

INSTABILITY OF THE SHERMAN-MORRISON FORMULA AND STABILIZATION BY ITERATIVE REFINEMENT

BEHNAM HASHEMI* AND YUJI NAKATSUKASA†

Abstract. Owing to its simplicity and efficiency, the Sherman-Morrison (SM) formula has seen widespread use across various scientific and engineering applications for solving rank-one perturbed linear systems of the form $(A + uv^T)x = b$. Although the formula dates back at least to 1944, its numerical stability properties have remained an open question and continue to be a topic of current research. We analyze the backward stability of the SM, demonstrate its instability in a scenario increasingly common in scientific computing and address an open question posed by Nick Higham on the proportionality of the backward error bound to the condition number of A . We then incorporate fixed-precision iterative refinement into the SM framework reusing the previously computed decompositions and prove that, under reasonable assumptions, it achieves backward stability without sacrificing the efficiency of the SM formula. While our theory does not prove the SM formula with iterative refinement always outputs a backward stable solution, empirically it is observed to eventually produce a backward stable solution in all our numerical experiments. We conjecture that with iterative refinement, the SM formula yields a backward stable solution provided that $\kappa_2(A), \kappa_2(A + uv^T)$ are both bounded safely away from ϵ_M^{-1} , where ϵ_M is the unit roundoff.

Key words. Sherman-Morrison formula; backward stability; iterative refinement.

AMS subject classifications. 65Gxx, 65F55

1. Introduction. The Sherman–Morrison–Woodbury (SMW) formula

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^TA^{-1}U)^{-1}V^TA^{-1}$$

has been used widely in scientific computing. Here U, V are $n \times r$ with $r \leq n$, usually $r \ll n$. When $r = 1$, it is often called the SM formula. Its primary usage is for solving linear systems of equations that undergo a low-rank update, as solving $(A + UV^T)x = b$ can be done efficiently if linear systems with respect to A can be solved efficiently, e.g. when a factorization of A is known or A is sparse or has other structures that can be exploited for efficient solution. We refer the readers to [10] for an overview of the key ideas related to the SMW formula, historical developments [4, 25, 2, 20] and its many applications from ordinary and partial differential equations [6, 18] to networks and optimization. It can also be extended to least-squares problems [9].

Despite its widespread use, the numerical stability of SMW has remained an open problem. Yip’s 1986 paper [26] remains the only reference to our knowledge to focus on its stability. Some more comments are in Hager [10] and Hao and Simoncini [12] etc. In these papers, the focus is on the conditioning of the matrix inversion $(I + V^TA^{-1}U)^{-1}$. Indeed, a poor choice of U (note that $UV^T = UM^{-1}MV^T$ for any nonsingular M) results in an ill-conditioned $(I + V^TA^{-1}U)$, and it is easy to see that the resulting use of SMW gives poor results.

Yip shows that $\kappa_2(I + V^TA^{-1}U) \leq \kappa_2(A)\kappa_2(A + UV^T)$ by choosing U, V such that U (or V) has orthonormal columns; this can be done simply by taking $M = R$ above, where $U = QR$ is the thin QR factorization. It follows that, if $A, A + UV^T$ are both well-conditioned, then it is easy to ensure the inversion of $I + V^TA^{-1}U$ is well-conditioned. Moreover, assuming A^{-1} (or linear systems with A) are computed in a

*School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, UK. b.hashemi@le.ac.uk.

†Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK. Supported by EPSRC grants EP/Y010086/1 and EP/Y030990/1. nakatsukasa@maths.ox.ac.uk.

backward stable fashion in each step of the SMW formula application, we conclude that the forward error is small, and hence so is the backward error. That is, SMW can be implemented in a stable manner if both A and $A + UV^T$ are well conditioned. In [14, Problem 26.2] Higham notes that examples can be found where SM is not backward stable, and that the backward error in practice appears to be bounded by $\mathcal{O}(\epsilon_M \kappa_\infty(A))$.

Ma, Boutsikas, Ghadiri and Drineas [16] recently studied the stability of the SMW formula deriving both backward and forward error bounds. Their analysis focuses on the errors in two steps: computing the inverse of A , and the inverse of the capacitance matrix $Z = I + V^T A^{-1} U$ whose size is smaller than that of A . Essentially, they assume that all other operations are performed in exact arithmetic and formulate sufficient conditions imposing bounds on the amount of forward error in computing the aforementioned matrix inverses, and on $\|A\|$ and $\|U\| \|V\|$ and $\|Z\|$ so as to bound the forward error and backward error. Their work studies the SMW formula's stability in terms of matrix inversion, not when used for linear systems; this is a different problem, and the stability (or otherwise) in one problem does not immediately imply (in)stability of the other¹. We also note that a classical work by Govaerts [8] examines the backward error of a $(n+1) \times (n+1)$ bordered linear system $\begin{bmatrix} A & u \\ v^T & -1 \end{bmatrix} \begin{bmatrix} x \\ \zeta \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$, which is related to the SM formula, and derives the backward error for this system. It is not entirely clear how the result can be translated into a backward error with respect to $A + uv^T$. We describe Govaerts' work in more detail in the appendix.

In this work we examine the backward stability of solving linear systems using the SM(W) formula. We first identify a source of instability. Perhaps contrary to intuition and most prior studies, it has nothing to do with the conditioning of $I + V^T A^{-1} U$; it is unstable even when $r = 1$. We then focus on the $r = 1$ case and carefully examine the backward error of solving

$$(A + uv^T)x = b \tag{1.1}$$

via the SM(W) formula, i.e., $x = A^{-1}b - A^{-1}u(1 + v^T A^{-1}u)^{-1}v^T A^{-1}b$. For definiteness we present the implementation in Algorithm 1.

Algorithm 1 SM: Sherman-Morrison formula for solving $(A + uv^T)x = b$. It is assumed that linear systems with A can be solved efficiently.

- 1: Solve $Ay = b$ for y to obtain $\hat{y}$.
 - 2: Solve $Az = u$ for z to obtain $\hat{z}$.
 - 3: Compute $\alpha = v^T \hat{y}$ and $\beta = 1 + v^T \hat{z}$, and let $\theta = \alpha/\beta$.
 - 4: Output $\hat{x} = \hat{y} - \theta \hat{z}$.
-

We show that under some conditions, a question raised by Nick Higham [14, p. 570] can be answered in the affirmative, namely that the backward error of the SM formula can be bounded by $\epsilon_M \kappa(A)$.

All these point towards a cautionary note: the SM(W) formula, while very attractive for computational efficiency, can be dangerously unstable. Fortunately, we suggest a simple remedy to the instability: iterative refinement (IR). IR is a classical idea in NLA but has witnessed a resurged interest, partly due to the growing interest

¹See the appendix for a more detailed discussion.

and availability of mixed precision arithmetic [3], but perhaps equally importantly, as a tool to correct or improve a fast but unstable algorithm [5].

Despite the widespread use of SM(W) and IR, the combination appears to not have been used in the literature. The key contribution of this paper is the introduction and analysis of iterative refinement (IR) for the SM formula. With IR we find a correction to a (possibly inaccurate) solution $\hat{x}$ obtained by SM as follows: compute the residual $r = b - (A + uv^T)\hat{x}$, and solve $(A + uv^T)\delta x = r$ for the correction δx using SM again, and output $x + \delta x$ as the solution. The resulting algorithm SM-IR is displayed in Algorithm 2. It is worth noting that SM-IR has the same complexity as SM; it only doubles the cost. If necessary, one can repeat IR more than once, which can further improve the accuracy.

Algorithm 2 SM-IR: Sherman-Morrison formula with iterative refinement for solving $(A + uv^T)x = b$.

- 1: Compute $\hat{x}$ using SM.
 - 2: Compute $r = b - A\hat{x} - (v^T\hat{x})u$.
 - 3: Solve $Ay_r = r$ for y_r to obtain $\hat{y}_r$.
 - 4: Compute $\theta_r := \alpha_r/\beta$ where $\alpha_r = v^Ty_r$ and β same as in Step 1.
 - 5: Output $\hat{w} = \hat{x} + y_r - \theta_r z$.
 - 6: If the accuracy is insufficient with $\hat{w}$, set $\hat{x} := \hat{w}$ and return to Step 2.
-

It is worth noting that IR is often used together with mixed precision. For example, the classical analysis and usage of IR assumed higher-precision arithmetic is used for the computation of the residual [14, Ch. 12]. More recently, Carson and Higham [3] study IR using multiple precisions. In this work, we focus on the simplest IR where fixed precision (e.g. double precision) is used throughout; mixed precision, while possibly helpful, is not observed to be necessary empirically, and our analysis assumes the same precision is used throughout. More generally, IR is often able to improve the accuracy of an unstable solution, even in fixed precision [15, 21].

We analyze the stability of SM-IR with one IR step, and show that under some mild conditions that can be verified on the fly, the algorithm outputs a backward stable solution $\hat{w}$, that is, one that satisfies $(A + uv^T + \Delta A)\hat{w} = b$ for some $\|\Delta A\| = \mathcal{O}(\epsilon_M \|A + uv^T\|)$. That is, while the SM formula is fundamentally unstable, the instability can be fixed with IR under mild assumptions.

However, the assumptions are not always satisfied, and we are currently unable to definitively prove that SM-IR will eventually (after a modest number of IR steps) produce a backward stable solution for any linear system. Despite this, our numerical experiments suggest strongly that this is the case. We hence conjecture that SM-IR is backward stable under very mild conditions, namely that both A and $A + uv^T$ have condition number bounded by ϵ_M^{-1} . Another open problem is to deal with a higher-rank case $r > 1$, that is, the SMW formula. One can clearly perform iterative refinement here, resulting in SMW-IR. An important open problem is to address the stability of SMW-IR.

Notation. Following Higham [14], quantities computed in floating-point arithmetic are denoted with a hat. To avoid confusing notation (to reserve u for the vector in the updated matrix $A + uv^T$), we denote unit roundoff by ϵ_M . Usually in stability analysis we suppress terms that grow like a modest polynomial in the matrix dimensions, so the choice of matrix norm is unimportant; for this reason we use $\|\cdot\|$ in most cases. Capital letters denote matrices, and vectors are represented by w, x, y and z .

Other lower-case letters denote scalars. When the norm choice matters or improves clarity, we use $\|\cdot\|_2$ for the spectral norm of a matrix. $\kappa(A) = \|A\| \|A^{-1}\|$ denotes the condition number of A .

2. A source of instability of the SMW formula. Here we identify a situation where the SMW formula would naturally lead to a large error, and is expected to be unstable.

To simplify the discussion here we assume $\|A + UV^T\|_2 = 1$ and $\|b\|_2 = 1$ without loss of generality, and also that $\|A\|_2 = \mathcal{O}(1)$, which is a nontrivial assumption. Here are the remaining assumptions:

1. *A is ill-conditioned:* $\kappa_2(A) \gg 1$. We make no assumptions on $\kappa_2(A + UV^T)$.
2. *Small-norm solution:* The linear system $(A + UV^T)x = b$ has a solution $\|x\|_2 = \mathcal{O}(1)$. By contrast, $Ay = b$ has solution $\|y\| = \mathcal{O}(\kappa_2(A)) \gg 1$.

The second assumption merits further explanation. It is essentially an assumption on the right-hand side b : if we write $b = \sum_{i=1}^n c_i v_i$ where $A + UV^T = \sum_{i=1}^n \sigma_i u_i v_i^T$ is the singular value decomposition, then the assumption is requiring that $|c_i| = \mathcal{O}(\sigma_i)$. We refer to this as a small-norm assumption, and x is a small-norm solution. We note that such situations are increasingly common in scientific computing: (i) in inverse problems, this is a standard phenomenon, and such assumptions are commonly made in the analysis, and called the *Picard condition* [11, §1.2.3]. (ii) In numerical computation with *frames* [1], it is an assumption made (and satisfied) in order to establish convergence and good numerical behavior of a numerically computed solution for approximation of functions in a domain. (iii) Indirectly related: the prominence of stochastic gradient descent comes (at least partially) from its ability to find a small-norm approximate solution to an (often highly underdetermined) optimization problem [13].

