\min}^{(k)}}$, see (2.8), (2.10).
- $w_k \gg 1$ happens probably when the *majority* of $\{\sigma_i^{(k)}\}$ is larger distinctly than α_k.

(ii) When $\delta_{k+1} \lesssim \eta_k$ then both factors, $\hat{d}_k$ and $\hat{\chi}_k$, can contribute significantly to the eventual smallness of χ_k. When $\delta_{k+1} \gg \eta_k$ then $\hat{d}_k$ is close to 1 and the eventual smallness of χ_k can result only when $\hat{\chi}_k$ is small.

(iii) When c_k is large, $\rho_k \ll 1$ and $\delta_{k+1} \gg \eta_k$ then the ability of χ_k to reduce significantly the multiplier $r_k, r_k \gg 1$, depends only on w_k:

$$\frac{w_k}{1 + w_k\rho_k} \lesssim \chi_k r_k \approx \hat{\chi}_k r_k \lesssim \min\{r_k, w_k\}.$$

All this means that with the danger, that in the backward induction large r_k would appear, exists an uncertain antidote for it: a *chance* (not *guarantee*!) that $\chi_k r_k \ll r_k$ holds. Unfortunately, this antidote not always acts sufficiently effectively when frequently large r_k appear.

In practical scaling $\rho_k \in [n^{-1/2}, n^{1/2}]$ holds, see (2.19), (2.20), (2.13). From observation, the frequency of the cases ρ_k close to $n^{1/2}$ or close to $n^{-1/2}$ is small. When c_0 is large then initially $\{c_k\}$ decrease very quickly, hence large r_k (close to $\sqrt{n}$) could appear only in a few last steps of the backward

induction, say: for $k = 4, 3, 2, 1, 0$. Regarding the *soothing influence* of the multipliers χ_k, the following conclusion seems to be justified:

Conclusion 2.2. In HS with practical scaling and good matrix-inversion (2.5) there are fair chances that all $\{\delta_k\}$ will be modest multiples of $\hat{\varepsilon}$. Hence $H_k \in \mathcal{HPD}$ for $k > 0$. Also $H_0 \in \mathcal{HPD}$ if $c_0\hat{\varepsilon} \ll 1$ holds. Otherwise, the positive definiteness of H_0 is *probable* but *not sure*.

Till now all our experiments confirm this optimistic hypothesis. This does not eliminate the theoretical danger that for some, not detected yet, matrix A (probably cond (A) and n large) the *multiplicative blow-up* in the last few steps of the backward-induction would succeed yielding δ_0 of the order $\hat{\varepsilon}n^q$ with q distinctly larger than $\frac{1}{2}$. There are some arguments contradicting this *pessimistic warning*, see section 5.

Corollary 2.2. The pair $\{X'_k, \gamma'_k\}$,

$$X'_k \stackrel{\text{df}}{=} X_k^{-H} = P_k \Sigma_k^{-1} Q_k^H, \quad \gamma'_k \stackrel{\text{df}}{=} \gamma_k^{-1}, \tag{2.50}$$

see (2.8), plays in (2.7) "symmetrically" the same role as the pair $\{X_k, \gamma_k\}$. Hence theorems 2.1, 2.2 and corollary 2.1 can be used to express explicitly also the accuracy δ'_k of U as an AUF of the matrix $X'_k \stackrel{\text{df}}{=} X_k^{-H}$. This means that theorems 2.1 and 2.2 can be tested on both matrices, $\tilde{X}_k$ and G_k^H. These experiments show good consistency with the theory. In the following we will report only the behaviour of the computed sequence $\{\delta_k\}$.

3 Tools for numerical experiments

Our experiments were performed only on real matrices, $A \in \mathbb{R}^{n\times n}, n \leqslant 35$. Some modifications of scaling allowed us to simulate the behaviour of the HS-process for larger dimensions, say $35 < n \leqslant 100$.

3.1 Numerical arithmetic, the epsilons

Our experiments were performed in MATLAB, which has a unit roundoff $\nu = \nu_d \approx 2.2 \times 10^{-16}$. We apply systematically the cumulation of "inner-products" on variables of the extended-type, $\nu = \nu_e \approx 10^{-19}$, what reduces distinctly the cumulation of errors "on the level 10^{-16}" and allows to compute some relative residuals with the errors not exceeding 10^{-18}. The computations with higher accuracy were done by means of MATLAB SYMBOLIC MATH Toolbox.

In the evaluation of the HSTEST-results a main role play the *epsilons*: $\varepsilon_x, \varepsilon_g, \hat{\varepsilon}, \varepsilon_l$, see (2.5), (2.7), (2.21), and the bound ε_s:

$$\frac{||X - P\Sigma Q^H||_2}{||X||_2} \leqslant \varepsilon_s, \quad X \in \mathbb{R}^{n\times n}, \tag{3.1}$$

where P, Σ, Q are the factors obtained in the numerical SVD of X. All these epsilons correspond to $0(n^3)$-operations processes. The observations of the computed values of the quantities $\{\delta_k, \delta'_k, e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\}$, see (2.29), corollary 2.2 and (3.6), seem to indicate the relation: $eps \approx \sqrt{n}10^{-16}$, for an *average bound on the errors* in such processes as GEPP or GECP matrix-inversion. Only ε_s is probably distinctly larger: $\varepsilon_s = eps * z$ (we chose $z = 2$). This influences the bounds $\varepsilon_x, \varepsilon_g$ (consequently also $\hat{\varepsilon}$) only in the subsection 4.5: $\varepsilon_x \approx \varepsilon_g \lesssim z * eps$.

3.2 The HSTEST-program

For given matrix $A \in \mathbb{R}^{n\times n}$ and the chosen number *eps* HSTEST performs the *double-sweep* process, computing in both sweeps the same iterates $\{\tilde{X}_k\}_{k=0}^{l}$ ($\tilde{X}_0 := A$) (that means: using the same computed inverses G_k, the same scaling parameters γ_k and the same *stopping criterion* defining the last computed iterate: $\tilde{X}_l = \tilde{U}$). In both sweeps essential quantities are computed and (eventually) printed, presenting only *three leading decimals* in each printed result.

In the first sweep $\{\sigma_{\max}^{(k)}, \sigma_{\min}^{(k)}, \gamma_k\}, k = 0, \ldots, l-1$, are computed (and stored to be used also in the second sweep). After finishing the first sweep the quantity $\Delta_l \stackrel{\mathrm{df}}{=} ||\tilde{X}_l^T \tilde{X}_l - I||_F$ is computed and printed. (If some other tested stopping criterion indicates $\tilde{X}_{l'}, l' \leqslant l$, as the *final iterate* then also $\Delta_{l'}$ is computed and printed).

The matrix $\tilde{U} = \tilde{X}_l$, computed in the first sweep, will be used in the second sweep for computing the quantities δ_k, δ'_k (see corollary 2.2).

In the second sweep in each step ($k = 0, \ldots, l-1$) the quantities: $c_k - 1, \sigma_{\max}^{(k)} - 1, \rho_k, e_k^{(L)}, e_k^{(R)}, \delta_k, \delta'_k$ are computed and, according to the chosen option, eventually printed, see (3.6).

Remarks 3.1.

(i) If the matrix $\tilde{H}$, see (2.31), will not pass the Cholesky *positivity-test*, then this test is repeated for the matrix $H_{\mathrm{I}} = \tilde{H} + (\tilde{\delta}_0 \sigma_{\max}^{(0)} n^{-1/2})I$. The result of the first or of both tests is signalled.

(ii) For computing of $\sigma_{\max}^{(k)}, \sigma_{\min}^{(k)}$ the `SVD` of $\tilde{X}_k$ is performed, yielding the computed values: $\tilde{\sigma}_{\max}^{(k)}, \tilde{\sigma}_{\min}^{(k)}$. If $\tilde{\sigma}_{\min}^{(k)} < 1000 eps * \tilde{\sigma}_{\max}^{(k)}$ holds then also the `SVD` of the matrix G_k is performed, yielding the *corrected value* of $\tilde{\sigma}_{\min}^{(k)}$: $\tilde{\sigma}_{\min}^{(k)} := ||G_k||_2^{-1}$.

`HSTEST` leaves several options to the user, in particular:

– *choosing the* `GEPP` *or* `GECP` *in computing the inverses* $\{G_k\}$, guaranteeing `LRS, RRS` or `NC-Property` of G_k, see section 4.1.

– choosing the inversion guaranteeing *only* the `Conj-Property` of G_k, see section 4.1.

– choosing the stopping criteria (the *used* and *tested* ones).

– choosing: $(1, \infty)$-scaling, (F)-scaling or optimal scaling (with or without some *special modifications*).

– choosing the switching-criterion for the transfer from $(1, \infty)$-scaling to unscaled iterations (for remaining iterations: $k = r+1, \ldots, l-1$).

Remark 3.2. The printed results allow to obtain further information on the tested process. For example, see the following relations (see (2.40)):

$$\left|\delta_{k+1} - |\varepsilon'_k|\right| \sigma_{\max}^{(k+1)} \lesssim \xi_k^* \lesssim (\delta_{k+1} + |\varepsilon'_k|)\sigma_{\max}^{(k+1)}. \tag{3.2}$$

3.3 The accuracy of the `HSTEST` results and their presentation

Let now $\tilde{b}$ ($\tilde{b} > 0$) mean: *the computed value of the quantity b, rounded to three leading figures*. This notation is not always *univocal* since sometimes $\tilde{b}$ can be considered also as the *computed value of some other quantity* $\hat{b}$. Regarding the rounding to three leading figures we can present the full error bounds in the form: $|b - \tilde{b}| \leqslant 5 \times 10^{-3}\tilde{b} + \upsilon(b), |\hat{b} - \tilde{b}| \leqslant 5 \times 10^{-3}\tilde{b} + \upsilon(\hat{b})$, where $\upsilon(b), \upsilon(\hat{b})$ mean, say, the *basic error-bounds*. If in the majority of cases $|b - \tilde{b}| \lesssim 10^{-2}\tilde{b}$ holds then we will present $\tilde{b}$ as b, signaling the eventual exceptions by marking b with the star ($b*$) if $10^{-2}\tilde{\lesssim}|b - \tilde{b}| \lesssim \tilde{b}$ holds *or* with the exclamation mark ($b!$) if there is no bound on $|b - \tilde{b}|$. In most cases $b*$ means that b has at least one good leading figure.

