CSE 6643 Homework 3

Karl Hiner

1 One-upping [25 pts]

Let $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ have full rank. Assume that we have already computed the QR decomposition of $\boldsymbol{A}$. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, we call the matrix $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T$ a rank-1 update of $\boldsymbol{A}$.

(a) [5 pts]

Prove that if $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$, then $\boldsymbol{B}$ is invertible.

Let $\boldsymbol{x} \in \mathbb{R}^m$ be a non-zero vector. Then,

$$\begin{aligned}
\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} &\neq -1 \\
1 &\neq -\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \\
\boldsymbol{v}^T\boldsymbol{x} &\neq -(\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u})(\boldsymbol{v}^T\boldsymbol{x}) \\
\boldsymbol{x} &\neq -\boldsymbol{A}^{-1}\boldsymbol{u}(\boldsymbol{v}^T\boldsymbol{x}) \\
\boldsymbol{A}\boldsymbol{x} &\neq -\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{x} \\
\big(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\big)\boldsymbol{x} &\neq \boldsymbol{0}. \\
\boldsymbol{B}\boldsymbol{x} &\neq \boldsymbol{0}
\end{aligned}$$

Thus, if $\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u} \neq -1$, there are no nontrivial solutions for $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$, and so $\boldsymbol{B}$ is invertible.

(b) [10 pts]

Design an algorithm that provably solves the system of equations $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}$ in $O(m^2)$ operations.

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{B}^{-1}\boldsymbol{b} && \text{(assume } \boldsymbol{B} \text{ is invertible)} \\
&= \big(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\big)^{-1}\boldsymbol{b} && \text{(since } \boldsymbol{B} = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T\text{)} \\
&= \left(\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}\right)\boldsymbol{b} && \text{(Sherman–Morrison formula [1])} \\
&= \boldsymbol{A}^{-1}\boldsymbol{b} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T}{1 + \boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}\boldsymbol{A}^{-1}\boldsymbol{b} && \text{(note: (1a above), } \boldsymbol{B} \text{ invertible } \implies \text{ den} \neq 0\text{)} \\
&= \tilde{\boldsymbol{x}} - \frac{\tilde{\boldsymbol{u}}\boldsymbol{v}^T}{1 + \boldsymbol{v}^T\tilde{\boldsymbol{u}}}\tilde{\boldsymbol{x}} && \text{(let } \tilde{\boldsymbol{x}} \equiv \boldsymbol{A}^{-1}\boldsymbol{b}, \tilde{\boldsymbol{u}} \equiv \boldsymbol{A}^{-1}\boldsymbol{u}\text{)} \\
&= \left[\boldsymbol{I} + \left(\frac{-1}{1 + \boldsymbol{v}^T\tilde{\boldsymbol{u}}}\right)\tilde{\boldsymbol{u}}\boldsymbol{v}^T\right]\tilde{\boldsymbol{x}} && \text{(rearrange)} \\
&= \big(\boldsymbol{I} + \alpha\tilde{\boldsymbol{u}}\boldsymbol{v}^T\big)\tilde{\boldsymbol{x}} && \text{(let } \alpha \equiv \frac{-1}{1 + \boldsymbol{v}^T\tilde{\boldsymbol{u}}}\text{)}
\end{aligned}$$

Thus, we have derived an expression for $\boldsymbol{x}$ in terms of a scalar α and vectors $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{x}}$.

Vectors $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{x}}$ are both defined in terms of $\boldsymbol{A}^{-1} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T$, where we assume $\boldsymbol{R}$ and $\boldsymbol{Q}$ have already been computed. Thus, both vectors can be computed using back substitution in $O(m^2)$. α can then be computed in $O(m)$, and the final expression for $\boldsymbol{x}$ involves a vector-scalar product, a vector outer product, an identity addition, and a matrix-vector multiplication, for a total of

$$O(m) + O(m^2) + O(m) + O(m^2) = O(2m) + O(m^2) = O(m^2)$$

operations.