More generally, one could argue that huge-norm solutions to ill-conditioned problems are unlikely to have practical/physical relevance; for example given a solution x with $\|x\| = 10^{16}$, the task of computing $x^T y$ for some y is difficult to do with accuracy better than $\mathcal{O}(1)$ in double-precision arithmetic.

For these reasons we regard the assumption (ii), while certainly a nontrivial one, as quite natural and an important case in practice. In experiments, one can form such b by simply taking a randomly generated (e.g. Gaussian) vector x and setting $b = (A + UV^T)x$. By contrast, if b is taken randomly, one gets the opposite situation² where $\|x\|_2 = \mathcal{O}(\kappa_2(A + UV^T))$. Gill, Wright, Murray [7, p. 102] describe how such small-norm solutions should arise naturally in practical problems.

Let us explain where the instability comes from. Recall that $x = (A + UV^T)^{-1}b = A^{-1}b - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1}b = y - \theta z$. In any implementation of the SMW formula (including Algorithm 1), one would first compute y , then find θ, z (using y). Then assuming a backward stable solver is used, with the computed $\hat{y}$ we have $(A + \Delta A)\hat{y} = b$, where $\|\Delta A\|/\|A\| = \epsilon$. By Neumann series this implies $\hat{y} - y = A^{-1}\Delta A A^{-1}b + \mathcal{O}(\|\Delta A\|^2)$. We next compute θ and z . Even if this is done exactly or with high relative accuracy (which in general we do not expect), the result will have error at least $\|\hat{x} - x\| \gtrsim \|\hat{y} - y\| = \|A^{-1}\Delta A A^{-1}b\| = \|A^{-1}\Delta A y\| = \mathcal{O}(\epsilon \kappa_2(A)\|y\|) = \mathcal{O}(\epsilon(\kappa_2(A))^2)$, by the assumption $\|y\| = \mathcal{O}(\kappa_2(A))$.

It remains to see that this is indeed not a backward stable solution. Note that by definition a backward stable solution $\hat{x}$ needs to satisfy $(A + UV^T + \epsilon)\hat{x} = b$, so the residual is $\|(A + UV^T)\hat{x} - b\| = \|\epsilon \hat{x}\| \leq \|\epsilon\| \|\hat{x}\| = \mathcal{O}(\epsilon_M)$, by the assumption that

²These statements hold with (exponentially) high probability.

$\|x\| = \mathcal{O}(1)$; which implies also that $\|\hat{x}\| = \mathcal{O}(1)$, as $\|(A + UV^T + \epsilon)(x - \hat{x})\| = \|\epsilon x\| = \mathcal{O}(\epsilon_M)$, and so $\|x - \hat{x}\| \leq \mathcal{O}(\epsilon_M / \sigma_{\min}(A + UV^T + \epsilon)) = \mathcal{O}(1)$; Here we are tacitly assuming $\epsilon_M \kappa_2(A) < 1$.

So if $\hat{x}$ is backward stable we need $\|\hat{x}\| = \mathcal{O}(1)$; however, $\hat{x}$ is computed as the difference $\hat{y} - \hat{z}$, where $\|\hat{y}\|_2 = \mathcal{O}(\kappa_2(A))$, so simply by representing $\hat{y}$ we incur error $\mathcal{O}(\epsilon_M \kappa_2(A))$. Thus $x - \hat{x}$ is at least $\mathcal{O}(\epsilon_M \kappa_2(A))$, and so the residual is

$$\|(A + UV^T)\hat{x} - b\| = \|(A + UV^T)(\hat{x} - x)\| \gtrsim \epsilon_M \|A + UV^T\| \kappa_2(A) = \mathcal{O}(\epsilon_M \kappa_2(A)),$$

where the estimate assumes that the error $\hat{x} - x$ is unstructured with respect to the singular vector expansion of $A + UV^T$. Note that we also derive an upper bound for the residual in (3.12), which (under the current assumption that $\kappa_2(A) \gg \kappa_2(A + UV^T)$ in addition to further assumptions set out there) shows that $\mathcal{O}(\epsilon_M \kappa_2(A))$ is a tight estimate of the residual; which also matches Higham's conjecture that the SM backward error is bounded from above by $\epsilon_M \kappa_2(A)$. Thus the backward error of $\hat{x}$ as a solution for $(A + UV^T)x = b$ is at least $\epsilon_M \kappa_2(A + UV^T)$, and so it is not a backward stable solution.

3. Higham's open problem: Bounding backward error in SM. In the remainder of the paper we restrict ourselves to the $r = 1$ case and consider the stability of the SM (not SMW) formula for solving $(A + uv^T)x = b$.

An open problem in Higham's magnum opus on numerical stability [14, p. 570] is whether SM has a backward error bound proportional to $\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$.

Our plan in what follows is to first derive a neat representation for the SM residual, bound it, and then invoke the Rigal-Gaches formula [19] for the backward error [14, Lem. 1.1]

$$\eta(\hat{x}) = \frac{\|r\|}{\|B\|\|\hat{x}\| + \|b\|}. \quad (3.1)$$

To bound the residual, we consider the solution obtained by SM as

$$\hat{x} := fl(y - \frac{\alpha}{\beta}z) = \hat{y} \ominus \frac{\alpha}{\beta}\hat{z} \quad (3.2)$$

where $\alpha = v^T \hat{y}$, and $\beta = 1 + v^T \hat{z}$. Here, a backward stable solver is used to compute $\hat{y}$ and $\hat{z}$ as solutions to $Ay = b$ and $Az = u$, respectively. Therefore,

$$(A + \Delta_1)\hat{y} = b, \quad \text{where } \|\Delta_1\| \leq c_1 \epsilon_M \|A\| \quad (3.3)$$

where c_1 is a modest multiple of 1; see [24, p. 171] or [23, Thm. 16.2]. Similarly,

$$(A + \Delta_2)\hat{z} = u, \quad \text{where } \|\Delta_2\| \leq c_2 \epsilon_M \|A\|. \quad (3.4)$$

We know from the standard floating point arithmetic model that for each $1 \leq i \leq n$ we have $\hat{y}_i \ominus \frac{\alpha}{\beta} \hat{z}_i = (\hat{y}_i - \frac{\alpha}{\beta} \hat{z}_i)(1 + \delta_i)$ where $|\delta_i| \leq \epsilon_M$. So, the above representation gives³

$$\hat{x} = (\hat{y} - \frac{\alpha}{\beta} \hat{z}) + \delta \hat{x} \quad (3.5)$$

³Here again we are focusing on the error in the subtraction and assume that the errors in $\hat{\alpha}, \hat{\beta}$, their division and multiplication with $\hat{z}_i$ are insignificant.

where $\delta\hat{x}$ is the vector whose entries are $(\hat{y}_i - \frac{\alpha}{\beta}\hat{z}_i)\delta_i$, i.e.,

$$|\delta\hat{x}| \leq \epsilon_M |\hat{y} - \frac{\alpha}{\beta}\hat{z}| \quad (3.6)$$

With this we now turn to r . We have

$$\begin{aligned} r &= b - B\hat{x} = b - (A + uv^T)(\hat{y} \ominus \frac{\alpha}{\beta}\hat{z}) = b - (A + uv^T)(\hat{y} - \frac{\alpha}{\beta}\hat{z} + \delta\hat{x}) \\ &= b - A\hat{y} + \frac{\alpha}{\beta}A\hat{z} - uv^T\hat{y} + \frac{\alpha}{\beta}(v^T\hat{z})u - (A + uv^T)\delta\hat{x} \\ &= b - A\hat{y} + \frac{\alpha}{\beta}\left(A\hat{z} + (v^T\hat{z})u\right) - uv^T\hat{y} - (A + uv^T)\delta\hat{x} \quad \text{and (3.4) yields } A\hat{z} = u - \Delta_2\hat{z} \\ &= \Delta_1\hat{y} + \frac{\alpha}{\beta}\left(u - \Delta_2\hat{z} + (v^T\hat{z})u\right) - uv^T\hat{y} - (A + uv^T)\delta\hat{x} \quad \text{where with (3.3) } b - A\hat{y} = \Delta_1\hat{y} \\ &= \Delta_1\hat{y} - \frac{\alpha}{\beta}(\Delta_2\hat{z}) + u\frac{\alpha}{\beta}\left(\underbrace{1 + (v^T\hat{z})}_{\beta}\right) - uv^T\hat{y} - (A + uv^T)\delta\hat{x} \\ &= \Delta_1\hat{y} - \frac{\alpha}{\beta}(\Delta_2\hat{z}) + u\alpha - uv^T\hat{y} - (A + uv^T)\delta\hat{x} \quad (\text{as } u\alpha - uv^T\hat{y} = 0) \\ &= \Delta_1\hat{y} - \frac{\alpha}{\beta}(\Delta_2\hat{z}) - (A + uv^T)\delta\hat{x} \end{aligned}$$

Using the bounds in (3.3), (3.4) and (3.6) yields

$$\begin{aligned} \|r\| &\leq \|\Delta_1\|\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\Delta_2\|\|\hat{z}\| + \|A + uv^T\|\|\delta\hat{x}\| \\ &\leq c \epsilon_M \|A\| \left(\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\hat{z}\|\right) + \epsilon_M \|A + uv^T\| \|\hat{y} - \frac{\alpha}{\beta}\hat{z}\| \\ &\leq c \epsilon_M \|A\| \left(\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\hat{z}\|\right) + \epsilon_M \|A + uv^T\| (\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\hat{z}\|) \\ &= \epsilon_M (c \|A\| + \|A + uv^T\|) \left(\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\hat{z}\|\right) \end{aligned} \quad (3.7)$$

where $c = \max\{c_1, c_2\}$ is again a modest multiple of 1.

It turns out that the last term $\|\hat{y}\| + \left|\frac{\alpha}{\beta}\right|\|\hat{z}\|$ plays a crucial role in the stability analysis of SM. It represents the (absolute) condition number of the subtraction of $\frac{\alpha}{\beta}\hat{z}$ from $\hat{y}$ when forming $\hat{x}$. We now seek a bound on this condition number as it tells us where cancellation may occur in SM. So, we now focus on finding upper bounds for $\|\hat{y}\|$ and $\left|\frac{\alpha}{\beta}\right|\|\hat{z}\|$. It is worth noting that these quantities are easy to compute on the fly; in particular if $\left|\frac{\alpha}{\beta}\right|\|\hat{z}\|$ is nicely bounded by $\max(\|A^{-1}\|_2, \|(A + uv^T)^{-1}\|_2)$, then it follows that the residual is bounded by $\|b\|_2$, that is, the SM formula gives a solution that at least has a residual smaller than that of $x = 0$. One can then employ iterative refinement (possibly more than one step) to reduce the residual until we arrive at a backward stable solution. Since $\|\hat{z}\|$ depends on the right-hand side the argument is not a trivial matter of repeating the same analysis. We study iterative refinement in detail in the next section.

LEMMA 3.1. *Assume that the errors in α, β , their division and multiplication with $\hat{z}_i$ are insignificant. If*

$$|v^T\hat{z}| > 1.1 \quad (3.8)$$

then

$$\|\hat{y}\| + \left| \frac{\alpha}{\beta} \right| \|\hat{z}\| \leq \check{c} \|\hat{y}\| \quad (3.9)$$

where $\check{c}$ is a constant.