The basic global bound on the error of the computed value $\tilde{\Delta}_l$ of $\Delta_l \stackrel{\text{df}}{=} ||\tilde{U}^H\tilde{U} - I||_F$ is: $\upsilon(\Delta_l) \approx q(n) \times 10^{-19}, q(n) \leqslant n^2$. We assume that in all our experiments $\upsilon(\Delta_l) \lesssim 10^{-18}$ holds since for larger n practically always $q(n) = n^2$ yields an unrealistic over-bound. Similar relation will be assumed

in the following whenever the computation of $\tilde{b}$ involves the matrix $*$ matrix multiplication.

Let us assume that $q_k \stackrel{\rm df}{=} \varepsilon_s \hat{c}_k \ll 0.5$ holds, where $\hat{c}_k \stackrel{\rm df}{=} {\rm cond}_2(\tilde{X}_k)$ and ε_s is defined in (3.1).

Let us consider the computed values $\tilde{x}_k, \tilde{g}_k$ of $x_k \stackrel{\rm df}{=} ||\tilde{X}_k||_2$ and $g_k \stackrel{\rm df}{=} ||G_k||_2$. The basic error bounds are, see remark 3.1 (ii),

$$\upsilon(x_k) \approx \varepsilon_s \tilde{x}_k \quad \upsilon(g_k) \approx \hat{q}_k \tilde{g}_k, \quad \hat{q}_k = \begin{cases} \varepsilon_s, & \text{if } q_k > 5 \times 10^{-3}, \\ q_k, & \text{otherwise.} \end{cases} \tag{3.3}$$

But $\tilde{x}_k, \tilde{g}_k$ can be considered also as the computed values of $\hat{x}_k \stackrel{\rm df}{=} ||X_k||_2, \hat{g}_k \stackrel{\rm df}{=} ||X_k^{-1}||_2$, see (2.5), with basic errors bounds

$$\upsilon(\hat{x}_k) \approx (\varepsilon_s + \varepsilon_x)\tilde{x}_k, \quad \upsilon(\hat{g}_k) \approx (\hat{q}_k + \varepsilon_g)\tilde{g}_k. \tag{3.4}$$

The quantity $\tilde{c}_k := \tilde{x}_k * \tilde{g}_k$ can be considered as the computed value of *both*: $c_k = {\rm cond}(X_k)$ and $\hat{c}_k \stackrel{\rm df}{=} {\rm cond}_2(\tilde{X}_k)$ with basic error bounds

$$\upsilon(c_k) \approx (\varepsilon_s + \varepsilon_x + \hat{q}_k + \varepsilon_g)\tilde{c}_k, \quad \upsilon(\hat{c}_k) \approx (\varepsilon_s + \hat{q}_k)\tilde{c}_k. \tag{3.5}$$

Similar bounds (as for c_k) can be presented for the computed values $\tilde{\rho}_k, \tilde{r}_k$ of ρ_k, r_k, see (2.13), (2.42).

For the computed values $\tilde{e}_k^{(\rm L)}, \tilde{e}_k^{(\rm R)}$ of the relative residuals

$$e_k^{(\rm L)} \stackrel{\rm df}{=} \frac{||I - G_k\tilde{X}_k||_F}{x_k g_k}, \quad e_k^{(\rm R)} \stackrel{\rm df}{=} \frac{||I - \tilde{X}_k G_k||_F}{x_k g_k}, \tag{3.6}$$

the basic error-bounds are:

$$\upsilon(e_k^{(\rm L)}) \approx (\varepsilon_s + \hat{q}_k)\tilde{e}_k^{(\rm L)} + 10^{-18}, \quad \upsilon(e_k^{(\rm R)}) \approx (\varepsilon_s + \hat{q}_k)\tilde{e}_k^{(\rm R)} + 10^{-18}. \tag{3.7}$$

The quantity $\tilde{\delta}_k := ||\tilde{B}_k - \tilde{B}_k^H||_F/(2 * \tilde{x}_k), \tilde{B}_k := \tilde{U}^H * \tilde{X}_k$, can be considered as the computed value of both: δ_k (see (2.29)) and $\hat{\delta}_k \stackrel{\rm df}{=} ||\hat{B}_k - \hat{B}_k^H||_F/(2x_k), \hat{B}_k \stackrel{\rm df}{=} U^H\tilde{X}_k$. The basic error bounds are here, see (2.21):

$$\upsilon(\delta_k) \approx \varepsilon_l\sqrt{n} + \varepsilon_s + \varepsilon_x + 10^{-18}, \quad \upsilon(\hat{\delta}_k) \approx \varepsilon_l\sqrt{n} + \varepsilon_s + 10^{-18}. \tag{3.8}$$

See that $\hat{\delta}_k = \mathtt{acc}(U, \tilde{X}_k)$ if $\hat{H}_k \stackrel{\rm df}{=} \frac{1}{2}(\hat{B}_k + \hat{B}_k^H) \in \mathcal{HPD}$.

Remarks 3.3.

(i) If the matrix G_k is not satisfying (2.5) then the quantities $\varepsilon_x, \varepsilon_g$ in (3.4), (3.5), (3.7), (3.8) should be replaced with $\hat{\varphi}_k$, see section 4.3.

(ii) The expressions (3.4), (3.5), (3.7), (3.8) yield in most cases serious over-bounds on actual error. This would imply categoric declining decision in situations when there are still good chances that the computed result $\tilde{b}$ is sufficiently close to b or $\hat{b}$. Therefore presenting $\tilde{b}$ as b or choosing the *exception-mark* ($*$ or !) we will accept an *optimistic rule*: the *maximum* instead of the *sum*. For example, $\upsilon(c_k) \approx \max\{\varepsilon_s, \varepsilon_x, \hat{q}_k, \varepsilon_g\}\tilde{c}_k, \upsilon(\hat{c}_k) \approx \max\{\varepsilon_s, \hat{q}_k\}\tilde{c}_k$ ($\varepsilon_x, \varepsilon_g$ eventually replaced with $\hat{\varphi}_k$).

(iii) The Cholesky positivity test of the computed Hermitian factor $\tilde{H}$ of A can be trusted at least when the bound $(eps + \tilde{\delta}_0)\mathrm{cond}_2(A) < 1$ holds.

Conclusions 3.1.

(i) The quantities $\tilde{c}_k, \tilde{\rho}_k, \tilde{r}_k$ can be presented as c_k, ρ_k, r_k when

$$q_k^* \stackrel{\mathrm{df}}{=} \max\{\varepsilon_s, \hat{\varphi}_k\}\tilde{c}_k \leqslant 5 \times 10^{-3}$$

holds ($\hat{\varphi}_k = eps/3$ when G_k and G_{k+1} satisfy (2.5), see section 4.3). The exceptions will be marked with $*$ *or* !, according to the cases: $5 \times 10^{-3} < q_k^* \lesssim 0.5$ *or* $0.5 < q_k^*$.

(ii) The quantities $\tilde{e}_k^{(\mathrm{L})}, \tilde{e}_k^{(\mathrm{R})}$ larger than 10^{-16} can be presented as $e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}$. Otherwise they will be marked with $*$ (due to the skipped term 10^{-18}).

(iii) The same (as in (ii)) holds for the quantities $\tilde{\delta}_k, \delta_k$, provided $\hat{\varphi}_k \leqslant 0,01$ holds. Eventually δ_k will be marked with $*$ or !, according to the cases $0.01 \leqslant \hat{\varphi}_k \leqslant 0.1$ or $0.1 < \hat{\varphi}_k$.

Example 3.1. In table 1 we present the results of the `HSTEST`-program for the 10×10 matrix $A_1 = \mathtt{tril}(\mathtt{rand}(10))^8\mathtt{rand}(U)$, see [4], applying (F)-scaling and `GEPP` matrix-inversion. The first sweep yields the computed `AUF` $\tilde{U} = \tilde{X}_9$ of A_1, $\tilde{\Delta}_9 = 5.14 \times 10^{-16}$. The computed factor $\tilde{H}$, see (2.31), of A_1 passed the positivity test.

Table 1

k	$c_k - 1$	ρ_k	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	δ_k
0	$8.74e+14*$	$0.930*$	$3.10e-17*$	$8.72e-09$	$5.12e-09*$
1	$1.66e+06$	0.708	$3.28e-17*$	$1.96e-15$	$1.19e-15$
2	$7.56e+02$	1.00	$5.90e-17*$	$7.52e-16$	$4.09e-16$
3	$1.19e+01$	0.732	$1.07e-16$	$1.44e-16$	$2.68e-16$
4	$1.17e+00$	1.07	$2.97e-16$	$2.95e-16$	$2.80e-16$
5	$8.38e-02$	1.03	$5.08e-16$	$5.16e-16$	$3.43e-16$
6	$1.51e-03$	1.00	$5.74e-16$	$5.74e-16$	$3.40e-16$
7	$7.01e-07$	1.00	$5.35e-16$	$5.35e-16$	$2.64e-16$
8	$2.46e-13$	1.00	$4.84e-16$	$4.84e-16$	$1.80e-16$

Remarks 3.4.