Here is the algorithm:

1. Solve $\boldsymbol{R}\tilde{\boldsymbol{x}} = \boldsymbol{Q}^T\boldsymbol{b}$ for $\tilde{\boldsymbol{x}}$ using back substitution. ($O(m^2)$)
2. Solve $\boldsymbol{R}\tilde{\boldsymbol{u}} = \boldsymbol{Q}^T\boldsymbol{u}$ for $\tilde{\boldsymbol{u}}$ using back substitution. ($O(m^2)$)
3. Compute the scalar $\alpha = \dfrac{-1}{1 + \boldsymbol{v}^T\tilde{\boldsymbol{u}}}$. ($O(m)$)
4. Finally, compute the solution $\boldsymbol{x} = \big(\boldsymbol{I} + \big((\alpha\tilde{\boldsymbol{u}})\boldsymbol{v}^T\big)\big)\tilde{\boldsymbol{x}}$. ($O(m^2)$)

Since the highest-order term across all steps is $O(m^2)$, the total number of operations is $O(m^2)$.

(c) r-upping [10 pts]

Extend the algorithm from the previous exercise to the case of $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^T$, for $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{m\times r}$ and $r \ll m$. Calculate the asymptotic complexity of the resulting algorithm.

The Sherman-Morrison formula was the key step above. Here, we will use the generalization of that formula, the Woodbury matrix identity [2]. Given the definitions above, the Woodbury matrix identity states that if $\big(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U}\big)$ is invertible,

$$\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\big(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U}\big)^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}.$$

Thus, we can express the solution $\boldsymbol{x}$ as

$$\begin{aligned}
\boldsymbol{x} &= \Big(\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}\big(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U}\big)^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}\Big)\boldsymbol{b} \\
&= \boldsymbol{A}^{-1}\boldsymbol{b} - \boldsymbol{A}^{-1}\boldsymbol{U}\big(\boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U}\big)^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{b} \\
&= \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{U}}\big(\boldsymbol{I} + \boldsymbol{V}^T\tilde{\boldsymbol{U}}\big)^{-1}\boldsymbol{V}^T\tilde{\boldsymbol{x}} && (\text{let } \tilde{\boldsymbol{x}} \equiv \boldsymbol{A}^{-1}\boldsymbol{b}, \tilde{\boldsymbol{U}} \equiv \boldsymbol{A}^{-1}\boldsymbol{U}) \\
&= \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{U}}\boldsymbol{C}^{-1}\boldsymbol{V}^T\tilde{\boldsymbol{x}} && (\text{let } \boldsymbol{C} = \boldsymbol{I} + \boldsymbol{V}^T\tilde{\boldsymbol{U}})
\end{aligned}$$

Note that, since $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times r}$, and $\boldsymbol{V}^T \in \mathbb{R}^{r\times m}$, we can compute the matrix product $\tilde{\boldsymbol{U}}\boldsymbol{V}^T$ using $O(m^2r)$ operations.

Here is the algorithm:

1. Solve $\boldsymbol{R}\tilde{\boldsymbol{x}} = \boldsymbol{Q}^T\boldsymbol{b}$ for $\tilde{\boldsymbol{x}}$ using back substitution (same as step 1 in (b)). ($O(m^2)$)
2. Compute $\tilde{\boldsymbol{U}} = \boldsymbol{A}^{-1}\boldsymbol{U}$ by solving each of the linear systems $\boldsymbol{R}\tilde{\boldsymbol{u}}_i = \boldsymbol{Q}^T\boldsymbol{u}_i$ for $\tilde{\boldsymbol{u}}_i$, where $\tilde{\boldsymbol{u}}_i$ is the ith column of $\tilde{\boldsymbol{U}}$ and $\boldsymbol{u}_i$ is the ith column of $\boldsymbol{U}$. This can be done using back substitution (as in step 2 in (b)) for each of the r columns. ($O(m^2r)$)
3. *I was unable to complete this.* I feel that this is very close. Finding $\boldsymbol{C}^{-1}$ should be possible using another linear system, similarly in $O(m^2r)$ time, and I think it can be done in a way that actually avoids an expensive final computation step by the correct formulation of this linear system (ideally using only $O(mr)$ matrix-vector/vector-vector multiplications).