Proof. We start by examining the size of $\frac{1}{\beta} = \frac{1}{1+v^T \hat{z}}$ where $\hat{z} = fl(A^{-1}u)$. We have

$$\frac{1}{1+v^T \hat{z}} = \frac{1}{v^T \hat{z}((v^T \hat{z})^{-1} + 1)} = \frac{1}{v^T \hat{z}} \frac{1}{1 + \zeta}$$

where from (3.8), it follows that $\zeta := (v^T \hat{z})^{-1}$ has a magnitude less than 1 hence $\frac{1}{1+\zeta} = 1 - \zeta + \zeta^2 - \dots$ yields

$$\frac{1}{1+v^T \hat{z}} = \frac{1}{v^T \hat{z}} \left(1 - \frac{1}{v^T \hat{z}} + \frac{1}{(v^T \hat{z})^2} - \dots \right) = \frac{1}{v^T \hat{z}} - \frac{1}{(v^T \hat{z})^2} + \frac{1}{(v^T \hat{z})^3} - \dots$$

and so

$$\frac{\alpha}{\beta} = \frac{v^T \hat{y}}{1+v^T \hat{z}} = \frac{v^T \hat{y}}{v^T \hat{z}} - \frac{v^T \hat{y}}{(v^T \hat{z})^2} + \frac{v^T \hat{y}}{(v^T \hat{z})^3} - \dots \quad (3.10)$$

Thus,

$$\begin{aligned} \left| \frac{\alpha}{\beta} \right| &\leq \frac{|v^T \hat{y}|}{|v^T \hat{z}|} (1 + |\zeta| + |\zeta|^2 + \dots) \\ &= \frac{1}{1-|\zeta|} \frac{\|v\| \|\hat{y}\| |\cos(\theta_{v,\hat{y}})|}{\|v\| \|\hat{z}\| |\cos(\theta_{v,\hat{z}})|} = \frac{1}{1-|\zeta|} \frac{\|\hat{y}\| |\cos(\theta_{v,\hat{y}})|}{\|\hat{z}\| |\cos(\theta_{v,\hat{z}})|} \end{aligned} \quad (3.11)$$

where $\theta_{v,\hat{y}}$ and $\theta_{v,\hat{z}}$ denote the angle between v and $\hat{y}$ and between v and $\hat{z}$, respectively. Multiplying both sides by $\|\hat{z}\| > 0$ gives

$$\left| \frac{\alpha}{\beta} \right| \|\hat{z}\| \leq \frac{1}{1-|\zeta|} \frac{|\cos(\theta_{v,\hat{y}})|}{|\cos(\theta_{v,\hat{z}})|} \|\hat{y}\|.$$

We thus obtain (3.9) with $\check{c} = 1 + \frac{1}{1-|\zeta|} \frac{|\cos(\theta_{v,\hat{y}})|}{|\cos(\theta_{v,\hat{z}})|}$. $\square$

The following result is an answer to Higham's question on the proportionality of the SM backward error with the condition number of A .

PROPOSITION 3.2. *Assume $Ay = b$ is solved with a normwise backward stable algorithm. Under the assumptions of the preceding lemma, we have*

$$\|r\| \leq \epsilon_M \frac{\check{c}}{1 - c_1 \epsilon_M \kappa(A)} \left(c \|A\| + \|A + uv^T\| \right) \|A^{-1}\| \|b\| + \mathcal{O}(\epsilon_M^2) \quad (3.12)$$

where c_1, c are modest multiples of 1, $c_1 \epsilon_M \kappa(A) < 1$, and $\check{c}$ is as in Lemma 3.1.

Proof. The previous lemma together with (3.7) give

$$\|r\| \leq \epsilon_M \check{c} (c \|A\| + \|A + uv^T\|) \|\hat{y}\| + \mathcal{O}(\epsilon_M^2)$$

In addition, from (3.3), $c_1 \epsilon_M \kappa(A) < 1$ and standard perturbation theory [22, p. 54]

we can show that⁴

$$\|\hat{y}\| \leq \frac{\|A^{-1}\|}{1 - c_1 \epsilon_M \kappa(A)} \|b\| \quad (3.13)$$

from which (3.12) follows. $\square$

This result shows that the backward error of the SM formula is bounded by $\mathcal{O}(\max(\kappa_2(A), \kappa_2(A + uv^T))\epsilon_M)$, if the assumptions hold. We emphasize that the assumptions, while verifiable on the fly, are nontrivial and not always true. It therefore remains an open problem to address whether the backward error of SM can always be bounded by $\mathcal{O}(\max(\kappa_2(A), \kappa_2(A + uv^T))\epsilon_M)$ or $\mathcal{O}(\kappa_2(A)\epsilon_M)$.

4. SM with iterative refinement. We now examine the stability of SM-IR as in Algorithm 2. The main goal of this section is to prove Theorem 4.4, which claims that the residual in SM admits the required form for invoking Theorem 12.3 in Higham's book [14]. This can be used to establish that, under natural conditions, iterative refinement enhances backward stability of SM.

Invoking Theorem 12.3 requires a *componentwise* bound on the residual which is more intricate than the *normwise* bound we established in the previous section. Moreover, here we account for *all* rounding errors without exception, even though they do not affect the overall outcome.

The table below outlines all steps in both exact and floating-point arithmetic, along with the relevant bounds. Below and in what follows, given a vector $v \in \mathbb{R}^n$, $|v| \in \mathbb{R}^n$ takes the absolute values elementwise.

It is assumed that all steps are performed using backward-stable algorithms. Specifically, the bounds in Steps 1-2 hold for solving the systems with QR factorization computed using Householder or Givens transformations [14, pp. 361-368]. We now begin bounding the residual

$$\begin{aligned} r &= b - (A + uv^T)\hat{x} = b - A\hat{x} - (v^T\hat{x})u \\ &= b - A\hat{y} + A\hat{w} - A\delta_- - (v^T\hat{x})u \quad \text{where } A\hat{w} = A(\hat{\theta}\hat{z} + \delta_\times) = \hat{\theta}A\hat{z} + A\delta_\times \\ &= b - (b - \Delta_1\hat{y}) + A(\hat{\theta}\hat{z} + \delta_\times) - A\delta_- - (v^T\hat{x})u \\ &= \Delta_1\hat{y} - \hat{\theta}\Delta_2\hat{z} + A\delta_\times - A\delta_- + u(\hat{\theta} - v^T\hat{x}) \end{aligned} \quad (4.1)$$

The only term in the last expression that does not explicitly contain a delta-type factor of order ϵ_M is the last one. Thus, we focus on bounding $\hat{\theta} - v^T\hat{x}$, whose exact-arithmetic counterpart, $\theta - v^Tx$, is zero.

LEMMA 4.1.

$$|\hat{\theta} - v^T\hat{x}| \leq \gamma_{n+1}|v|^T|\hat{y}| + \gamma_{n+4}\frac{|\hat{\alpha}|}{|\hat{\beta}|}|v|^T|\hat{z}| + 2\epsilon_M\frac{|\hat{\alpha}|}{|\hat{\beta}|} + \mathcal{O}(\epsilon_M^2)$$

Proof. Step 7 yields

$$v^T\hat{x} = v^T\hat{y} - v^T\hat{w} + v^T\delta_- = v^T\hat{y} - v^T(\hat{\theta}\hat{z} + \delta_\times) + v^T\delta_- = v^T\hat{y} - \hat{\theta}v^T\hat{z} - v^T\delta_\times + v^T\delta_-.$$

⁴Let $\delta y := y - \hat{y}$ denote the forward error in $\hat{y}$. We have

$$A\delta y = A\hat{y} - Ay = (b - \Delta_1\hat{y}) - Ay = -\Delta_1\hat{y} \rightarrow \delta y = -A^{-1}\Delta_1\hat{y} \rightarrow \|\delta y\| \leq \|A^{-1}\|\|\Delta_1\|\|\hat{y}\|$$

and so from (3.3) we get

$$\|\delta y\| \leq c_1 \epsilon_M \kappa(A) \|\hat{y}\|. \quad \text{forward error in terms of } \|\hat{y}\|$$

Now, $\|\hat{y}\| = \|y + \delta y\| \leq \|y\| + \|\delta y\| \leq \|y\| + c_1 \epsilon_M \kappa(A) \|\hat{y}\|$ yields $\|\hat{y}\| \leq \frac{\|y\|}{1 - c_1 \epsilon_M \kappa(A)} = \frac{\|A^{-1}b\|}{1 - c_1 \epsilon_M \kappa(A)}.$

exact arithmetic op.	floating-point op.	backward error bound
1. $y = A \setminus b$	$\hat{y} = fl(y)$	$(A + \Delta_1)\hat{y} = b$, where $ \Delta_1 \leq \tilde{\gamma}_{n^2} e e^T A $
2. $z = A \setminus u$	$\hat{z} = fl(z)$	$(A + \Delta_2)\hat{z} = u$, where $ \Delta_2 \leq \tilde{\gamma}_{n^2} e e^T A $
3. $\alpha = v^T y$	$\hat{\alpha} = fl(\alpha) = v^T \otimes \hat{y}$	$\hat{\alpha} = v^T \hat{y} + \delta_{\hat{y}}$ with $ \delta_{\hat{y}} \leq \gamma_n v ^T \hat{y} $
4. $\beta = 1 + v^T z$	$\hat{\beta} = fl(\beta) = 1 \oplus (v^T \otimes \hat{z})$	$\hat{\beta} = 1 + (v^T \hat{z} + \delta_{\hat{z}}) + \delta_+$ with $ \delta_{\hat{z}} \leq \gamma_n v ^T \hat{z} $, $ \delta_+ \leq \epsilon_M (1 + v^T \otimes \hat{z})$ and $v^T \otimes \hat{z} = v^T \hat{z} + \delta_{\hat{z}}$
5. $\theta = \alpha / \beta$	$\hat{\theta} = fl(\theta) = \hat{\alpha} \ominus \hat{\beta}$	$\hat{\theta} = \frac{\hat{\alpha}}{\hat{\beta}} + \delta_{\hat{\theta}}$ with $ \delta_{\hat{\theta}} \leq \frac{\epsilon_M \hat{\alpha} }{ \hat{\beta} }$
6. $w = \theta z$	$\hat{w} = fl(w) = \hat{\theta} \otimes \hat{z}$	$\hat{w} = \hat{\theta} \hat{z} + \delta_{\times}$ with $ \delta_{\times} \leq \epsilon_M \hat{\theta} \hat{z} $
7. $x = y - w$	$\hat{x} = fl(x) = \hat{y} \ominus \hat{w}$	$\hat{x} = \hat{y} - \hat{w} + \delta_-$ with $ \delta_- \leq \epsilon_M \hat{y} - \hat{w} \leq \epsilon_M \hat{y} + \epsilon_M \hat{w} $ $\leq \epsilon_M \hat{y} + \epsilon_M \hat{\theta} \hat{z} + \epsilon_M^2 \hat{\theta} \hat{z} $

Hence

$$\hat{\theta} - v^T \hat{x} = \hat{\theta} - v^T \hat{y} + \hat{\theta} v^T \hat{z} + v^T \delta_{\times} - v^T \delta_- = (1 + v^T \hat{z}) \hat{\theta} - v^T \hat{y} + v^T (\delta_{\times} - \delta_-).$$

Substituting $\hat{\theta} = \frac{\hat{\alpha}}{\hat{\beta}} + \delta_{\hat{\theta}}$ from Step 5 gives

$$\hat{\theta} - v^T \hat{x} = (1 + v^T \hat{z}) \frac{\hat{\alpha}}{\hat{\beta}} - v^T \hat{y} + (1 + v^T \hat{z}) \delta_{\hat{\theta}} + v^T (\delta_{\times} - \delta_-). \quad (4.2)$$