(i) The quantities marked with $*$ have probably still at least one correct leading figure. In the case of c_0, ρ_0 it is implied by additional information on $\mathrm{cond}_2(A_1)$ gained in other experiments (judging only from this experiment c_0 should be marked with !).

(ii) $e_o^{(\mathrm{R})} \approx 8.72 \times 10^{-9}$ indicates that G_0 has only the `LRS-Property` (see the next section for the definition)). This is the reason why $\tilde{U}$ is a poor `AUF` of A: if $\tilde{H}$ is really positive-definite then $\mathtt{acc}(U, A) \approx \delta_0 \approx 5.12 \times 10^{-9}$. The computed polar factors $\{\tilde{U}, \tilde{H}\}$ are not acceptable. See subsection 4.5 for further discussion.

(iii) The computed values for $k > 3$ are typical for all our experiments and fully consistent with the presented theory. In the following we will present only the relevant part of the experimental results, skipping the trivial part of them.

4 The quality problem of the matrix-inversion in the numerical Higham's algorithm (`HS`)

Contemporary standard matrix-inversion procedures use the Gaussian triangular factorization with partial pivoting (`GEPP`) of the inverted matrix, see [4]. Using these procedures in the `HS`-process yields frequently (but not always) acceptable results.

We will show experimentally that for some special matrices A the quality of the `GEPP`-inversion is not sufficient to guarantee the acceptability (1.4) of the unitary polar factor $\tilde{U}$ computed in the `HS`-process.

Some other (more expensive) algorithms compute always inverses with sufficient quality not impending the good behaviour of the HS-process (for example the inversion: via triangular factorization with complete pivoting (GECP) **or** via Householder qr-factorization with column pivoting (QRCP)). We should recognize the properties of the computed inverses $\{G_k\}$ not impending the good behaviour of the HS-process and those properties which can spoil the accuracy in this process.

4.1 Properties of computed inverses

Let G be the computed inverse of the nonsingular matrix X. We introduce auxiliary quantities

$$x \stackrel{\text{df}}{=} ||X||_2, \quad g \stackrel{\text{df}}{=} ||G||_2, \quad c \stackrel{\text{df}}{=} \text{cond}_2(X) \quad (= x||X^{-1}||_2) \tag{4.1}$$

and consider the following four eventual properties of G:

$$||G - X^{-1}||_F \leqslant \varepsilon g c, \tag{4.2}$$

$$||GX - I||_F \leqslant \varepsilon g x, \tag{4.3}$$

$$||XG - I||_F \leqslant \varepsilon g x, \tag{4.4}$$

$$\exists \Delta', \Delta: \quad G + \Delta' = (X + \Delta)^{-1}, \quad ||\Delta'||_F \leqslant \varepsilon_g g, \quad ||\Delta||_F \leqslant \varepsilon_x x. \tag{4.5}$$

The same relations define the *properties of inversion procedures* as follows: Let $\mathbb{M}$ be a subset of nonsingular $n \times n$ matrices X. We say that an inversion algorithm Inv is *numerically stable* (NS) in $\mathbb{M}$ if for each $X \in \mathbb{M}$ the computed inverse G satisfies (4.2). In the same way:

– (4.3) defines the left-residual stability (LRS) of Inv in $\mathbb{M}$,

– (4.4) defines the right-residual stability (RRS) of Inv in $\mathbb{M}$,

– (4.5) defines the numerical correctness (NC) of Inv in $\mathbb{M}$.

We shall use the same notation: NS, LRS, RRS, NC for the properties (4.2)–(4.5) of the matrix G, no matter what is the *official property* of the algorithm which computed G (for some matrices $X \in \mathbb{M}$ the computed inverse G can have also some stronger property than the property guaranteed by Inv for the whole subset $\mathbb{M}$).

We define also two *combined* properties of G:

$$\texttt{Alt} \stackrel{\text{df}}{=} \texttt{LRS} \textit{ or } \texttt{RRS}, \quad \texttt{Conj} \stackrel{\text{df}}{=} \texttt{LRS} \textit{ and } \texttt{RRS}. \tag{4.6}$$

Assuming $\varepsilon_x + \varepsilon_g + \varepsilon_x\varepsilon_g \leqslant \varepsilon$ and $\varepsilon xg < 1$ we find the following implications:

$$\texttt{NC} \Longrightarrow \texttt{Conj} \Longrightarrow \texttt{Alt} \Longrightarrow \texttt{NS} \tag{4.7}$$

and the bounds

$$||GX - I||_F \leqslant c||XG - I||_F, \quad ||XG - I||_F \leqslant c||GX - I||_F. \tag{4.8}$$

Let us note further that for small c, say $c \leqslant 10$, `NS` implies `NC` with $\Delta = 0$ and $\varepsilon_g = 10\varepsilon$ (provided 10ε can be accepted as a quantity of the order ν). Hence the listed properties of G: `NS`, `LRS`, `RRS`, `NC`, `Alt`, `Conj` can differ distinctly only when $c = \mathrm{cond}_2(X)$ is sufficiently large.

Further definitions: We will say that G has *only* the property `LRS` (**or**: G has the `LRS-Property-Only`) if G has the `LRS-Property` but has not the `RRS-Property`. In this case, due to (4.6) and (4.7), G has the `Alt-Property` (hence also `NS`) but has *neither* property `Conj` *nor* `NC`. In the same way, using the term: *to have only* (or: `Property-Only`), we define other eventual *highest properties* of G in the hierarchical system defined by (4.6), (4.7).

Let us note at last that the `NC-Property` is really the highest general quality (which can be achieved) of an inverse G computed in a constant finite precision. According to the formulation of W. Kahan, in this `NC`-case: G is *a slightly wrong inverse of a slightly wrong matrix* X. We show in our experiments that only the `NC-Property` of matrices $\{G_k\}$ guarantees good behaviour of the `HS`-process. The considered hierarchical system of properties does not include of course all possible properties of computed inverses. For example we do not consider here the elementwise properties, see [4]. The only exception is the following example 4.1, where elementwise bounds allow to visualise the relations between considered norm-properties.

Example 4.1. Let us consider the matrices: $X = \mathrm{diag}(c, \sqrt{c}, 1), G = X^{-1} + \Gamma, \Gamma = (\gamma_{ij})$ with $c > 1$ and $\varepsilon c \ll 1$ and their norms (compare (4.1)): $x \stackrel{\mathrm{df}}{=} ||X||_2 = c = \mathrm{cond}_2(X), g \stackrel{\mathrm{df}}{=} ||G||_2 = 1 + \theta\gamma$, where $\gamma \stackrel{\mathrm{df}}{=} ||\Gamma||_2 = \beta^{-1}||\Gamma||_F, 1 \leqslant \beta \leqslant \sqrt{3}$. We want to give realistic (closely achievable) upper bounds on $|\gamma_{ij}|$ for $i, j = 1, 2, 3$: $|\Gamma| \leqslant Z = [z_{ij}]$ for G with considered properties (4.2)–(4.6). Due to (4.7) the weakest property (4.2) is always fulfilled, hence the bound $||\Gamma||_F \leqslant \varepsilon cg = \varepsilon c(1 + \theta\beta^{-1}||\Gamma||_F)$ holds, what implies

$$\gamma \leqslant ||\Gamma||_F \leqslant \varepsilon' c, \quad \varepsilon' \stackrel{\mathrm{df}}{=} \frac{\varepsilon}{1 - \varepsilon c}, \quad \frac{1}{1 + \varepsilon c} < g \leqslant \frac{1}{1 - \varepsilon c}.$$

For the properties NC, LRS, RRS, Conj of G we obtain hence the following upper bounds Z on $|\Gamma|$:

$$Z_{\mathtt{NS}} = \varepsilon' \begin{bmatrix} c & c & c \\ c & c & c \\ c & c & c \end{bmatrix}, \quad Z_{\mathtt{LRS}} = \varepsilon' \begin{bmatrix} 1 & \sqrt{c} & c \\ 1 & \sqrt{c} & c \\ 1 & \sqrt{c} & c \end{bmatrix},$$

$$Z_{\mathtt{RRS}} = \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ \sqrt{c} & \sqrt{c} & \sqrt{c} \\ c & c & c \end{bmatrix}, \quad Z_{\mathtt{Conj}} = \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ 1 & \sqrt{c} & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix}.$$

For the NC-Property of G with $\varepsilon_x + \varepsilon_g + \varepsilon_x\varepsilon_g \leqslant \varepsilon$ the bound is

$$Z_{\mathtt{NC}} = \frac{\varepsilon_x}{1-\varepsilon_x c} \begin{bmatrix} c^{-1} & c^{-1/2} & 1 \\ c^{-1/2} & 1 & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix} + \frac{\varepsilon_g}{1-\varepsilon c} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Hence

$$Z_{\mathtt{NC}} < \varepsilon' \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & \sqrt{c} \\ 1 & \sqrt{c} & c \end{bmatrix}.$$

Let us note that each individual bound $|\gamma_{ij}| \leqslant z_{ij}$ is realistic. The elementwise bounds $|\Gamma| \leqslant Z$ are qualitatively equivalent to corresponding norm-properties. For example, $|\Gamma| \leqslant Z_{\mathtt{LRS}}$ implies $||GX - I||_F < 3\varepsilon'(1+\varepsilon c)gx$, compare with (4.3). For all other considered properties similar implications hold.