This solution should be computable in $O(m^2r)$ time.

2 You Factor [25 pts]

In class we have seen that if $\tilde{\boldsymbol{x}} \in \mathbb{R}^m$ is the solution to the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, as computed by unpivoted LU factorization, we have

$$(\boldsymbol{A} + \boldsymbol{E})\tilde{\boldsymbol{x}} = \boldsymbol{b}. \tag{1}$$

For u the unit roundoff error and $\tilde{\boldsymbol{L}}, \tilde{\boldsymbol{U}}$ the LU factors computed in finite precision, we have

$$|\boldsymbol{E}| \leq mu\big(2|\boldsymbol{A}| + 4\big|\tilde{\boldsymbol{L}}\big|\big|\tilde{\boldsymbol{U}}\big|\big) + O(u^2). \tag{2}$$

Here, $|\cdot|$ signifies the element-wise absolute values and $\leq$ is interpreted element-wise, as well. In this problem, we investigate the conclusions from this bound in the case of row-pivoted LU factorization.

(a) [7.5 pts]

Deduce that under row-pivoted LU factorization and taking $\|\cdot\|_\infty$ to signify the vector-infinity norm, we have

$$\|\boldsymbol{E}\|_\infty \leq mu\big(2\|\boldsymbol{A}\|_\infty + 4m\big\|\tilde{\boldsymbol{U}}\big\|_\infty\big) + O(u^2). \tag{3}$$

This prompts us to investigate the growth factor $\rho \coloneqq \frac{\|\boldsymbol{U}\|_\infty}{\|\boldsymbol{A}\|_\infty}$ of row-pivoted LU factorization.

During row-pivoted LU factorization, we pick the maximum value over a column during pivot selection. Thus, every element of $\tilde{\boldsymbol{L}}$ has magnitude ≤ 1, by construction. This leads to the following inequality:

$$\big(\big|\tilde{\boldsymbol{L}}\big|\big|\tilde{\boldsymbol{U}}\big|\big)_{ij} = \sum_{k=1}^{m} \big|\tilde{\boldsymbol{L}}_{ik}\big|\big|\tilde{\boldsymbol{U}}_{kj}\big| \leq m\big|\tilde{\boldsymbol{U}}\big|_{ij} \tag{4}$$

Thus, we have

$$\begin{aligned} |\boldsymbol{E}| &\leq mu\big(2|\boldsymbol{A}| + 4\big|\tilde{\boldsymbol{L}}\big|\big|\tilde{\boldsymbol{U}}\big|\big) + O(u^2) \\ &\leq mu\big(2|\boldsymbol{A}| + 4m\big|\tilde{\boldsymbol{U}}\big|\big) + O(u^2). \end{aligned}$$

Since all operators here apply element-wise, the desired result then follows immediately, by the definition of the vector-infinity norm applied to matrices, which applies an element-wise maximum.

(b) [5 pts]

Verify that the rows $\boldsymbol{u}_i^T, \boldsymbol{a}_i^T$ of $\boldsymbol{U}, \boldsymbol{A}$ satisfy

$$\boldsymbol{u}_i^T = \boldsymbol{a}_i^T - \sum_{j=1}^{i-1} \boldsymbol{L}_{ij}\boldsymbol{u}_j^T. \tag{5}$$

Expressing the LU factorization of $\boldsymbol{A}$ in terms of rows and columns:

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$$

$$\begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \cdots \\ \boldsymbol{a}_m^T \end{bmatrix} = \boldsymbol{L} \begin{bmatrix} \boldsymbol{u}_1^T \\ \boldsymbol{u}_2^T \\ \cdots \\ \boldsymbol{u}_m^T \end{bmatrix}$$

By the definition of matrix multiplication, with $\boldsymbol{X}_i$ denoting the ith row of matrix $\boldsymbol{X}$,

$$\boldsymbol{A}_{ij} = \boldsymbol{L}_i \cdot \left(\boldsymbol{U}^T\right)_j = \sum_{k=1}^{m} \boldsymbol{L}_{ik}\boldsymbol{U}_{kj}.$$