On the other hand, from Step 4 we have $\hat{\beta} = 1 + v^T \hat{z} + \delta_{\hat{z}} + \delta_+$. Substituting $1 + v^T \hat{z} = \hat{\beta} - \delta_{\hat{z}} - \delta_+$ into (4.2) yields

$$\begin{aligned} \hat{\theta} - v^T \hat{x} &= (\hat{\beta} - \delta_{\hat{z}} - \delta_+) \frac{\hat{\alpha}}{\hat{\beta}} - v^T \hat{y} + (1 + v^T \hat{z}) \delta_{\hat{\theta}} + v^T (\delta_{\times} - \delta_-) \\ &= \hat{\alpha} - v^T \hat{y} - \frac{\hat{\alpha}}{\hat{\beta}} \delta_{\hat{z}} - \frac{\hat{\alpha}}{\hat{\beta}} \delta_+ + (1 + v^T \hat{z}) \delta_{\hat{\theta}} + v^T (\delta_{\times} - \delta_-) \\ &= \delta_{\hat{y}} - \frac{\hat{\alpha}}{\hat{\beta}} \delta_{\hat{z}} - \frac{\hat{\alpha}}{\hat{\beta}} \delta_+ + (1 + v^T \hat{z}) \delta_{\hat{\theta}} + v^T (\delta_{\times} - \delta_-) \end{aligned} \quad (4.3)$$

which implies

$$\begin{aligned} |\hat{\theta} - v^T \hat{x}| &\leq |\delta_{\hat{y}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| |\delta_{\hat{z}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| |\delta_+| + (1 + |v|^T |\hat{z}|) |\delta_{\hat{\theta}}| + |v|^T (|\delta_{\times}| + |\delta_-|) \\ &= |\delta_{\hat{y}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| (|\delta_{\hat{z}}| + |\delta_+|) + (1 + |v|^T |\hat{z}|) \epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} + |v|^T (|\delta_{\times}| + |\delta_-|) \\ &= |\delta_{\hat{y}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \left(|\delta_{\hat{z}}| + |\delta_+| + (1 + |v|^T |\hat{z}|) \epsilon_M \right) + |v|^T (|\delta_{\times}| + |\delta_-|) \end{aligned} \quad (4.4)$$

Looking at the last term, recall from Steps 6 and 7 that

$$|\delta_{\times}| \leq \epsilon_M |\hat{\theta}| |\hat{z}| \quad \text{and} \quad |\delta_-| \leq \epsilon_M |\hat{y}| + \epsilon_M |\hat{\theta}| |\hat{z}| + \mathcal{O}(\epsilon_M^2)$$

which give

$$|\delta_{\times}| + |\delta_{-}| \leq 2\epsilon_M |\hat{\theta}| |\hat{z}| + \epsilon_M |\hat{y}| + \mathcal{O}(\epsilon_M^2).$$

Next, from Step 5 we have

$$|\hat{\theta}| \leq \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| + |\delta_{\hat{\theta}}| \leq \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| + \epsilon_M \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \quad (4.5)$$

which yields

$$|\delta_{\times}| + |\delta_{-}| \leq 2\epsilon_M \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| |\hat{z}| + \epsilon_M |\hat{y}| + \mathcal{O}(\epsilon_M^2). \quad (4.6)$$

Substituting this into (4.4) gives

$$|\hat{\theta} - v^T \hat{x}| \leq |\delta_{\hat{\theta}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \left(|\delta_{\hat{z}}| + |\delta_{+}| + (1 + |v|^T |\hat{z}|) \epsilon_M + 2\epsilon_M |v|^T |\hat{z}| \right) + \epsilon_M |v|^T |\hat{y}| + \mathcal{O}(\epsilon_M^2)$$

Also we know from Step 4 that

$$|\delta_{+}| \leq \epsilon_M (1 + |v|^T |\hat{z}| + |\delta_{\hat{z}}|) \leq \epsilon_M (1 + |v|^T |\hat{z}| + \gamma_n |v|^T |\hat{z}|) = \epsilon_M (1 + |v|^T |\hat{z}|) + \mathcal{O}(\epsilon_M^2) \blacksquare$$

which gives

$$\begin{aligned} |\hat{\theta} - v^T \hat{x}| &\leq |\delta_{\hat{\theta}}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \left(|\delta_{\hat{z}}| + 2\epsilon_M (1 + |v|^T |\hat{z}|) + 2\epsilon_M |v|^T |\hat{z}| \right) + \epsilon_M |v|^T |\hat{y}| + \mathcal{O}(\epsilon_M^2) \\ &\leq \gamma_n |v|^T |\hat{y}| + \epsilon_M |v|^T |\hat{y}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \left(\gamma_n |v|^T |\hat{z}| + 2\epsilon_M (1 + |v|^T |\hat{z}|) + 2\epsilon_M |v|^T |\hat{z}| \right) + \mathcal{O}(\epsilon_M^2) \\ &= (\gamma_n + \epsilon_M) |v|^T |\hat{y}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \left((\gamma_n + 4\epsilon_M) |v|^T |\hat{z}| + 2\epsilon_M \right) + \mathcal{O}(\epsilon_M^2) \blacksquare \end{aligned}$$

Since $(\gamma_n + \epsilon_M) \leq \gamma_{n+1}$ and $(\gamma_n + 4\epsilon_M) \leq \gamma_{n+4}$, the proof is complete. $\square$

To obtain a bound on the residual, we require the following two auxiliary results.

LEMMA 4.2.

$$|\hat{y}| \leq (1 + \epsilon_M) |\hat{x}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| (1 + 3\epsilon_M) |\hat{z}| + \mathcal{O}(\epsilon_M^2). \quad (4.7)$$

Proof. Applying the reverse triangle inequality to $\hat{x} = \hat{y} - \hat{w} + \delta_{-}$ gives $|\hat{y}| \leq |\hat{x}| + |\hat{w}| + |\delta_{-}|$. From $\hat{w} = \hat{\theta} \hat{z} + \delta_{\times}$ with $|\delta_{\times}| \leq \epsilon_M |\hat{\theta}| |\hat{z}|$ and (4.5) we obtain

$$|\hat{w}| \leq |\hat{\theta}| |\hat{z}| (1 + \epsilon_M) \leq \left(\left| \frac{\hat{\alpha}}{\hat{\beta}} \right| (1 + \epsilon_M) + \epsilon_M \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| \right) |\hat{z}| + \mathcal{O}(\epsilon_M^2).$$

Also, $|\delta_{-}| \leq \epsilon_M |\hat{y}| + \epsilon_M \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| |\hat{z}| + \mathcal{O}(\epsilon_M^2)$ leading to

$$|\hat{y}| \leq |\hat{x}| + \left| \frac{\hat{\alpha}}{\hat{\beta}} \right| (1 + 3\epsilon_M) |\hat{z}| + \epsilon_M |\hat{y}| + \mathcal{O}(\epsilon_M^2).$$

Applying $(1 - \epsilon_M)^{-1} = 1 + \epsilon_M + \mathcal{O}(\epsilon_M^2)$ yields the result. $\square$

LEMMA 4.3.

$$|\hat{z}| \leq |A^{-1} u| + \tilde{\gamma}_{n^2} |A^{-1} e e^T A| |A^{-1} u| + \mathcal{O}(\epsilon_M^2). \quad (4.8)$$

Proof. From Step 2 we have

$$\hat{z} = (A + \Delta_2)^{-1}u = (A(I + A^{-1}\Delta_2))^{-1}u =: (I + F)A^{-1}u$$

where $|F| \leq \tilde{\gamma}_{n^2}|A^{-1}|ee^T|A| + \mathcal{O}(\epsilon_M^2)$. $\square$

Our main result is the following.

THEOREM 4.4. *The SM residual satisfies*

$$|b - A\hat{x} - (v^T\hat{x})u| \leq \epsilon_M (g(A, u, v)|\hat{x}| + h(A, u, v, b)) \quad (4.9)$$

where

$$g(A, u, v) = (cn^2ee^T + I)|A| + (n+1)|u||v|^T$$

and

$$h(A, u, v, b) = \frac{|\hat{\alpha}|}{|\hat{\beta}|} \left((2dn^2ee^T + 3I)|A| + 2(n+4)|u||v|^T \right) |A^{-1}u| + 2 \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u|. \quad (4.10)$$

in which c and d are small integer constants⁵.

Proof. Beginning with (4.1),

$$|r| \leq |\Delta_1||\hat{y}| + |\hat{\theta}||\Delta_2||\hat{z}| + |A|(|\delta_{\times}| + |\delta_{-}|) + |u||\hat{\theta} - v^T\hat{x}|$$

we apply Lemma 4.1 to obtain

$$\begin{aligned} |r| &\leq |\Delta_1||\hat{y}| + |\hat{\theta}||\Delta_2||\hat{z}| + |A|(|\delta_{\times}| + |\delta_{-}|) \\ &\quad + \gamma_{n+1}|u||v|^T|\hat{y}| + \gamma_{n+4} \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u||v|^T|\hat{z}| + 2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u| + \mathcal{O}(\epsilon_M^2). \end{aligned}$$

Applying (4.5) and (4.6) yield

$$\begin{aligned} |r| &\leq |\Delta_1||\hat{y}| + \left(\frac{|\hat{\alpha}|}{|\hat{\beta}|} + \epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} \right) |\Delta_2||\hat{z}| \\ &\quad + |A| \left(2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |\hat{z}| + \epsilon_M |\hat{y}| + \mathcal{O}(\epsilon_M^2) \right) \\ &\quad + \gamma_{n+1}|u||v|^T|\hat{y}| + \gamma_{n+4} \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u||v|^T|\hat{z}| + 2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u| + \mathcal{O}(\epsilon_M^2) \\ &= \left(|\Delta_1| + \epsilon_M |A| + \gamma_{n+1}|u||v|^T \right) |\hat{y}| \\ &\quad + \frac{|\hat{\alpha}|}{|\hat{\beta}|} \left(|\Delta_2| + 2\epsilon_M |A| + \gamma_{n+4}|u||v|^T \right) |\hat{z}| + 2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u| + \mathcal{O}(\epsilon_M^2) \\ &\leq \left((\tilde{\gamma}_{n^2}ee^T + \epsilon_M I)|A| + \gamma_{n+1}|u||v|^T \right) |\hat{y}| \\ &\quad + \frac{|\hat{\alpha}|}{|\hat{\beta}|} \left((\tilde{\gamma}_{n^2}ee^T + 2\epsilon_M I)|A| + \gamma_{n+4}|u||v|^T \right) |\hat{z}| + 2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u| + \mathcal{O}(\epsilon_M^2). \end{aligned}$$

⁵appearing in the notation $\tilde{\gamma}_k = \gamma_{ck} = \frac{ck\epsilon_M}{1-ck\epsilon_M}$.

Next, we apply (4.7) and (4.8), pushing higher-order terms to $\mathcal{O}(\epsilon_M^2)$ to obtain

$$|r| \leq \left((\tilde{\gamma}_{n^2} e e^T + \epsilon_M I) |A| + \gamma_{n+1} |u| |v|^T \right) |\hat{x}| \\ + \frac{|\hat{\alpha}|}{|\hat{\beta}|} \left((2\tilde{\gamma}_{n^2} e e^T + 3\epsilon_M I) |A| + (\gamma_{n+1} + \gamma_{n+4}) |u| |v|^T \right) |A^{-1} u| + 2\epsilon_M \frac{|\hat{\alpha}|}{|\hat{\beta}|} |u| + \mathcal{O}(\epsilon_M^2). \blacksquare$$

Note that $\gamma_{n+1} + \gamma_{n+4} \leq 2\gamma_{n+4}$. $\square$

In addition to Theorem 4.4, we require a bound on the difference between the residual r and the computed residual $\hat{r}$ in SM to be able to apply [14, Thm. 12.3].