4.2 Wilkinson's conjecture

There are several versions of computing the inverse G from the GEPP-triangular factorization of matrix X, see [4], which are either left-residual or right-residual stable in a broad subset $\mathbb{M}$ of $n \times n$ matrices. Hence such GEPP-inversion algorithms guarantee the Alt-Property of computed inverses, see (4.6). For well-conditioned matrices X, say $c \leqslant 10$, it means practically Conj-Property of G, see (4.8). But also for badly-conditioned matrices X we can check directly that quite frequently both computed residuals $||GX - I||_F$ and $||XG - I||_F$ are small, are bounded by εxg, see [14], [4]. This means that G has the Conj-Property, in spite of (4.8) with large c.

J.H. Wilkinson explained this phenomenon in [14], pp. 110–111, showing that the matrix G (computed via GEPP-factorization) has the NC-Property, provided the triangular systems - involved in the computation of G from

GEPP - are solved to high accuracy. This happens frequently though not always, see also [13]. Let us formulate this as follows:

Wilkinson's conjecture. If an inverse G, computed via GEPP-factorization of X, has the Conj-Property then, probably, G has also the (stronger) NC-Property.

We do not know whether the GEPP-inversion can produce the computed inverse G with Conj-Property-Only.

The experiments given in subsection 4.5 show that the smallness of both relative residuals $\{e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\}$, see (3.6), (the Conj-Property) is not sufficient to guarantee good behaviour of HS. In all our experiments with GEPP-inversion, subsection 4.4, the good behaviour of HS (smallness of $\{\delta_k\}$) succeeds *iff* both residuals $\{e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\}$ are small. This is our experimental contribution in justifying the Wilkinson's conjecture.

4.3 HS with inverses not always satisfying (2.5)

In (2.5) we postulate in fact the NC-Property (4.5) of the computed inverses G_k of $\tilde{X}_k$ in the whole HS-process, $k = 0, \ldots, l-1$, see below remark 4.1. This implies the essential relations (2.6), (2.7) in our description of HS in section 2. But in some experiments in this section we obtain (in several steps only, when c_k is large) the computed inverses G_k having only the Alt-Property or only the Conj-Property. To interpret the results of such experiments it is convenient to incorporate such *deviations* (from the normality of (2.5)) into our general description of the HS-process.

The first step is to uniformize the description of the computed inverses. We are doing this *par force*, assuming (4.9) with *false epsilon* $\check{\varepsilon}_k$:

$$X_k = \tilde{X}_k + \mathbf{\Delta}_k,\ X_k^{-1} = G_k + \mathbf{\Delta}'_k,\ \check{\varepsilon}_k \stackrel{\mathrm{df}}{=} \max\Big\{\frac{||\mathbf{\Delta}_k||_F}{||X_k||_2}, \frac{||\mathbf{\Delta}'_k||_F}{||X_k^{-1}||_2}\Big\}. \quad (4.9)$$

If G_k has NC-Property then $\check{\varepsilon}_k$ is a *true epsilon*, otherwise $\check{\varepsilon}_k$ is distinctly larger than could be normally accepted as a modest multiple of ν. We want to keep $\check{\varepsilon}_k$ as small as possible. We will see that in the considered cases exist such matrices $\mathbf{\Delta}_k, \mathbf{\Delta}'_k$ that (4.9) holds with $\check{\varepsilon}_k$ smaller than 1. Let us now learn how this quantity can exceed the admissible level ε of the *true epsilons*.

Remark 4.1. (4.9) implies the NC-Property of G_k if $\check{\varepsilon}_k/(1-\check{\varepsilon}_k) \leqslant \varepsilon$ holds. The NC-Property of G_k implies (4.9) with $\check{\varepsilon}_k \lesssim \varepsilon/(1-\varepsilon)$. Hence (2.5) is practically equivalent to the NC-Property of G_k.

Let us assume that c_k is large and $\varepsilon c_k \ll 1$ holds. We present below a *simplified version* of theorem 4.1, using approximate equalities $a \approx b$ (a, b non-negative), meaning some of the following possibilities: $|a-b| \leqslant O(\varepsilon), |a-b| \leqslant O(\varepsilon c_k)\max\{a,b\}, |a-b| \leqslant O(c_k^{-1})\max\{a,b\}$.

Let G_k be the computed inverse of $\tilde{X}_k$ and $e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}$, see (3.6), the relative residuals, and let us define the quantities

$$\varepsilon_k^{(\mathrm{A})} \stackrel{\mathrm{df}}{=} \min\{e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\}, \quad \check{\varepsilon}_k^{(\mathrm{A})} \stackrel{\mathrm{df}}{=} \max\{e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}\}, \quad c_k \stackrel{\mathrm{df}}{=} \mathrm{cond}_2(X_k). \tag{4.10}$$

Theorem 4.1. (simplified version). Let G_k be the computed inverse of $\tilde{X}_k$ and let (4.10) hold. If G_k has at least the `Alt-Property` (hence $\varepsilon_k^{(\mathrm{A})}$ is a true epsilon) then we can choose such matrices $\boldsymbol{\Delta}_k, \boldsymbol{\Delta}'_k$ in (4.9) that the following approximate equalities are fulfilled:

$$\check{\varepsilon}_k \approx \check{\varepsilon}_x^{(\mathrm{k})} \approx \check{\varepsilon}_g^{(\mathrm{k})} \approx \hat{\varphi}_k, \quad \text{where } \check{\varepsilon}_x^{(\mathrm{k})} \stackrel{\mathrm{df}}{=} \frac{||\boldsymbol{\Delta}_k||_F}{||X_k||_2},\ \check{\varepsilon}_g^{(\mathrm{k})} \stackrel{\mathrm{df}}{=} \frac{||\boldsymbol{\Delta}'_k||_F}{||X_k^{-1}||_2} \tag{4.11}$$

and the quantity $\hat{\varphi}_k$ is specified according to the assumed property of G_k:

(i) if G_k has the `Alt-Property` then

$$\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{Alt})} \approx \frac{1}{2}\varepsilon_k^{(\mathrm{A})}\beta_k^{(\mathrm{Alt})}c_k, \quad \beta_k^{(\mathrm{Alt})} \in [c_k^{-1}, 1], \tag{4.12}$$

$$\frac{1}{2}\check{\varepsilon}_k^{(\mathrm{A})} \lesssim \hat{\varphi}_k^{(\mathrm{Alt})} \lesssim \sqrt{\frac{1}{2}\varepsilon_k^{(\mathrm{A})}\check{\varepsilon}_k^{(\mathrm{A})}c_k}, \tag{4.13}$$

(ii) if G_k has the `Conj-Property` then with $\beta_k^{(\mathrm{Conj})} \in [c_k^{-1/2}, 1]$

$$\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{Conj})} \approx \frac{1}{2}\varepsilon_k^{(\mathrm{Conj})}\beta_k^{(\mathrm{Conj})}c_k^{1/2}, \quad \varepsilon_k^{(\mathrm{A})} \leqslant \varepsilon_k^{(\mathrm{Conj})} \leqslant \sqrt{(e_k^{(\mathrm{L})})^2 + (e_k^{(\mathrm{R})})^2}, \tag{4.14}$$

(iii) if G_k has the `NC-Property` then

$$\hat{\varphi}_k = \hat{\varphi}_k^{(\mathrm{NC})} \leqslant \max\{\varepsilon_x, \varepsilon_g\} + \nu\sqrt{n} \approx \frac{1}{3}\hat{\epsilon}. \tag{4.15}$$

Proof. We present here only a main idea of the proof. Let use the following index-free notation (not identical with such notation in section 2): $X \stackrel{\mathrm{df}}{=} \tilde{X}_k, G \stackrel{\mathrm{df}}{=} G_k, c \stackrel{\mathrm{df}}{=} \mathrm{cond}_2(X), \varepsilon_{\mathrm{A}} \stackrel{\mathrm{df}}{=} \varepsilon_k^{(\mathrm{A})}, \check{\varepsilon}_{\mathrm{A}} \stackrel{\mathrm{df}}{=} \check{\varepsilon}_k^{(\mathrm{A})}$. Not lessening the

generality of considerations, let us assume: $x \stackrel{\mathrm{df}}{=} ||X||_2 = c$, what implies (with P, Q unitary and $p_0 \stackrel{\mathrm{df}}{=} p(X^{-1} - G)$): $X = P\Sigma Q, \Sigma = \mathrm{diag}(\sigma_i), 1 \leqslant \sigma_i \leqslant c, g \stackrel{\mathrm{df}}{=} ||G||_2 = (1 + \theta_0 p_0 \varepsilon_{\mathrm{A}})^{-1}$. The quantity $\check{\varepsilon} \stackrel{\mathrm{df}}{=} \check{\varepsilon}_k$, see (4.9), can be presented hence in the form:

$$\check{\varepsilon} = \hat{\varphi}(\mathbf{\Phi}) \stackrel{\mathrm{df}}{=} \max\Big\{\frac{||\mathbf{\Phi}||_F}{||\Sigma + \mathbf{\Phi}||_2}, \frac{||\mathbf{\Psi}||_F}{||\hat{G} + \mathbf{\Psi}||_2}\Big\}, \quad \mathbf{\Psi} = (\Sigma + \mathbf{\Phi})^{-1} - \hat{G},$$

where $\mathbf{\Phi} \stackrel{\mathrm{df}}{=} P^H \mathbf{\Delta} Q$, $\hat{G} \stackrel{\mathrm{df}}{=} Q^H G P$, $\mathbf{\Psi} \stackrel{\mathrm{df}}{=} Q^H \mathbf{\Delta}' P$, and, under the assumption $\zeta_2 = \zeta_2(\mathbf{\Phi}) \stackrel{\mathrm{df}}{=} \min\{||\Sigma^{-1}\mathbf{\Phi}||_2, ||\mathbf{\Phi}\Sigma^{-1}||_2\} < 1$, we have $\hat{G} + \mathbf{\Psi} = (\Sigma + \mathbf{\Phi})^{-1} = \Sigma^{-1} - \Sigma^{-1}\mathbf{\Phi}\Sigma^{-1} + Z(\mathbf{\Phi}), ||Z(\mathbf{\Phi})||_2 \leqslant \zeta_2^2/(1 - \zeta_2)$. Minimizing the *linear part* $\varphi^*(\mathbf{\Phi})$ of $\hat{\varphi}(\mathbf{\Phi})$:

$$\varphi^*(\mathbf{\Phi}) \stackrel{\mathrm{df}}{=} \max\Big\{\frac{||\mathbf{\Phi}||_F}{c}, \frac{||\mathbf{\Gamma} - \Sigma^{-1}\mathbf{\Phi}\Sigma^{-1}||_F}{g}\Big\}, \quad \mathbf{\Gamma} \stackrel{\mathrm{df}}{=} \Sigma^{-1} - \hat{G},$$

we obtain the *minimizer* $\mathbf{\Phi}^*$ such that the equalities

$$\varphi^* \stackrel{\mathrm{df}}{=} \min_{\mathbf{\Phi}} \varphi^*(\mathbf{\Phi}) = \frac{||\mathbf{\Phi}^*||_F}{c} = \frac{||\mathbf{\Gamma} - \Sigma^{-1}\mathbf{\Phi}^*\Sigma^{-1}||_F}{g}$$

hold. Specifying $\mathbf{\Gamma}$ for the cases: (i) and (ii) we find that in both cases, for $\mathbf{\Phi}^* = \mathbf{\Phi}^*_{\mathrm{Alt}}$ and $\mathbf{\Phi}^* = \mathbf{\Phi}^*_{\mathrm{Conj}}, ||\Sigma^{-1}\mathbf{\Phi}^*||_F \lesssim \varepsilon_{\mathrm{A}} c$ holds. Hence $||Z(\mathbf{\Phi}^*)||_F \leqslant O(\varepsilon c)^2$ and (with $\mathbf{\Delta} = P\mathbf{\Phi}^* Q^H, \mathbf{\Delta}' = (X + \mathbf{\Delta})^{-1} - G$) we have $\mathrm{cond}_2(X + \mathbf{\Delta}) = \mathrm{cond}_2(G + \mathbf{\Delta}') \approx c$. This means that $c_k \stackrel{\mathrm{df}}{=} \mathrm{cond}_2(X_k) \approx c \stackrel{\mathrm{df}}{=} \mathrm{cond}_2(\tilde{X}_k)$. Ultimately the quantities $\hat{\varphi}_k^{(\mathrm{Alt})} \stackrel{\mathrm{df}}{=} \hat{\varphi}(\mathbf{\Phi}^*_{\mathrm{Alt}}), \hat{\varphi}_k^{(\mathrm{Conj})} \stackrel{\mathrm{df}}{=} \hat{\varphi}(\mathbf{\Phi}^*_{\mathrm{Conj}})$ satisfy the relations (4.11)–(4.14). (Only the proof of (4.13) needs some more argumentation). The case (iii) needs no proof. □

Remarks 4.2.

(i) (4.13) is valid also when G_k has the `Conj-Property`. Let us notice that both (4.13) and (4.14) imply the same bound: $\hat{\varphi}_k^{(\mathrm{Conj})} \lesssim \varepsilon(c_k/2)^{1/2}$.

(ii) When G_k has only the `Alt-Property` then the upper bound on $\hat{\varphi}_k^{(\mathrm{Alt})}$ in (4.13) is pointless as an *a priori* bound. But can be useful (and often is) as an *a posteriori* bound, if effectively computed quantity $\check{\varepsilon}_k^{(\mathrm{A})}$ is much smaller than $\varepsilon_k^{(\mathrm{A})} c_k$. A very generous way (in the proof) of obtaining this bound indicates that in the case of $\check{\varepsilon}_k^{(\mathrm{A})}$ distinctly larger than ε we can expect the relation

$$\hat{\varphi}_k^{(\mathrm{Alt})} = \check{\varepsilon}_k^{(\mathrm{A})} z, \quad z \geqslant 0.5 \tag{4.16}$$

with a modest number z: $z \leqslant 1?, z \leqslant 3?, \ldots$.

(iii) With large c_k the quantities $\beta_k^{(\mathrm{Alt})}, \beta_k^{(\mathrm{Conj})}$ in most cases will be much smaller than 1. Let "β close to 1" mean: $\beta \in [0.05, 1]$. Then the construction of $\beta_k^{(\mathrm{Alt})}, \beta_k^{(\mathrm{Conj})}$ in the proof suggests that:

- $\beta_k^{(\mathrm{Alt})}$ close to 1 is probable (but not warranted) when a significant part of pairs $\{\tilde{\sigma}_i^{(\mathrm{k})}, \tilde{\sigma}_j^{(\mathrm{k})}\}$ of the singular values of $\tilde{X}_k$ is close to $\{\tilde{\sigma}_{\min}^{(\mathrm{k})}, \tilde{\sigma}_{\max}^{(\mathrm{k})}\}$.
- $\beta_k^{(\mathrm{Conj})}$ close to 1 is probable (but not warranted) when a significant part of $\{\tilde{\sigma}_i^{(\mathrm{k})}\}$ is close to $\tilde{\alpha}_k \overset{\mathrm{df}}{=} (\tilde{\sigma}_{\min}^{(\mathrm{k})} \tilde{\sigma}_{\max}^{(\mathrm{k})})^{1/2}$.

(iv) Comparing in example 4.1 the bound-matrices $Z_{\mathtt{LRS}}, Z_{\mathtt{RRS}}, Z_{\mathtt{Conj}}$ with $Z_{\mathtt{NC}}$ we obtain essentially the same information as in (*iii*). See that $\varepsilon' \approx \varepsilon_x + \varepsilon_g$ in this example corresponds to $\check{\varepsilon}_x^{(\mathrm{k})} + \check{\varepsilon}_g^{(\mathrm{k})} \approx 2\hat{\varphi}_k$.

From (2.6), (4.9), (4.11) we obtain upper bounds on $|\mathbf{\Phi}_{k+1}|$ and $||\mathbf{\Phi}_{k+1}||_F$:

$$|\mathbf{\Phi}_{k+1}| \leqslant \frac{1}{2}\Big[(|\mathbf{\Delta}_k| + \nu|\tilde{X}_k|)\gamma_k + \frac{(|\mathbf{\Delta}'_k| + \nu|G_k|)^T}{\gamma_k}\Big] + \Big(|\mathbf{\Delta}_{k+1}| + \nu|\tilde{X}_{k+1}|\Big), \tag{4.17}$$

$$||\mathbf{\Phi}_{k+1}||_F \lesssim \frac{1}{2}\Big[(\check{\varepsilon}_x^{(\mathrm{k})} + \nu\sqrt{n})x_k\gamma_k + \frac{(\check{\varepsilon}_g^{(\mathrm{k})} + \nu\sqrt{n})\upsilon_k}{\gamma_k}\Big] + (\check{\varepsilon}_x^{(\mathrm{k}+1)} + \nu\sqrt{n})x_{k+1}, \tag{4.18}$$

where $x_k \overset{\mathrm{df}}{=} ||X_k||_2, \upsilon_k \overset{\mathrm{df}}{=} ||X_k^{-1}||_2$. Due to $x_{k+1} \approx f_k > \frac{1}{2}\max\{x_k\gamma_k, \upsilon_k/\gamma_k\}$ the relation $||\mathbf{\Phi}_{k+1}||_F \lesssim f_k\check{\varepsilon}_k^*, \check{\varepsilon}_k^* \overset{\mathrm{df}}{=} 2\hat{\varphi}_k + \hat{\varphi}_{k+1}$ holds. (We are skipping here the $\nu\sqrt{n}$-terms when $\check{\varepsilon}_x^{(\mathrm{k})}, \check{\varepsilon}_g^{(\mathrm{k})}$ are not *true epsilons*).

Conclusion 4.1. If the computed inverses $\{G_k\}$ in the `HS`-process have: the `Alt-`, the `Conj-` or the `NC-Property` then the main relation describing the process can be presented as follows:

$$X_{k+1} = Z_{k+1} + \mathbf{\Phi}_{k+1}, \quad ||\mathbf{\Phi}_{k+1}||_F \lesssim \check{\varepsilon}_k^* f_k, \quad \check{\varepsilon}_k^* = 2\hat{\varphi}_k + \hat{\varphi}_{k+1} \tag{4.19}$$

(Z_{k+1}, f_k defined in (2.6), (2.7), quantities $\{\hat{\varphi}_k\}$ defined in theorem 4.1).

Theorems 2.1 and 2.2 remain valid, but if the matrix G_k or G_{k+1} is not satisfying (2.5) then $\hat{\varepsilon}$ should be replaced with $\check{\varepsilon}_k^*$ in (2.40), (2.46). In particular, assuming $H_{k+1} \in \mathcal{HPD}$ and $\xi'_k \ll 1$, we have

$$\delta_k \approx |\delta_{k+1} + \varepsilon'_k|\chi_k r_k, \quad |\varepsilon'_k| \leqslant \check{\varepsilon}_k^* = 2\hat{\varphi}_k + \hat{\varphi}_{k+1}. \tag{4.20}$$

This means that G_k or G_{k+1} not satisfying (2.5) can spoil seriousely the quality of U as an AUF of X_k: $\delta_k \gg \delta_{k+1}$. The experiments in the subsections 4.4 and 4.5 show that this can really happen.