Since $\boldsymbol{L}$ is lower-triangular and $\boldsymbol{U}$ is upper-triangular, $\boldsymbol{L}_{ij} = \boldsymbol{U}_{ji} = 0$ for $i < j$. Thus, we have

$$\begin{aligned}
\boldsymbol{A}_{ij} &= \sum_{k=1}^{i} \boldsymbol{L}_{ik}\boldsymbol{U}_{kj} \\
\boldsymbol{A}_{ij} &= \sum_{k=1}^{i-1} \boldsymbol{L}_{ik}\boldsymbol{U}_{kj} + \boldsymbol{L}_{ii}\boldsymbol{U}_{ij} \\
&= \sum_{k=1}^{i-1} \boldsymbol{L}_{ik}\boldsymbol{U}_{kj} + \boldsymbol{U}_{ij} && \text{(since } \boldsymbol{L}_{ii} = 1\text{)} \\
\boldsymbol{U}_{ij} &= \boldsymbol{A}_{ij} - \sum_{k=1}^{i-1} \boldsymbol{L}_{ik}\boldsymbol{U}_{kj} && \text{(rearranging terms)} \\
\boldsymbol{u}_i^T &= \boldsymbol{a}_i^T - \sum_{k=1}^{i-1} \boldsymbol{L}_{ik}\boldsymbol{u}_k^T && \text{(restate in terms of rows. QED)}
\end{aligned}$$

(c) [5 pts]

Use part (b) to show that $\|\boldsymbol{U}\|_\infty \leq 2^{m-1}\|\boldsymbol{A}\|_\infty$.

Note: I got some guidance from fellow student Samuel Talkington about this problem.

First, we can use the result from (b) to find the max value of the first row of $\boldsymbol{U}$:

$$\begin{aligned}
\left\|\boldsymbol{u}_1^T\right\|_\infty &= \left\|\boldsymbol{a}_1^T - \sum_{j=1}^{0} \boldsymbol{L}_{ij}\boldsymbol{u}_j^T\right\|_\infty \\
&= \left\|\boldsymbol{a}_1^T\right\|_\infty.
\end{aligned}$$

We can use this result to find a bound on the second row:

$$\begin{aligned}
\left\|\boldsymbol{u}_2^T\right\|_\infty &= \left\|\boldsymbol{a}_2^T - \sum_{j=1}^{1} \boldsymbol{L}_{2j}\boldsymbol{u}_j^T\right\|_\infty && \text{(from (b))} \\
&\leq \left\|\boldsymbol{a}_2^T\right\|_\infty + \left\|\sum_{j=1}^{1} \boldsymbol{L}_{2j}\boldsymbol{u}_j^T\right\|_\infty && \text{(triangle inequality, def. of } \infty\text{-norm)} \\
&= \left\|\boldsymbol{a}_2^T\right\|_\infty + \left\|\boldsymbol{L}_{21}\boldsymbol{u}_1^T\right\|_\infty && \text{(only one element in sum)} \\
&\leq \left\|\boldsymbol{a}_2^T\right\|_\infty + \left\|\boldsymbol{u}_1^T\right\|_\infty && \text{(since } \max|\boldsymbol{L}_{ij}| \leq 1\text{)} \\
&= \left\|\boldsymbol{a}_2^T\right\|_\infty + \left\|\boldsymbol{a}_1^T\right\|_\infty && \text{(from above result for row 1)}
\end{aligned}$$