LEMMA 4.5. *Let $r := b - A\hat{x} - (v^T \hat{x})u$ and $\hat{r} = fl(r)$. Then, we have*

$$|r - \hat{r}| \leq \epsilon_M t(A, u, v, \hat{x}, b) \quad (4.11)$$

where

$$t(A, u, v, \hat{x}, b) = \frac{\gamma_{n+2}}{\epsilon_M} \left(|b| + (|A| + |u| |v|^T) |\hat{x}| \right). \quad (4.12)$$

Proof. First, consider $p = b - A\hat{x}$ and $\hat{p} = fl(p)$. It is easy to see that $\hat{p} = p + \delta p$ where $|\delta p| \leq \gamma_{n+1}(|b| + |A||\hat{x}|)$. Next, let $q = (v^T \hat{x})u$ and $\hat{q} = fl(q) = (v^T \otimes \hat{x}) \otimes u$. We know that $\hat{q} = q + \delta q$ with (see [14, p. 74] for instance)

$$|\delta q| \leq (\gamma_n + \epsilon_M(1 + \gamma_n)) |u| |v|^T |\hat{x}| \leq \gamma_{n+1} |u| |v|^T |\hat{x}| + \mathcal{O}(\epsilon_M^2).$$

Finally, we have $\hat{r} = \hat{p} \ominus \hat{q}$ which satisfies $\hat{r} = \hat{p} - \hat{q} + \check{\delta}_-$ with

$$|\check{\delta}_-| \leq \epsilon_M(|\hat{p}| + |\hat{q}|) \leq \epsilon_M(|p| + |\delta p|) + \epsilon_M(|q| + |\delta q|) \leq \epsilon_M(|b| + |A||\hat{x}|) + \epsilon_M |u| |v|^T |\hat{x}| + \mathcal{O}(\epsilon_M^2). \blacksquare$$

We have

$$\hat{r} = (p + \delta p) - (q + \delta q) + \check{\delta}_- = (p - q) + \delta p - \delta q + \check{\delta}_- = r + \delta r$$

with $\delta r := \delta p - \delta q + \check{\delta}_-$ which yields

$$|r - \hat{r}| = |\delta r| \leq |\delta p| + |\delta q| + |\check{\delta}_-| \\ \leq (\gamma_{n+1} + \epsilon_M)(|b| + |A||\hat{x}|) + (\gamma_{n+1} + \epsilon_M) |u| |v|^T |\hat{x}| + \mathcal{O}(\epsilon_M^2) \\ \leq \gamma_{n+2}(|b| + |A||\hat{x}|) + \gamma_{n+2} |u| |v|^T |\hat{x}| + \mathcal{O}(\epsilon_M^2).$$

$\square$

We are now ready to invoke [14, Thm. 12.3] to establish our main result demonstrating that each SM iteration improves the stability of SM. See Alg.2.

THEOREM 4.6. *Let A and $B := A + uv^T$ be nonsingular $n \times n$ matrices. Suppose the linear system $Bx = b$ is solved using SM in floating-point arithmetic following one step of iterative refinement as in Algorithm 2, where y , z and y_r are computed with a backward stable algorithm. Assume that the computed SM solution $\hat{x}$ satisfies (4.9), and the computed residual satisfies (4.11). Then, the corrected solution $\hat{w}$ satisfies*

$$|b - A\hat{w} - (v^T \hat{w})u| \leq \epsilon_M \left(h(A, u, v, \hat{r}) + t(A, u, v, \hat{w}, b) + (|A| + |u| |v|^T) |\hat{w}| \right) + \epsilon_M q, \quad (4.13)$$

where $q = \mathcal{O}(\epsilon_M)$ if $t(A, u, v, \hat{x}, b) - t(A, u, v, \hat{w}, b) = \mathcal{O}(\|\hat{x} - \hat{w}\|_\infty)$.

Next we adapt Higham's explanation of [14, Thm. 12.3] to our linear system $Bx = b$. Theorem 4.13 indicates that, to first order, the componentwise relative backward error

$$\text{Berr}_{|B|,|b|}(\hat{w}) = \max_i \frac{|r|_i}{(|B||\hat{w}| + |b|)_i}$$

(see [17] and [14, Thm. 7.3]) will be small after one step of iterative refinement as long as $t(A, u, v, \hat{w}, b)$ and $h(A, u, v, \hat{r})$ are bounded by a modest scalar multiple of $|B||\hat{w}| + |b|$, the denominator of $\text{Berr}_{|B|,|b|}(\hat{w})$. Following (4.12) we have

$$t(A, u, v, \hat{w}, b) = \frac{\gamma_{n+2}}{\epsilon_M} \left(|b| + (|A| + |u||v|^T)|\hat{w}| \right)$$

which is exactly what is required. In the case of h , note first that $h(A, u, v, \hat{r})$ is obtained by replacing b in (4.10) with $\hat{r}$. This means that $\hat{\alpha}$ in (4.10) is replaced with $\hat{\alpha}_r$, the quantity computed in Step 4 of Alg. 2. We have therefore shown that

THEOREM 4.7. *If*

$$h(A, u, v, \hat{r}) = \left((2dn^2ee^T + 3I)|A| + 2(n+4)|u||v|^T \right) \frac{|\hat{\alpha}_r|}{|\hat{\beta}|} |A^{-1}u| + 2 \frac{|\hat{\alpha}_r|}{|\hat{\beta}|} |u| \quad (4.14)$$

is bounded by a modest scalar multiple of $|A + uv^T||\hat{w}| + |b|$, then one step of SM-IR guarantees (componentwise relative) backward stability.

This is a general theorem; to be able to say anything more specific we would require knowledge of the size of

$$\frac{|\hat{\alpha}_r|}{|\hat{\beta}|} |A^{-1}u|$$

against $\hat{w}$. In trying gain some insight, let us consider the two cases of large- and small-norm solutions.

- If the solution has a large norm $\mathcal{O}(\sigma_{\min}(A + uv^T)\|b\|)$, it is likely that SM is already backward stable in which case IR is not required but even if it is applied, assuming the SM-IR solution $\hat{w}$ has a similar norm to the SM solution $\hat{x}$ as well as the exact solution x , we can expect $\frac{|\hat{\alpha}_r|}{|\hat{\beta}|} \|A^{-1}u\|$ to be bounded by $\|\hat{w}\|$ hence backward stability is maintained.
- On the other hand, consider the small-norm scenario in which A and $A + uv^T$ are ill-conditioned, $\|x\| = \mathcal{O}(1)$ (e.g., $\|A\| = \|b\| = \mathcal{O}(1)$), but $\|y\|$ and $\|z\|$ are $\mathcal{O}(\kappa(A)) \gg 1$. In a nutshell, Lemma 3.1 says $\frac{\hat{\alpha}}{\hat{\beta}} = \mathcal{O}(1)$, as follows from (3.11) and assumptions such as $\frac{\cos(\theta_{v,\hat{y}})}{\cos(\theta_{v,\hat{z}})} = \mathcal{O}(1)$, and that the SM residual $\hat{r}$ before any IR steps is proportional to $\epsilon_M \kappa(A)$; similarly, assuming Lemma 3.1 applies when $\hat{y}_r$ is used instead of $\hat{y}$, we can argue that

$$\frac{|\hat{\alpha}_r|}{|\hat{\beta}|} \|\hat{z}\| \leq \check{c} \|\hat{y}_{\hat{r}}\| + \mathcal{O}(\epsilon_M)$$

where $\hat{y}_{\hat{r}} = fl(y_{\hat{r}}) = fl(A^{-1}\hat{r})$ whose norm could be bounded analogous with (3.13) where b is replaced with $\hat{r}$. Assuming $\hat{w} = \mathcal{O}(1)$, we have

$$\begin{aligned} \frac{|\hat{\alpha}_r|}{|\hat{\beta}|} \max_i \frac{|A^{-1}u|_i}{|\hat{w}|_i} &\approx \frac{\|A^{-1}\hat{r}\|}{\|A^{-1}u\|} \|A^{-1}u\| = \|A^{-1}\hat{r}\| \leq \|A^{-1}\| \|\hat{r}\| \\ &= \mathcal{O}(\kappa(A)) \mathcal{O}(\epsilon_M \kappa(A)) = \mathcal{O}(\epsilon_M \kappa^2(A)). \end{aligned}$$

Roughly speaking, this implies that in the small-norm solution scenario, for SM to achieve backward stability with one step of fixed-precision IR, we would require $\kappa(A) < \epsilon_M^{-1/2}$, which corresponds to 6.7×10^7 in double precision. In such cases the original SM residual r for the SM solution $\hat{x}$ to $Bx = b$ (i.e., before any IR steps) has a norm bounded by $\epsilon_M \kappa(A) \|b\| \approx 10^{-8} \|b\|$. With the first IR step, as discussed above, the deciding quantity is essentially $\|A^{-1} \hat{r}\|$ which corresponds to the solution norm of a linear system $Ax = \hat{r}$. This can be bounded by $\epsilon_M \kappa(A) \|\hat{r}\| \lesssim \epsilon_M^2 \kappa^2(A) \|b\| \lesssim \epsilon_M \|b\|$ thereby achieving backward stability.

5. Numerical experiments. In our experiments we compare the following algorithms for solving (1.1):

1. MATLAB backslash applied directly to the sum $B := A + uv^T$: This is GEPP (Gaussian elimination with partial pivoting) based on computing the LU decomposition with partial pivoting.
2. SM-LU: Algorithm 1 where the A -solves ($Ay = b$ and $Az = u$) are done using GEPP.
3. SM-QR: Algorithm 1 using the QR factorization of A for A -solves.
4. SM-LU-IR: This is SM-IR (Algorithm 2) using GEPP for A -solves.

The complexity of the first algorithm is cubic $\mathcal{O}(n^3)$ in the matrix dimension, $A \in \mathbb{R}^{n \times n}$. Assuming the factors of an LU or QR decomposition of A are already available and the number of IR steps is $\mathcal{O}(1)$, the complexity of the rest of the methods is quadratic $\mathcal{O}(n^2)$.

While GEPP/LU is known to have adversarial examples where $Ax = b$ would not be solved in a stable fashion, such matrices are known to be extremely rare [23, Ch. 22], and empirically LU performs exceptionally well (QR, by contrast, is always backward stable). For this reason we do not present SM-QR-IR, as (expectedly) its empirical performance is almost identical to SM-LU-IR.

In our experiments, the normwise relative backward error of an approximate solution $\tilde{x}$ is computed using the bound in (3.1). We take $\max_{i=1}^n \frac{|r_i|}{(|B||\tilde{x}| + |b|)_i}$ as the componentwise relative backward error. In SM-LU-IR we repeated IR until the relative backward error fell below $5\epsilon_M \approx 5 \times 10^{-16}$.

Based on the insight obtained in Section 2, and because the empirical performance differ based on the parameters, we split the experiments in terms of the magnitude of $\kappa(A)$, $\kappa(A + uv^T)$, and whether or not the norm of the solution x is small $\mathcal{O}(1)$ or large $\mathcal{O}(1/\sigma_{\min}(A + uv^T)) = \mathcal{O}(\kappa(A + uv^T))$. Throughout this section, κ denotes the 2-norm condition number $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$. All reported backward and forward errors are relative. We consider four cases based on whether or not $\kappa(A)$, $\kappa(A + uv^T) \gg 1$, and when $\kappa(A + uv^T) \gg 1$, we further split into two cases: (i) the solution x has small norm $\|x\| = \mathcal{O}(1)$, and (ii) large norm $\|x\| = \mathcal{O}(\kappa(A + uv^T))$.