4.4 Experiments with inverses *having only* the Alt-Property

In the experiments of this section we apply in the HS-process the optimal or practical scaling and we use the matrix-inversion via GEPP-triangular decomposition of the inverted matrix. It is essentially the B-method, see [4] and [8], completed eventually with some transpositions of matrices, to yield *either* the version guaranteeing the LRS-Property *or* (with transpositions) the version guaranteeing the RRS-Property of the computed inverses. For some *special matrices* these computed inverses have *only* the Alt-Property and only in such cases we can observe that this quality of inversion is not sufficiently good for the HS-process. In all other cases (what happens frequently) these computed inverses have the Conj-Property, hence probably also the NC-Property (according to the Wilkinson's-conjecture, see subsection 4.2).

In these experiments we never have met the case of the computed inverse G_k having the Conj-Property, correlated with an evident deterioration of the quantity $\delta_k = \mathtt{acc}(U, X_k) : \delta_k \gg \delta_{k+1}$ (this could indicate that G_k has *only* the Conj-Property). In this way our experiments contribute to justifying the Wilkinson's conjecture.

As the test matrices we choose here the matrices from [4]:

$$A_1 = QL^T, \quad Q \text{ orthogonal, random}, \quad L = \mathtt{tril}(\mathtt{rand}(10))^8, \quad A_2 = A_1^T,$$

$$A_3^{(n)} = QL_3^T, \quad Q \text{ orthogonal, random}, \quad L_3 = \mathtt{qr}(\mathtt{vand}(n))^T, \ n \geqslant 15.$$

That are special matrices yielding G_0 with *only* the Alt-Property (when inverted via appropriate GEPP-factorization). In some cases also G_1 has *only* the Alt-Property.

In all presented below experiments (with GEPP-triangularizations) the computed $\tilde{U} = \tilde{X}_l$ cannot be accepted as a sufficiently good AUF of the matrix A_p $(p = 1, 2, 3)$.

Conclusion 4.3. Alt-Property Only is not a sufficiently good quality of the matrix-inversion in the HS-process.

Examples 4.3. We apply here GEPP-inversion; in (ii) the version guaranteeing the RR-Property of the computed inverses, in (i), (iii), (iv) the version

guaranteeing the `LR-Property`. In (i), (ii), (iii) the matrices $\tilde{H}$ passed the positivity test. In (iv) only the matrix H_{I} passed the positivity test, see remark 3.1(i).

(i) see example 3.1.

(ii) The experiment "symmetric" to (i) with the matrix A_2, $\tilde{\Delta}_9 = 6.2 \times 10^{-16}$, is presented in table 5.

Table 5

k	c_k	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	δ_k
0	$8.75e+14*$	$8.79e-09$	$3.25e-17*$	$5.45e-09*$
1	$1.86e+06$	$5.57e-15$	$6.12e-17*$	$2.69e-15$
2	$2.96e+02$	$6.39e-16$	$3.46e-16$	$3.46e-16$

(iii) Experiment with the matrix $A_3^{(15)}, \tilde{\Delta}_{10} = 9.17 \times 10^{-16}$, is presented in table 6.

Table 6

k	c_k	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	δ_k
0	$1.58e+13$	$3.68e-17*$	$3.91e-14$	$2.13e-14$
1	$1.11e+06$	$8.92e-17*$	$1.65e-14$	$8.23e-15$
2	$4.82e+02$	$1.38e-16$	$1.21e-15$	$7.12e-16$
3	$1.15e+01$	$2.22e-16$	$3.01e-16$	$5.47e-16$

(iv) Experiment with the matrix $A_3^{(25)}, \tilde{\Delta}_{10} = 2.46e-15$, is presented in table 7.

Table 7

k	c_k	$e_k^{(\mathrm{L})}$	$e_k^{(\mathrm{R})}$	δ_k
0	$1.87e+18!$	$2.93e-17*$	$1.39e-10$	$8.55e-11!$
1	$4.25e+08$	$8.65e-17*$	$1.67e-12$	$7.67e-13$
2	$1.10e+04$	$1.15e-16$	$6.69e-15$	$3.75e-15$
3	$5.26e+01$	$3.47e-16$	$6.38e-16$	$1.09e-15$

Remarks 4.3. In (iv) we have no bound on $|\tilde{c}_0 - c_0|$, hence also no bound on $\hat{\varphi}_0$ and on $|\tilde{\delta}_0 - \delta_o|$, see (4.13), (3.8). In (i), (ii) $\hat{\varphi}_0$ can be larger than 0.01

but it is probably less than 0.1, hence δ_0 is marked with $*$ (in these cases from (3.2) follows that the relation $\xi_0' < 1$ is doubtful). But in (i), (ii), (iv) we have the bounds on $|\tilde{\delta}_0 - \hat{\delta}_0|$, see (3.8). In particular, if $\frac{1}{2}(U^T A_2 + A_2^T U) \in \mathcal{HPD}$, then the bound $|\mathtt{acc}(U, A_2) - 5.45 \times 10^{-9}| \lesssim 2.73 \times 10^{-11}$ holds. All experimental results illustrate well the theory presented in subsection 4.3. See in particular the relations (4.13), (4.16) (also in the case (iv)!).

4.5 Experiments with inverses *having only* the `Conj-Property`

We use here the special procedure `INVCONJ`(X). This procedure computes at first the `SVD` of the real matrix X:

$$X =: P * \Sigma * Q^T, \quad \Sigma = \mathrm{diag}(\sigma_i), \quad P, Q \text{ orthogonal},$$

and constructs then the *computed inverse* G of X:

$$G := Q * (\Sigma^{-1} + \mathbf{\Psi}) * P^T, \quad \mathbf{\Psi} = eps * \frac{\sigma_{\max}}{\sigma_{\min}} \Big[\frac{rd_{ij}}{\max\{\sigma_i, \sigma_j\}} \Big],$$

where $\{rd_{ij}\}$ are random numbers, $|rd_{ij}| \leqslant 1$.

G has evidently the `Conj-Property` (both norms: $||GX - I||_F, ||XG - I||_F$ are bounded with $(n\ eps + \varepsilon_x)xg$). But G has *only* the `Conj-Property` if $c = \sigma_{\max}/\sigma_{\min}$ is large and some σ_i are close to $\alpha \stackrel{\mathrm{df}}{=} \sqrt{\sigma_{\max}\sigma_{\min}}$. (Only in this case the `Conj-Property` differs distinctly from the `NC-Property`).

We will use here the optimal scaling: $\gamma_k := \hat{\alpha}_k^{-1}, \hat{\alpha}_k \stackrel{\mathrm{df}}{=} \sqrt{\hat{\sigma}_{\max}^{(k)} \hat{\sigma}_{\min}^{(k)}}$, where $\{\hat{\sigma}_i^{(k)}\}_{i=1}^n$ are singular values of $\tilde{X}_k$. We need in our experiment some iterates $\tilde{X}_k$ with large $\hat{c}_k \stackrel{\mathrm{df}}{=} \mathrm{cond}_2(\tilde{X}_k)$ and with several, say: m_k, singular values close to $\hat{\alpha}_k$. Choosing appropriate $\{\hat{\sigma}_i\}$ in the following construction:

$$A = \tilde{X}_0 := P * \mathrm{diag}(\hat{\sigma}_i) * Q^T, \quad P, Q \text{ orthogonal, random}, \tag{4.21}$$

we obtain: $\hat{c}_0$ large, $m_0 \geqslant 2$ and eventually also $\hat{c}_1$ large, $m_1 \geqslant 2$. Example F.2 in [12] explains why in real computations we must have $m_0 \geqslant 2$ (eventually also $m_1 \geqslant 2$). With $m_0 = 1$ the perturbation $G_0 - \tilde{X}_0^{-1}$ there could not produce the next iterate $\tilde{X}_1$ with orthogonal factor $\hat{U}_1$ distinctly different from $\hat{U}_0$ (the orthogonal factor of $\tilde{X}_0$). Hence δ_0 could not be much larger than $\delta_1 = \mathtt{acc}(U, X_1)$. This correlates with theorem 2.3 in [10].

In all presented below experiments the computed result $\tilde{U} = \tilde{X}_l$ cannot be accepted as a sufficiently good `AUF` of A.

Conclusion 4.2. `Conj-Property Only` is not a sufficiently good quality of the matrix-inversion in the `HS`-process.

Examples 4.2. In experiments presented in tables 2–4 all matrices $\tilde{H}$ passed the positivity test, all $e_k^{(\mathrm{L})}, e_k^{(\mathrm{R})}$ were less than 2.7×10^{-15}. We have generated matrices A as in (4.21).