Continuing to the third row:

$$\begin{aligned}
\left\|\boldsymbol{u}_3^T\right\|_\infty &\le \left\|\boldsymbol{a}_3^T\right\|_\infty + \left\|\sum_{j=1}^{2} \boldsymbol{L}_{3j}\boldsymbol{u}_j^T\right\|_\infty \\
&\le \left\|\boldsymbol{a}_3^T\right\|_\infty + \left\|\boldsymbol{L}_{32}\boldsymbol{u}_2^T\right\|_\infty + \left\|\boldsymbol{L}_{31}\boldsymbol{u}_1^T\right\|_\infty && \text{(expand sum, triangle inequality)} \\
&\le \left\|\boldsymbol{a}_3^T\right\|_\infty + \left\|\boldsymbol{u}_2^T\right\|_\infty + \left\|\boldsymbol{u}_1^T\right\|_\infty && \text{(since } \max|\boldsymbol{L}_{ij}| \le 1\text{)} \\
&= \left\|\boldsymbol{a}_3^T\right\|_\infty + \left\|\boldsymbol{a}_2^T\right\|_\infty + 2\left\|\boldsymbol{a}_1^T\right\|_\infty && \text{(from above results for rows 1/2)}
\end{aligned}$$

and the fourth:

$$\begin{aligned}
\left\|\boldsymbol{u}_4^T\right\|_\infty &\le \left\|\boldsymbol{a}_4^T\right\|_\infty + \left\|\boldsymbol{u}_3^T\right\|_\infty + \left\|\boldsymbol{u}_2^T\right\|_\infty + \left\|\boldsymbol{u}_1^T\right\|_\infty \\
&= \left\|\boldsymbol{a}_4^T\right\|_\infty + \left\|\boldsymbol{a}_3^T\right\|_\infty + 2\left\|\boldsymbol{a}_2^T\right\|_\infty + 4\left\|\boldsymbol{a}_1^T\right\|_\infty
\end{aligned}$$

We can see the pattern emerging:

$$\begin{aligned}
\left\|\boldsymbol{u}_i^T\right\|_\infty &\le \left\|\boldsymbol{a}_i^T\right\|_\infty + \sum_{j=1}^{i-1}\left\|\boldsymbol{u}_j^T\right\|_\infty \\
&= \left\|\boldsymbol{a}_i^T\right\|_\infty + \sum_{j=1}^{i-1} 2^{(j-1)}\left\|\boldsymbol{a}_{i-j}^T\right\|_\infty
\end{aligned}$$

Thus, the bound for the mth row is

$$\begin{aligned}
\left\|\boldsymbol{u}_m^T\right\|_\infty &\le \left\|\boldsymbol{a}_m^T\right\|_\infty + \sum_{j=1}^{m-1} 2^{(j-1)}\left\|\boldsymbol{a}_{m-j}^T\right\|_\infty \\
&\le 2^{m-1}\left(\max_i \left\|\boldsymbol{a}_i^T\right\|_\infty\right) \\
&= 2^{m-1}\|\boldsymbol{A}\|_\infty,
\end{aligned}$$

where the last equality is from the definition of the vector ∞-norm applied to matrices. This last row, $\boldsymbol{u}_m^T$, has the largest ∞-norm of all rows, and thus will contain the largest element in the matrix, and so we have shown that $\|\boldsymbol{U}\|_\infty \le 2^{m-1}\|\boldsymbol{A}\|_\infty$.

(d) [7.5 pts]

Consider matrices of the form

$$\boldsymbol{A} = \begin{pmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix}. \tag{6}$$

Derive the growth factor in this case as a function of m. How does this relate to part (c)?

The $\boldsymbol{PA} = \boldsymbol{LU}$ factorization for $\boldsymbol{A}$ is:

$$\begin{pmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ -1 & -1 & -1 & 1 & \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & & & & 1 \\ & 1 & & & 2 \\ & & 1 & & 4 \\ & & & 1 & 8 \\ & & & & 16 \end{pmatrix} \tag{7}$$

The growth factor is then

$$\rho = \frac{\|\boldsymbol{U}\|_\infty}{\|\boldsymbol{A}\|_\infty} = \frac{16}{1} = 16 = 2^4 = 2^{m-1}.$$

In part (c), we showed that an upper bound for the maximum value of $\boldsymbol{U}$ can be related to the matrix dimension and the maximum value of a matrix $\boldsymbol{A}$, with

$$\|\boldsymbol{U}\|_\infty \leq 2^{m-1}\|\boldsymbol{A}\|_\infty.$$

Rearranging terms, we can see that this also provides an upper bound for the growth factor of an $m \times m$ matrix $\boldsymbol{A}$:

$$\frac{\|\boldsymbol{U}\|_\infty}{\|\boldsymbol{A}\|_\infty} \leq 2^{m-1}.$$

Since $\rho = 2^{m-1}$ for $\boldsymbol{A}$, we can conclude that $\boldsymbol{A}$ exemplifies a worst-case growth factor for a square matrix of its size.