5.1. Case 1(i): both $\kappa(A), \kappa(A + uv^T) \gg 1$, small-norm solution. **EXAMPLE 5.1.** We generate sparse random matrices A of size $n = 8000$ using the `sprandn` routine in MATLAB with condition numbers of $10^6, 10^8, 10^{10}$ and 10^{12} . The density of A is set to 0.0001 resulting in a matrix with about 8000 nonzero entries. In this example we had $\kappa(A) = 10^6$. The vectors u and v are generated with `randn`. In all four matrices, the condition number of $B = A + uv^T$ was about 10^{10} ; all of these are reported on the horizontal axes in Figure 5.1. We take the exact solution to also be a `randn` vector for which $\|x\| \approx 90$ and then form $b := Bx$.

As the condition number of A grows, the number of IR steps increases from

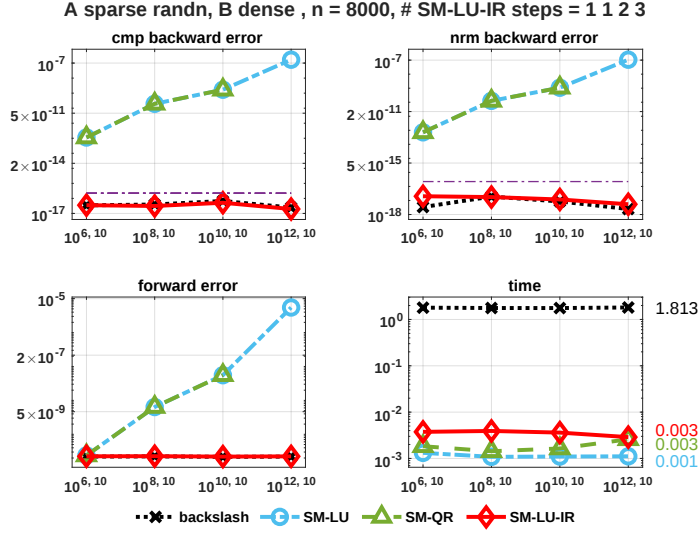


FIG. 5.1. Results for Example 5.1 with sparse A . Case 1(i): $\kappa(A), \kappa(A + uv^T) \gg 1$, small-norm solution.

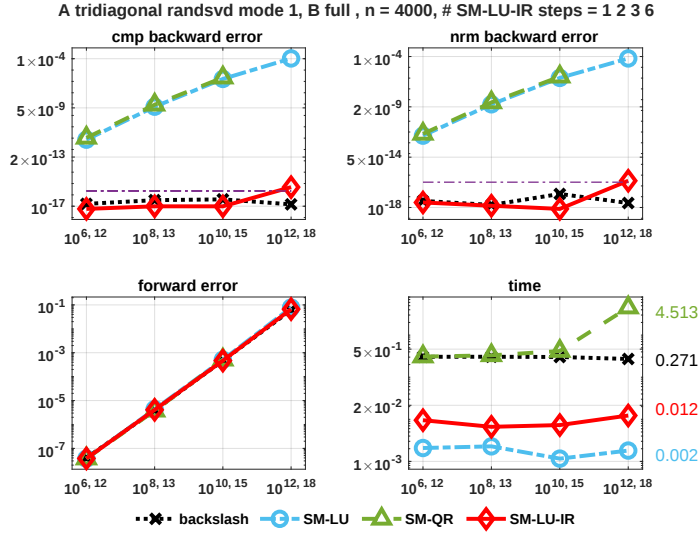


FIG. 5.2. Results for Example 5.2. Case 1(i): $\kappa(A), \kappa(A + uv^T) \gg 1$, small-norm solution.

1 to 3 for the most ill-conditioned matrix. The number of steps are also reported in the title of Figure 5.1. To be clear, for the first test where $\kappa(A) = 10^6$ and $\kappa(B) = 5.1 \times 10^9 \approx 10^{10}$, SM-LU-IR needs one IR step, and in the last test which is the most ill-conditioned example, it requires three IR steps. Note that SM-LU-IR exhibits backward stability while the additional IR iterations required for the more ill-conditioned tests incur no significant time penalty. Also, as A is sparse, all variants of SM can take advantage of structure to speed up their computation. B is not sparse, which is why backslash applied to B is the slowest method in all four tests.

We also note that SM-QR does not perform well for the most ill-conditioned ex-

ample for which the backward and forward errors are NaN. This is due to infinity entries in y and z computed with the QR factors when applied to the sparse matrix A . The loss of accuracy could be mitigated at the cost of slowing down **SM-QR** if A is first converted to a full matrix, thereby neglecting its sparsity. Even then, as with the previous three tests, **SM-QR** without IR would still fail to achieve backward stability.

EXAMPLE 5.2. We generate A of size $n = 4000$ using **randsvd** matrices of mode 1 in the MATLAB gallery with condition numbers of $10^6, 10^8, 10^{10}$ and 10^{11} . We also generate u and v as **randn** vectors. The condition number of matrices $B = A + uv^T$ are $10^{11}, 10^{13}, 10^{15}$ and 10^{18} , respectively; all of these are reported on the horizontal axes in Figure 5.2. We take the exact solution to also be a **randn** vector for which $\|x\| \approx 30$ and then form $b := Bx$.

The backward error of **SM-LU-IR** is comparable with backslash applied to B , but it is faster even in the last case where 6 IR steps are required.

5.2. Case 1(ii): both $\kappa(A), \kappa(A + uv^T) \gg 1$, large-norm solution. **EXAMPLE 5.3.** **SM** performs stably here. See Figure 5.3. Here we set A to be tridiagonal as it is easier to control $\kappa(B)$. In addition, b is set to be a **randn** vector and the exact solution is considered the one obtained with backslash.

5.3. Case 2(i): $\kappa(A) = \mathcal{O}(1), \kappa(A + uv^T) \gg 1$, small-norm solution. **EXAMPLE 5.4.** We generate A , tridiagonal of size $n = 1000$ using **randsvd** matrices of mode 5 in the MATLAB gallery with condition numbers of $10^1, 10^2, 10^3$ and 10^4 . We also generate u and v as **randn** vectors. The condition number of matrices $B = A + uv^T$ are $10^5, 10^6, 10^7$ and 10^8 , respectively. We take the exact solution to also be a **randn** vector for which $\|x\| \approx 30$ and then form $b := Bx$. See Figure 5.4 suggesting instability of **SM** at least to some extent.

5.4. Case 2(ii): $\kappa(A) = \mathcal{O}(1), \kappa(A + uv^T) \gg 1$, large-norm solution. **EXAMPLE 5.5.** In this problem where solution has a large-norm, again **SM** performs stably. See Figure 5.5.

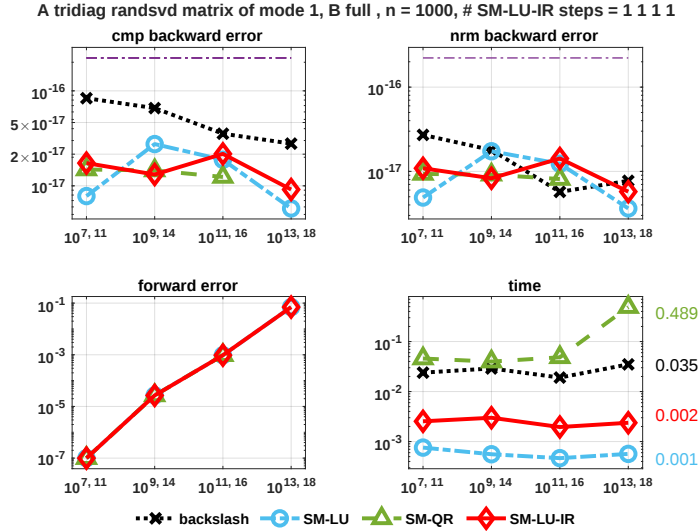


FIG. 5.3. Results for Example 5.3. Case 1(ii): $\kappa(A), \kappa(A + uv^T) \gg 1$, large-norm solution.

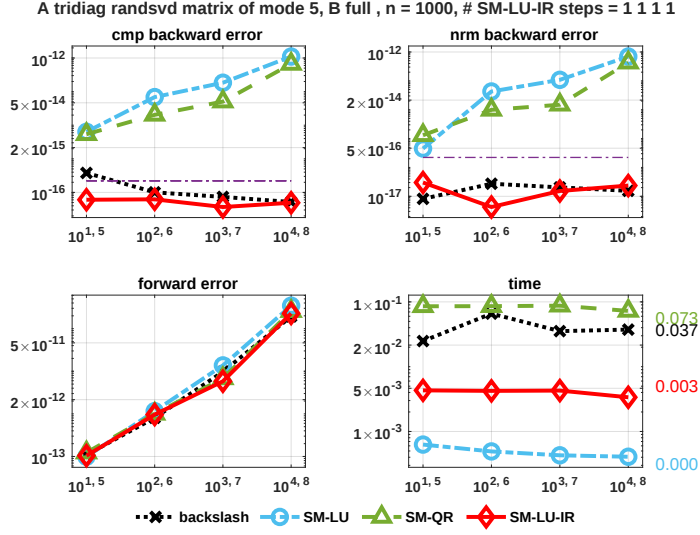


FIG. 5.4. Results for Example 5.4. Case 2(i): $\kappa(A) = \mathcal{O}(1)$, $\kappa(A + uv^T) \gg 1$, small-norm solution

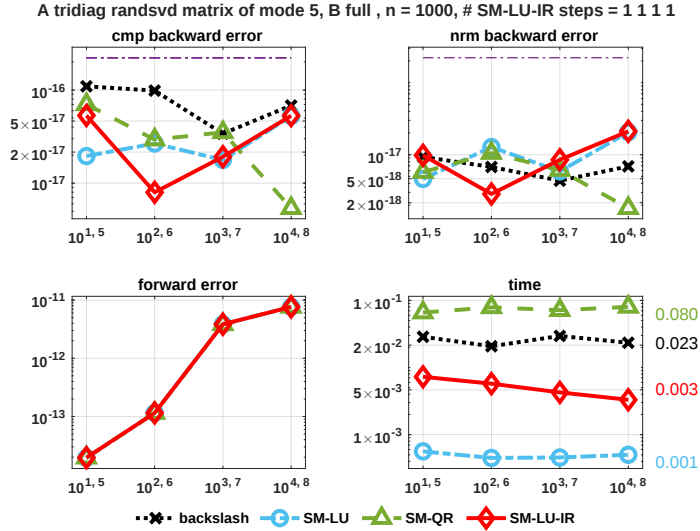


FIG. 5.5. Results for Example 5.5. Case 2(ii): $\kappa(A) = \mathcal{O}(1)$, $\kappa(A + uv^T) \gg 1$, large-norm solution.

5.5. Case 3: $\kappa(A) \gg 1$, $\kappa(A + uv^T) = \mathcal{O}(1)$, small-norm solution. EXAM-
 PLE 5.6. We generate A , pentadiagonal of size $n = 1000$ using `randsvd` matrices of
 mode 2 in the MATLAB gallery with condition numbers of $\kappa = 10^7, 10^9, 10^{11}$ and 10^{13} .
 Matrices A of this type have one small singular value that is equal to κ^{-1} and all the
 remaining $n - 1$ singular values are equal to 1. In order to make B well-conditioned,
 we take u to be a random multiple of the right singular vector of A corresponding
 to its smallest singular value and analogously for v . Therefore the smallest singular
 value of A is replaced with one of the same order as the rest of its singular values
 drastically improving the condition number of $B = A + uv^T$ to $\mathcal{O}(1)$.