(i) $n = 6, \hat{\sigma}_1 = \hat{\sigma}_6^{-1} := 10^7, \hat{\sigma}_2 = \hat{\sigma}_5^{-1} := 1000\sqrt{20}, \hat{\sigma}_3 = \hat{\sigma}_4 := 1, \tilde{\Delta}_6 = 5.76 \times 10^{-16}$

Table 2

k	c_k	$\sqrt{c_k}$	δ_k	m_k
0	$1.00e+14$	$1.00e+07$	$5.49e-10$	2
1	$5.06e+06$	$2.25e+03$	$1.01e-13$	2
2	$1.06e+03$	$3.26e+01$	$8.74e-16$	–
3	$1.01e+00$	$1.01e+00$	$5.91e-16$	–

(ii) $n = 20, \{\hat{\sigma}_k\} = \{10^{14}, 10^7, \ldots, 10^7, 1\}, \tilde{\Delta}_6 = 1.99 \times 10^{-15}$

Table 3

k	c_k	$\sqrt{c_k}$	δ_k	m_k
0	$9.99e+13$	$1.00e+07$	$7.04e-09$	18
1	$5.17e+06$	$2.27e+03$	$1.72e-15$	–
2	$1.07e+00$	$1.04e+00$	$1.74e-15$	–

(iii) $n = 20, \{\hat{\sigma}_i\}_{i=1}^{20} = \{q^{i-1}\}_{i=1}^{20}, q = 10^{14/19}, \tilde{\Delta}_8 = 1.87 * 10^{-15}$

Table 4

k	c_k	$\sqrt{c_k}$	$\tilde{\delta}_k$	m_k
0	$1.00e+14$	$1.00e+07$	$4.39e-10$	2
1	$3.61e+06$	$1.90e+03$	$1.31e-13$	2
2	$7.27e+02$	$8.50e+01$	$6.62e-15$	1
3	$1.35e+01$	$3.67e+00$	$2.10e-15$	–

5 The problem of too small scaling parameters

The good behaviour of the HS-process is equivalent to the smallness of all entries of the sequence $\{\delta_k\}_{k=0}^{l-1}, \delta_k = \mathtt{acc}(U, X_k)$ provided $H_k \in \mathcal{HPD}$. If all δ_k are at most of the order ε then the computed unitary factor $\tilde{U} = \tilde{X}_l$ is a good AUF of all matrices $\{X_k\}, \{\tilde{X}_k\}$ (provided $H_0 \in \mathcal{HPD}$) and - in the consequence - also of A, see lemma 2.1.

Assuming: $\xi'_k \ll 1, H_{k+1} \in \mathcal{HPD}$ and the *good matrix-inversion* (2.5), we can use the simplified form of the backward-induction:

$$\delta_k \approx |\delta_{k+1} + \varepsilon'_k| z_k, \quad |\varepsilon'_k| \leqslant \hat{\varepsilon}, \quad z_k \stackrel{\mathrm{df}}{=} \chi_k r_k, \quad \chi_k \leqslant 1$$

where $r_k = \max\{1, (c_k + \rho_k)/(c_k\rho_k + 1)\}, \quad \rho_k = (\gamma_k/\gamma_k^{(\mathrm{opt})})^2$. If $\gamma_k < \gamma_k^{(\mathrm{opt})}$ then $\rho_k < 1$ and $r_k > 1$. It can happen that also $z_k = \chi_k r_k > 1$ holds (though χ_k essentially decreases with ρ_k, see remark 2.5) and can result the *multiplicative increase (deterioration)* of δ_k: $\delta_k \approx |\delta_{k+1} + \varepsilon'_k| z_k > d_{k+1}$. This can continue if also $z_{k-1} = \chi_{k-1} r_{k-1} > 1$ holds.

That is the *problem of too small scaling parameters: the danger of a serious multiplicative deterioration of accuracy in the* HS*-process if we use the scaling parameters* $\{\gamma_k\}$ *smaller than the optimal ones.* The multipliers $\chi_k < 1$ can act here *soothingly*, but there is no guaranty that it implies always $z_k = \chi_k r_k \lesssim 1$ *or* yields at the end $\delta_0 = \mathtt{acc}(U, X_0) \approx \mathtt{acc}(U, A)$ sufficiently small. Our experiments seem to indicate that the HS-process with practical scaling ($n^{-1/2} \leqslant \rho_k \leqslant n^{1/2}$) is *immune* to this *danger* of such deterioration of accuracy. (We simulated also the computations on 100×100 matrices, *blowing-up* the *actual* ρ_k to the interval $[0.1, 10]$.) This phenomenon of the immunity of the practical scaling can be explained probably as follows.

Let us consider the approximate equality

$$\tilde{X}_{k+1} \approx B_k + F_k, \quad B_k \stackrel{\mathrm{df}}{=} \frac{1}{2}\tilde{X}_k\gamma_k, \quad F_k \stackrel{\mathrm{df}}{=} \frac{1}{2\gamma_k} G_k^H. \tag{5.1}$$

If $\gamma_k = \gamma^{(\mathrm{opt})}(\tilde{X}_k, G_k) \stackrel{\mathrm{df}}{=} \sqrt{||G_k||_2/||\tilde{X}_k||_2}$, $(\gamma_k \approx \gamma_k^{(\mathrm{opt})}, \rho_k \approx 1)$ then $||B_k||_2 = ||F_k||_2$ holds. Optimal scaling *equilibrates* the matrices B_k, F_k in (5.1) (in the sense of the 2-norm). If $\gamma_k < \gamma_k^{(\mathrm{opt})}$ then $\rho_k < 1$ and $||B_k||_2 \approx ||F_k||_2\rho_k$ holds; $||B_k||_2$ is by a factor ρ_k smaller than $||F_k||_2$.

In the assignment statement: $\tilde{X}_{k+1} := B_k + F_k$ the next iterate $\tilde{X}_{k+1}$ "obtains" practically full numerical information on G_k^H and only the "upper parts" of the entries $[\gamma_k\tilde{X}_k]_{(i,j)}$ of B_k rounded on the level $\nu_d|\tilde{X}_{k+1}|_{(i,j)}$. Therefore the matrix $\tilde{U}$ computed in the HS-process from $\tilde{X}_{k+1}$ can be in a

similar degree a good AUF for G_k^H, as it is for $\tilde{X}_{k+1}$, but can be a distinctly worse AUF for $\tilde{X}_k$ (by a factor up to ρ_k^{-1}?). This explains the presence of the multipliers r_k, $1 < r_k \lesssim \rho_k^{-1}$, in the basic backward -induction recursion: $\delta_k \approx |\delta_{k+1} + \varepsilon'_k|\chi_k r_k$. The factor $r_k > 1$ is here *a warning*, $\chi_k \leqslant 1$ can "soothe" it, if the information on $\tilde{X}_k$ delivered to $\tilde{X}_{k+1}$ (hence to $\tilde{U}$) was not so much disturbed as it could be supposed regarding only the formal quantity $\rho_k = (\gamma_k/\gamma_k^{(\mathrm{opt})})^2$.

Let us notice now that the (F)-scaling:

$$\gamma_k = \gamma^{(\mathrm{F})}(\tilde{X}_k, G_k) = \sqrt{\frac{||G_k||_F}{||\tilde{X}_k||_F}}$$

equilibrates the (F)-norms: $||B_k||_F = ||F_k||_F$ and the $(1,\infty)$-scaling: $\gamma_k = \gamma^{(1,\infty)}(\tilde{X}_k, G_k) = \sqrt{\ll G_k \gg / \ll \tilde{X}_k \gg}$ equilibrates the pseudo-norms: $\ll B_k \gg = \ll F_k \gg$, where $\ll B \gg \stackrel{\mathrm{df}}{=} \sqrt{||B||_1||B||_\infty}$. Hence both rules of practical scaling also equilibrate the matrices B_k, F_k in (5.1). There is no reason to believe that one of these three rules of equilibrating is always and distinctly better than the other two.

Working in our analysis with the 2-norm we must be prepared to obtain the eventual warning: $r_k > 1$, but if the equilibration of B_k, F_k was sufficiently good (that means: the matrix $\tilde{X}_{k+1}$, hence also $\tilde{U}$, obtained sufficient information on $\tilde{X}_k$), then the factor χ_k should act "soothingly", yielding $z_k = \chi_k r_k \lesssim 1$. That seems to explain the phenomenon of the good behaviour of the HS-process with practical scaling.

There remains *rather academic problem*: whether using too small scaling parameters can really "produce" a serious multiplicative deterioration of the accuracy? Some experiments presented below give a positive answer to this question.

Example 5.1. We are using here GECP-inversion and essentially practical or optimal scaling. Only in several initial steps of HS-process we apply *special-scaling*, choosing γ_k distinctly smaller than $\gamma^{(\mathrm{opt})}(\tilde{X}_k, G_k)$. This retards the decreasing of the sequence $\{c_k\}$ in those steps and spoils the quality of $\tilde{U}$ as an AUF of $\tilde{X}_k$ (hence also of $\tilde{X}_0$ and A). In (i) $A = A_1$, see subsection 4.5, in (ii) $A = A_5$ random and ill-conditioned 10×10 matrix, in (iii) A_6, specially constructed 10×10 matrix. All matrices $\tilde{H}$ passed the positivity-test. In tables 8–10 we present additionally the quantities $\{\tilde{\chi}_k\}$, $\tilde{\chi}_k := \delta_k/[r_k*(\delta_{k+1} + 10^{-16})]$, probably a lower bound on χ_k.

(i)

Table 8

k	c_k	ρ_k	r_k	δ_k	$\hat{\chi}_k$
0	$8.75e+14$	$8.27e-04$	$1.21e+03$	$5.82e-13$	0.078
1	$4.35e+08$	$1.19e-03$	$8.99e+02$	$6.09e-15$	0.036
2	$2.65e+05$	$1.11e-03$	$8.40e+02$	$1.90e-14$	0.026
3	$6.00e+03$	$9.44e-04$	$9.01e+02$	$7.96e-15$	0.041
4	$1.24e+03$	$1.12e+00$	$1.00e+00$	$1.16e-16$	0.431
5	$1.51e+01$	$9.26e-01$	$1.07e+00$	$1.69e-16$	0.720

(ii)

Table 9

k	c_k	ρ_k	r_k	δ_k	$\hat{\chi}_k$
0	$9.61e+14*$	$8.21e-05*$	$1.21e+04*$	$3.96e-13$	0.0013
1	$1.12e+09$	$1.12e+00$	1	$2.46e-14$	0.422
2	$1.17e+04$	$1.27e-04$	$5.13e+03$	$5.85e-14$	0.013
3	$5.17e+03$	$1.08e+00$	1	$6.71e-16$	0.647
4	$3.15e+01$	$3.25e-02$	$1.55e+01$	$8.36e-16$	0.154
5	$1.64e+01$	$1.37e+00$	1	$1.51e-16$	0.302

(iii)

Table 10

k	c_k	ρ_k	r_k	δ_k	$\hat{\chi}_k$
0	$1.00e+11$	0.01	100	$7.30e-14$	0.145
1	$1.58e+06$	0.01	100	$4.95e-15$	0.261
2	$6.29e+03$	1.43	1	$8.94e-17*$	0.281
3	$4.63e+01$	1.07	1	$2.18e-16$	0.522

Remarks 5.1. Experimental results above demonstrate the tendency of the quantities $\{\chi_k\}$ to decrease with $\{\rho_k\}$ and the "soothing" role of χ_k in the eventual blowing-up of $\{\delta_k\}$ in the backward induction. Example (iii) demonstrates the multiplicative blowing-up of $\{\delta_k\}$ in the last two steps of the backward induction. Special choice of the singular values of the matrix A_6 yields χ_0, χ_1 not very small.

Conclusion 5.1. Using in the `HS`-process the scaling parameters γ_k distinctly smaller than the optimal ones retards the convergence and spoils the quality of the computed `AUF` $\tilde{U}$ of A.

6 The switching criteria in HS

In [12] we presented two *switching criteria*: criterion for accepting the last constructed iterate $\tilde{X}_l$ as the computed unitary factor $\tilde{U} := \tilde{X}_l$ and criterion for switching from $(1,\infty)$-scaling to unscaled iterations. Their aim was to guarantee the monotonic decrease of the sequences $\{c_k\}_{k=0}^{l}, \{\sigma_{\max}^{(k)}\}_{k=1}^{l}$ and to yield the last iterate achieving the limiting accuracy $\tilde{\Delta}_l := ||\tilde{X}_l^H * \tilde{X}_l - I||_F \leqslant \varepsilon$. In our experiments (aimed essentially to study the problems of sections 4 or 5) we tested additionally these criteria, comparing their performance with the performance of the corresponding criteria for $(1,\infty)$-scaling or (F)-scaling, see [6], [11]. More than sixty HS-test-processes supplied appropriate experimental data (none special experiments for testing our criteria were performed).

Our criteria depend on the sequence $\{\beta_k\}_{k=1}^{l}$, where the quantity $\beta_k := ||\tilde{X}_k - G_k^H||_F (1 \leqslant k \leqslant l)$ is computed in the kth iteration in the same loop in which the next iterate $\tilde{X}_{k+1}$ is constructed (compare lemma 2.1 in [11]). In appendix B [12] we show that: $e_k \stackrel{\mathrm{df}}{=} \sigma_{\max}^{(k)} - 1 > \varepsilon$ implies $\beta_k \approx 2p_k e_k, p_k \stackrel{\mathrm{df}}{=} p(\tilde{X}_k - G_k^H)$.

The stopping criterion. We accept $\tilde{X}_{k+1}$ as the last computed iterate, $l = l_N \stackrel{\mathrm{df}}{=} k+1$, provided the following both conditions hold: the kth iteration is unscaled or optimally or (F)-scaled and $\beta_k \leqslant \sqrt{2\nu n^{1/2}}$. In all our experiments $\tilde{X}_l$ with $l = l_N$ was achieving acceptable limiting accuracy: $\tilde{\Delta}_l \leqslant \varepsilon$.

Let l_H and l_F be the indices of the last iterates indicated, respectively, for unscaled [6] and (F)-scaled [11] iterations ($k > 0$, $x_{k+1} \geqslant x_k$ implies $l_F \stackrel{\mathrm{df}}{=} k$, where $x_p := ||\tilde{X}_p||_F$ for $p = k, k+1$). In all our experiments we noticed the relations: $l_N \leqslant l_H \leqslant l_N + 2$, $l_F \leqslant l_N \leqslant l_F + 1$. The relation $l_H = l_N + 1$ was noticed frequently, the relation $l_H = l_N + 2$ was noticed twice. The relation $l_N = l_F + 1$ was noticed once. The HS-process with stopping criterion [6] is performing frequently (at least) one redundant iteration.

The criterion for switching to unscaled iterations. We propose to switch to unscaled iterations for $k > r = r_N$ provided the following both conditions hold: the iterations for $k \leqslant r_N$ are $(1,\infty)$-scaled and $\beta_k \leqslant 1.5$ or $\beta_k \geqslant \beta_{k-1}$.

Let r_H be the index of the last $(1,\infty)$-scaled iteration according to the criterion in [6]. The relations observed in our experiments are: $r_N < r_H \leqslant r_N + 4$. Frequently $r_H = r_N + 3$ or $r_H = r_N + 2$. The case $r_H = r_N + 4$ was noticed once. Our proposal of an earlier switch to unscaled iterations is a result of *cautiousness*. Assuming the bound (2.19) on τ_k we can not

guarantee the monotonic decrease of the sequences $\{c_k\}, \{\sigma_{\max}^{(k)}\}$ in the HS-process with $(1, \infty)$-scaling. But in all observed cases (of HS with $(1, \infty)$-scaling) the quantities τ_k with $r_N < k \leqslant r_H$ were distinctly smaller than the bound (2.19) , hence the mentioned sequences were nicely decreasing (faster than for unscaled iterations). The criterion in [6] performed better than ours. This competition between our *cautious criterion* and (probably) the *risky criterion* in [6] remains hence unsolved. We have no experimental proof that this criterion is too risky.

Conclusion 6.1. Our stopping criterion is performing better than the criterion in [6] and equally well as the criterion for (F)-scaling. Our switching criterion to unscaled iterations is performing correctly (guarantees monotonic convergence), but is probably unnecessarily too cautious. Replacing the condition $\beta_k \leqslant 1.5$ with, say, $\beta_k \leqslant 0.5$ could sometimes reduce the number of needed iterations by one.

7 Final conclusions

We now summarize conclusions.

(i) Matrix-inversion in the HS-process should yield the computed inverse G of the matrix X (the inverse of the current iterate) satisfying the condition (4.5) (the NC-Property). This property is warranted by the inversion *via* GECP-triangularization of X. Using the standard inversion via GEPP, see [4], can fail, yielding for some special matrices A a poor unitary factor $\tilde{U}$. This will never occur for well-conditioned matrices A, say: $\mathrm{cond}_2(A) \leqslant 10^2$.

(ii) Using in the HS-process a good matrix-inversion, see (i), and either the (F)-scaling [11] or the $(1, \infty)$-scaling [6] (with appropriate switch to unscaled iterations) practically guarantees good quality of the computed unitary factor $\tilde{U}$ of A. If $\varepsilon \mathrm{cond}_2(A) < 1$ holds then $\tilde{U}$ has similar quality as the factor computed *via* SVD of A.

(iii) An appropriate stopping criterion, see section 6, in most cases guarantees that $\tilde{U} = \tilde{X}_l$ is the first iterate reaching the limiting accuracy. With the stopping-criterion in [6] frequently one redundant step is performed.

(iv) The formal cost (the number of arithmetic operations) of the HS-process in the standard double precision is at most of the same order as for

`SVD` (is smaller for well-conditioned matrices or matrices with large gaps in the spectrum of the singular values).

(v) Using in the `HS`-process scaling parameters $\{\gamma_k\}$ distinctly larger or smaller than the optimal-ones, see (2.12), can spoil the convergence. Using $\{\gamma_k\}$ distinctly smaller is spoiling also the quality of $\tilde{U}$ as an approximate unitary factor of A. Practical scaling, see (ii), is not involving such impendency.

References

[1] R. Bhatia, *Matrix Analysis* (Springer, Berlin, 1996).

[2] A. Barrlund, Perturbation bounds on the polar decomposition, BIT 31 (1990) 101–113.

[3] Å. Björck, C. Bowie, An iterative algorithm for computing the best estimate of an orthogonal matrix, SIAM J. Numer. Anal. 8 (1971) 358–364.

[4] J.J. Du Croz, N.J. Higham, Stability of methods for matrix inversion, IMA J. Numer. Anal. 12 (1992) 1–19.

[5] W. Gander, Algorithms for the polar decomposition, SIAM J. Sci. Stat. Comput. 11 (1990) 1102–1115.

[6] N.J. Higham, Computing the polar decomposition - with applications, SIAM J. Sci. Stat. Comput. 7 (1986) 1160–1174.

[7] N.J. Higham, P. Papadimitriou, A parallel algorithm for computing the polar decomposition, Parallel Comput. 20 (1994) 1161-1173.

[8] N.J. Higham, *Accuracy and Stability of Numerical Algorithms* (SIAM, Philadelphia, 1996).

[9] N.J. Higham, R.S. Schreiber, Fast polar decomposition of an arbitrary matrix, SIAM J. Sci. Stat. Comput. 11 (1990) 648–655.

[10] Ch. Kenney, A.J. Laub, Polar decomposition and matrix sign function condition estimates, SIAM J. Sci. Stat. Comput. 12 (1991) 488–504.

[11] Ch. Kenney, A.J. Laub, On scaling Newton's method for polar decomposition and the matrix sign function, SIAM J. Matrix Anal. Appl. 13 (1992) 688–706.

[12] A. Kiełbasiński, K. Ziętak, Numerical behaviour of Higham's scaled method for polar decomposition, Numerical Algorithms 32 (2003), 105–140.

[13] G.W. Stewart, The triangular matrices of Gaussian elimination and related decomposition, IMA J. Numer. Anal. 17 (1997) 7–16.

[14] J.H. Wilkinson, *Rounding Errors in Algebraic Process* (Her Majesty's Stationery Office, London, 1963).

[15] P. Zieliński, K. Ziętak, The polar decomposition – properties, applications and algorithms, Applied Mathematics, Annals of Polish Math. Soc. 38 (1995) 23–49.