We take x a random vector and set $b = Bx$. Therefore, both x and b have

a small norm. See Figure 5.6. Interestingly, SM is unstable once again; reflecting the discussion in Section 2. Perhaps even more surprising is that not only SM-IR behaves in a backward stable manner, even the forward error is $\mathcal{O}(\epsilon_M)$ although this is a fixed-precision IR! Given that the plain SM is not backward stable, it has no chance of getting close to SM-IR in terms of forward stability, even though B is well-conditioned. In such cases, if SM is required (for speed), it is strongly advisable to use IR.

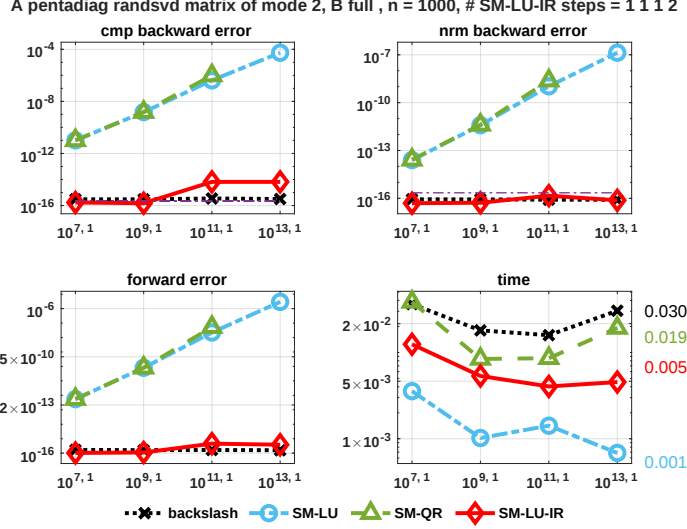


FIG. 5.6. Results for Example 5.6. Case 3(i): $\kappa(\mathbf{A}) \gg 1, \kappa(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \mathcal{O}(1)$, small-norm solution

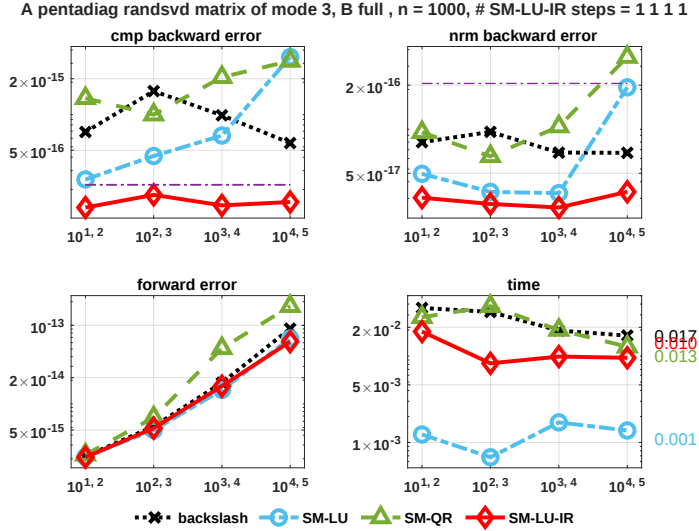


FIG. 5.7. Results for Example 5.7. Case 4: both $\kappa(\mathbf{A}), \kappa(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \mathcal{O}(1)$, small-norm solution

5.6. Case 4: both $\kappa(\mathbf{A}), \kappa(\mathbf{A} + \mathbf{u}\mathbf{v}^T) = \mathcal{O}(1)$, small-norm solution. EXAMPLE 5.7. We generate pentadiagonal A of size $n = 1000$ using randsvd matrices of

default mode 3 in the MATLAB gallery with condition numbers of $10^1, 10^2, 10^3$ and 10^4 .

We take u and v as random vectors, but also normalized to ensure that $\kappa(A + uv^T)$ does not deviate significantly from the scenario under consideration. The condition number of matrices A and $B = A + uv^T$ can be seen on the horizontal axes of Figure 5.7. We take the exact solution to also be a `randn` vector for which $\|x\| \approx 30$ and then form $b := Bx$ with $\|b\| \approx 10$. This is a situation where SM should be stable as discussed in the introduction; and the backward errors are all small. It is still worth noting that the backward error is observed to be improved by IR to $\mathcal{O}(\epsilon_M)$ from $\mathcal{O}(\kappa(A)\epsilon_M)$.

As in this case both A and $A + uv^T$ are well-conditioned, with a random b the solution cannot have a large norm. So, we skip the other sub-case again.

5.7. Effect of IR. Above we have shown the results of SM-IR after the final IR step. It is interesting to see how convergence is taking place as the IR steps proceed. We present these in the figures below, for Case 1(i) where we required the most number of IR iterations. We observe (as did in all cases with varying speed of improvement by IR) that both the residual and backward error converge steadily, until the backward error reaches $\mathcal{O}(\epsilon_M)$.

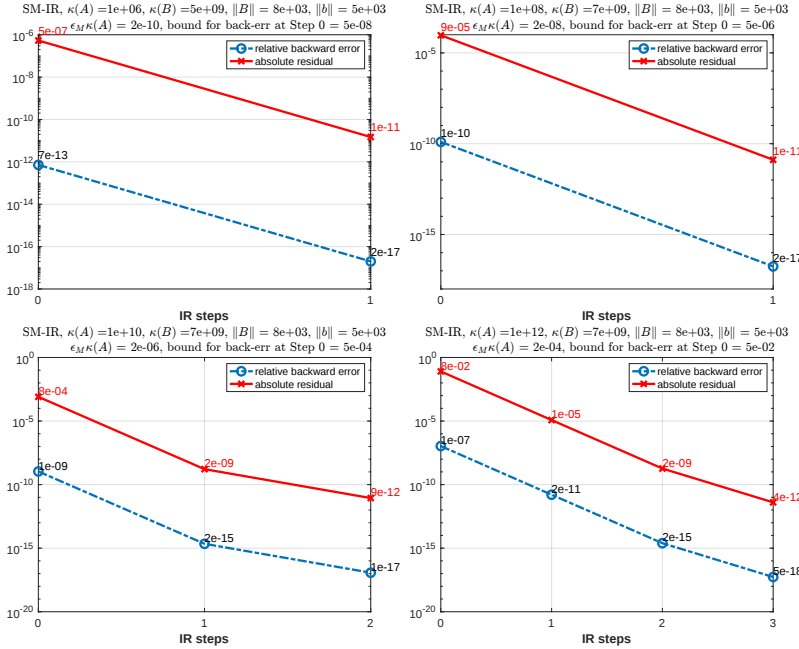


FIG. 5.8. Results for Example 5.1. A sparse, $n = 8000$. Case 1(i): $\kappa(A), \kappa(A + uv^T) \gg 1$, small-norm solution

Summary of experiments. In all our experiments, SM(-LU)-IR was able to compute a backward stable solution after $\mathcal{O}(1)$ IR steps, without sacrificing speed. Based on these promising empirical results, we conjecture that SM-IR is backward stable when applied to numerically nonsingular linear systems. As noted in the introduction, our theory requires the assumptions to hold, so is not sufficient to prove this in full. A full proof is left as an open problem.

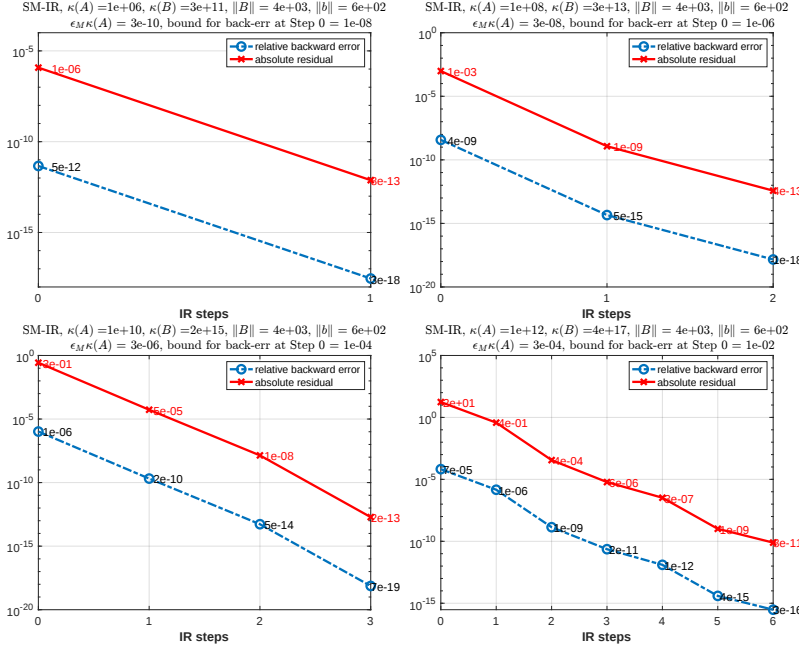


FIG. 5.9. Results for Example 5.2, A tridiagonal, $n = 4000$. Case 1(i): $\kappa(A), \kappa(A + uv^T) \gg 1$, small-norm solution.

Acknowledgments. We are grateful to David Bindel for pointing out the connection to bordered linear systems and for bringing [8] to our attention. We also thank Alex Townsend for highlighting the relevance of structured-plus-low-rank linear systems in ultraspherical spectral methods.

REFERENCES

- [1] B. ADCOCK AND D. HUYBRECHS, *Frames and numerical approximation*, SIAM Rev., 61 (2019), pp. 443–473.
- [2] M. S. BARTLETT, *An inverse matrix adjustment arising in discriminant analysis*, Ann. Math. Stat., 22 (1951), pp. 107–111.
- [3] E. CARSON AND N. J. HIGHAM, *Accelerating the solution of linear systems by iterative refinement in three precisions*, SIAM J. Sci. Comput., 40 (2018), pp. A817–A847.
- [4] W. J. DUNCAN, *Some devices for the solution of large sets of simultaneous linear equations: with an appendix on the reciprocation of partitioned matrices*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 35 (1944), pp. 660–670.
- [5] E. N. EPPERLY, M. MEIER, AND Y. NAKATSUKASA, *Fast randomized least-squares solvers can be just as accurate and stable as classical direct solvers*, arXiv preprint arXiv:2406.03468, to appear in Comm. Pure Appl. Math., (2024).
- [6] D. FORTUNATO, N. HALE, AND A. TOWNSEND, *The ultraspherical spectral element method*, Journal of Computational Physics, 436 (2021), p. 110087.
- [7] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Numerical Linear Algebra and Optimization*, SIAM, 2021.
- [8] W. GOVAERTS, *Stable solvers and block elimination for bordered systems*, SIAM Journal on Matrix Analysis and Applications, 12 (1991), pp. 469–483.
- [9] S. GÜTTEL, Y. NAKATSUKASA, M. WEBB, AND A. B. RILEY, *A Sherman–Morrison–Woodbury approach to solving least squares problems with low-rank updates*, arXiv preprint arXiv:2406.15120, (2024).
- [10] W. W. HAGER, *Updating the inverse of a matrix*, SIAM Review, 31 (1989), pp. 221–239.
- [11] P. C. HANSEN, *Discrete Inverse Problems: Insight and Algorithms*, SIAM, 2010.

- [12] Y. HAO AND V. SIMONCINI, *The Sherman–Morrison–Woodbury formula for generalized linear matrix equations and applications*, Numer. Lin. Alg. Appl., 28 (2021), p. e2384.
- [13] T. HASTIE, A. MONTANARI, S. ROSSET, AND R. J. TIBSHIRANI, *Surprises in high-dimensional ridgeless least squares interpolation*, Annals of Statistics, 50 (2022), p. 949.
- [14] N. J. HIGHAM, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2002.
- [15] M. JANKOWSKI AND H. WOŹNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT Numerical Mathematics, 17 (1977), pp. 303–311.
- [16] L. MA, C. BOUTSIKAS, M. GHADIRI, AND P. DRINEAS, *A note on the stability of the Sherman–Morrison–Woodbury formula*, arXiv preprint 2504.04554v1, (2025).
- [17] W. OETTLI AND W. PRAGER, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numerische Mathematik, 6 (1964), pp. 405–409.
- [18] S. OLVER AND A. TOWNSEND, *A fast and well-conditioned spectral method*, SIAM Review, 55 (2013), pp. 462–489.
- [19] J.-L. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, Journal of the ACM, 14 (1967), pp. 543–548.
- [20] J. SHERMAN AND W. J. MORRISON, *Adjustment of an inverse matrix corresponding in one element of a given matrix*, Ann. Math. Stat., 21 (1950), pp. 124–127.
- [21] R. D. SKEEL, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.
- [22] G. W. STEWART, *Afternotes goes to graduate school: lectures on advanced numerical analysis*, SIAM, 1998.
- [23] L. N. TREFETHEN AND D. BAU, *Numerical Linear Algebra, Twenty-fifth Anniversary Edition*, SIAM, Philadelphia, PA, 2022.
- [24] D. S. WATKINS, *Fundamentals of Matrix Computations*, John Wiley & Sons, 2004.
- [25] M. A. WOODBURY, *Inverting modified matrices*, Department of Statistics, Princeton University, 1950.
- [26] E. L. YIP, *A note on the stability of solving a rank- p modification of a linear system by the Sherman–Morrison–Woodbury formula*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 507–513.

Appendix A. Govaerts' work. The starting point is to introduce the scalar variable $\zeta = v^T x$, which serves as a convenient intermediate step in the computation. We can now regard (1.1) as equivalent to the bordered linear system

$$\begin{bmatrix} A & u \\ v^T & -1 \end{bmatrix} \begin{bmatrix} x \\ \zeta \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \quad (\text{A.1})$$

Govaerts [8] proposed algorithms for solving general $(n+1) \times (n+1)$ bordered linear systems, where the $(1,1)$ block is a matrix A for which a black-box solver is assumed to be available. He discussed variants of block Gaussian elimination, both without and with IR. For instance, the BEC (block elimination Crout) algorithm of [8], when applied to (A.1), employs the following LU decomposition

$$\begin{bmatrix} A & u \\ v^T & -1 \end{bmatrix} = \begin{bmatrix} A & 0 \\ v^T & -1 - v^T A^{-1}u \end{bmatrix} \begin{bmatrix} I_n & A^{-1}u \\ 0 & 1 \end{bmatrix}.$$

Backward and forward substitutions then give the SM formula. The following mixed forward-backward error bound is how Govaerts's analysis of BEC [8, Prop. 3.3] adapts to (A.1).

PROPOSITION A.1. *Let S be a normwise backward stable solver for A with stability constant C_S ⁶. Then, $\hat{x}, \hat{\zeta}$ satisfy*

$$\begin{bmatrix} A + \Delta A & u + \delta u \\ (v + \delta v)^T & -1 + \delta \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{\zeta} \end{bmatrix} = \begin{bmatrix} b + \delta b \\ 0 \end{bmatrix} + \begin{bmatrix} T \\ q^T \end{bmatrix} \hat{y} \quad (\text{A.2})$$

where

$$\begin{aligned} \|\Delta A\| &\leq (2 + C_S)\epsilon_M \|A\| + \mathcal{O}(\epsilon_M^2), \\ \|\delta u\| &\leq C_S \epsilon_M \|u\|, \\ \|\delta v\| &\leq (5 + C_{IP})\epsilon_M \|v\| + \mathcal{O}(\epsilon_M^2), \\ |\delta| &\leq 3\epsilon_M \exp(3\epsilon_M), \\ \|\delta b\| &\leq C_S \epsilon_M \|b\|, \\ \|T\| &\leq (2C_S + 1)\epsilon_M \|A\| + \mathcal{O}(\epsilon_M^2), \\ \|q\| &\leq (4 + 2C_{IP})\epsilon_M \|v\| + \mathcal{O}(\epsilon_M^2), \end{aligned}$$

in which $C_{IP} \leq \frac{n}{1-n\epsilon_M}$ is the constant appearing in the backward error for the inner product of two vectors.

Govaerts [8] also discusses two other algorithms (BED and BEM) which, when applied to (A.1) are not equivalent to the SM formula as it is commonly implemented, for example in Alg 1 or in [14, p. 487]. Directions for future work include investigating the potential implications of these alternative algorithms for the SM(W) framework.

Appendix B. Related backward error bounds. Although this paper is *not* concerned with the stability of the SMW formula for matrix inversion, in this subsection, we discuss how the backward error bounds derived by Ma, Boutsikas, Ghadiri and Drineas [16] for a computed inverse of B using SMW relate to Higham's question. Although it is common knowledge that multiplying by an explicit inverse

⁶i.e., when S is applied for solving $Ay = b$ in floating-point arithmetic, then there exist a modest constant C_S , a matrix ΔA and a vector δb such that $(A + \Delta A)\hat{y} = b + \delta b$ with $\|\Delta A\|_2 \leq C_S \epsilon_M \|A\|_2$ and $\|\delta b\|_2 \leq C_S \epsilon_M \|b\|_2$.

is not an advisable way of solving a linear system in practice, one might argue that, at-least in theory, SMW could first be applied to invert $A + UV^T$, followed by a multiplication with b . Could the existing error bounds for matrix inversion via SMW then lead to alternative solutions to Higham's question? Here, we present an attempt in this direction that yields a negative answer. We first outline backward error bounds derived in [16] for the low-rank perturbed matrix inversion and then examine their implications when combined with ideal forward error bounds.

PROPOSITION B.1. [16, Thm. 6] Let $\lambda := \|U\|_2\|V\|_2$ and

$$\epsilon_1 := \|\tilde{A}^{-1} - A^{-1}\|_2, \quad (\text{B.1})$$

$$\epsilon_2 := \|Z^{-1} - (I + V^T \tilde{A}^{-1} U)^{-1}\|_2 \quad (\text{B.2})$$

where Z is a computed approximate inverse of $I + V^T \tilde{A}^{-1} U$. If

$$\begin{aligned} \|I + V^T A^{-1} U\|_2 &\leq \beta, \\ \tilde{A} \text{ and } (Z^{-1})^{-1} - V^T \tilde{A}^{-1} U &\text{ are invertible,} \\ \epsilon_1 &< \frac{1}{2\|A\|_2}, \\ \epsilon_2 &< \frac{1}{2(\beta + \lambda \epsilon_1)}, \\ 2(\beta + \lambda \epsilon_1)^2 \epsilon_2 &< \frac{1}{2}. \end{aligned} \quad (\text{B.3})$$

Then

$$\|B - \left(\tilde{A}^{-1} - \tilde{A}^{-1} U Z^{-1} V^T \tilde{A}^{-1}\right)^{-1}\|_2 \leq 2\epsilon_1 \|A\|_2^2 + 4\lambda \epsilon_2 (\beta + \lambda \epsilon_1)^2. \quad (\text{B.4})$$

Note that ϵ_1 and ϵ_2 are absolute forward errors in computing the inverse of A and the inverse of $I + V^T A^{-1} U$, respectively. In particular, ϵ_2 assumes that the addition of matrices and matrix-matrix multiplications in $I + V^T \tilde{A}^{-1} U$ are performed in exact arithmetic such that ϵ_2 solely reflects the forward error in the operation of inverting $I + V^T \tilde{A}^{-1} U$. Assuming further that the update UV^T has a small norm relative to that of A , the paper [16] proves that the backward error (B.4) simplifies to the following bound.

PROPOSITION B.2. [16, Cor. 7] Assume that ϵ_1 , ϵ_2 and $\sigma_{\min}(A)$ are all upper bounded by one and also assume that ϵ_1 and ϵ_2 satisfy the assumptions of Proposition B.1. Additionally, assume that $\lambda \leq \frac{\sigma_{\min}(A)}{2}$. Then, the right-hand side of the bound of eqn. (B.4) simplifies to:

$$\|B - \left(\tilde{A}^{-1} - \tilde{A}^{-1} U Z^{-1} V^T \tilde{A}^{-1}\right)^{-1}\|_2 \leq 2\epsilon_1 \|A\|_2^2 + 8\epsilon_2.$$

Next, let us recall the *ideal* forward error bound for matrix inversion. This follows the derivation in [14, p. 261] where we replace componentwise perturbations with those in the 2-norm. Suppose A is perturbed to $A + \Delta A$ with $\|\Delta A\|_2 \leq \epsilon \|A\|_2$. If $\tilde{A}^{-1}$, an approximate inverse of A , is the exact inverse of a perturbation of A , i.e., if it satisfies $\tilde{A}^{-1} = (A + \Delta A)^{-1}$, then $(A + \Delta A)\tilde{A}^{-1} = \tilde{A}^{-1}(A + \Delta A) = I$ and so

$$\tilde{A}^{-1} = (A + \Delta A)^{-1} = A^{-1} - A^{-1} \Delta A A^{-1} + \mathcal{O}(\epsilon^2)$$

which gives the following standard forward error bound

$$\|\tilde{A}^{-1} - A^{-1}\|_2 \leq \|A^{-1} \Delta A\|_2 \|A^{-1}\|_2 + \mathcal{O}(\epsilon^2) \leq \epsilon \kappa_2(A) \|A^{-1}\|_2 + \mathcal{O}(\epsilon^2).$$

Arguing similarly for the inverse of $I + V^T \tilde{A}^{-1} U$ and replacing ϵ with machine epsilon ϵ_M , we can replace ϵ_1, ϵ_2 in (B.1) and (B.2) by

$$\epsilon_1 \leq \epsilon_M \kappa_2(A) \|A^{-1}\|_2 + \mathcal{O}(\epsilon_M^2) \quad (\text{B.5})$$

$$\epsilon_2 \leq \epsilon_M \kappa_2(Z) \|Z^{-1}\|_2 + \mathcal{O}(\epsilon_M^2) \quad (\text{B.6})$$

which are typically less than one, but could be even larger if A and Z are ill-conditioned. Since (B.5) and (B.6) are ideal bounds, in view of the assumption (B.3), we can see that in practice Propositions B.1 and B.2 are applicable to matrices A for which

$$\kappa_2(A) \leq (2\epsilon_M)^{-1/2}.$$

In addition, incorporating (B.5) and (B.6) into Proposition B.2 yields the following backward error bound

$$\|B - (\tilde{A}^{-1} - \tilde{A}^{-1} U Z^{-1} V^T \tilde{A}^{-1})^{-1}\|_2 \leq \epsilon_M (2\kappa_2(A)^2 \|A\|_2 + 8\kappa_2(Z)^2 \|Z^{-1}\|_2) + \mathcal{O}(\epsilon_M^2).$$

Applying the reverse triangle inequality to $B = A + UV^T$, the definition of λ and the bound on it imposed in Proposition B.2, it is easy to cast the above absolute bound into the following relative backward error bound

$$\frac{\|B - (\tilde{A}^{-1} - \tilde{A}^{-1} U Z^{-1} V^T \tilde{A}^{-1})^{-1}\|_2}{\|B\|_2} \leq \epsilon_M \frac{4\kappa_2(A)^3 \|A\|_2 + 16\kappa_2(Z)^2 \|Z^{-1}\|_2 \|A^{-1}\|_2}{2\kappa_2(A) - 1} + \mathcal{O}(\epsilon_M^2). \quad \blacksquare$$

where the bound is, once more, essentially proportional to the square of the condition number of A .