3 [25 pts]

Suppose that $\boldsymbol{A} \in \mathbb{K}^{m\times m}$ is strictly column diagonally dominant, meaning that for all $1 \leq k \leq m$,

$$|\boldsymbol{A}_{kk}| > \sum_{j\neq k} |\boldsymbol{A}_{jk}|. \tag{8}$$

Show that if LU factorization with row pivoting is applied to $\boldsymbol{A}$, no row interchange takes place.

To show that no row interchange takes place, we can show the following:

1. Show that $\boldsymbol{A}$'s strict column diagonal property implies no row interchange takes place during the first factorization step. Then the first permutation matrix $\boldsymbol{P}^1 = \boldsymbol{I}$, and we have $\boldsymbol{L}^1\boldsymbol{P}^1\boldsymbol{A} = \boldsymbol{L}^1\boldsymbol{A} = \boldsymbol{U}^1$, with $\boldsymbol{U}^1$ being the first intermediate matrix on the way to $\boldsymbol{U}$ after completing factorization step $k = 1$. Express $\boldsymbol{U}^1$ as

$$\boldsymbol{U}^1 = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{1k>1} \\ \\ \boldsymbol{0} & \boldsymbol{A}^1 \end{bmatrix},$$

   with $\boldsymbol{A}^1 = \boldsymbol{A}_{2:m,2:m} \in \mathbb{K}^{m-1\times m-1}$.

2. Given (1), show that $\boldsymbol{A}^1$ is also strictly diagonally dominant. Then, apply LU facorization with row pivoting to $\boldsymbol{A}^1$, with $\hat{\boldsymbol{P}}\boldsymbol{A}^1 = \hat{\boldsymbol{L}}\hat{\boldsymbol{U}}$. We conclude from (1) that $\hat{\boldsymbol{P}}^1 = \boldsymbol{I}$, and we have $\hat{\boldsymbol{L}}^1\boldsymbol{A}^1 = \hat{\boldsymbol{U}}^1$, where

$$\hat{\boldsymbol{U}}^1 = \begin{bmatrix} \boldsymbol{A}^1_{11} & \boldsymbol{A}^1_{1,2:m} \\ \\ \boldsymbol{0} & \boldsymbol{A}^2 \end{bmatrix},$$

   with $\boldsymbol{A}^2 = \boldsymbol{A}^1_{2:m,2:m} \in \mathbb{K}^{m-2\times m-2}$.

If we show (1) and (2) hold, then we could apply step (1) again to matrix $\boldsymbol{A}^2$ to show that no row interchange takes place at step $k = 2$, and again apply step (2) to show $\boldsymbol{A}^3$ is also strictly diagonally dominant, etc. Thus, if we show that (1) and (2) hold, we can inductively infer that no row interchanges take place during all factorization steps $1 \leq k \leq m-1$.

1. Show that $|\boldsymbol{A}_{kk}| > \sum_{j \neq k} |\boldsymbol{A}_{jk}| \implies \boldsymbol{P}^1 = \boldsymbol{I}$: During the first factorization step, we choose a pivot row $j_1 = \underset{j}{\operatorname{argmax}} |\boldsymbol{A}_{j1}|$.

$$
\begin{aligned}
|\boldsymbol{A}_{kk}| > \sum_{j \neq k} |\boldsymbol{A}_{jk}| \geq \max_{j \neq k} |\boldsymbol{A}_{jk}| \implies \\
|\boldsymbol{A}_{kk}| = \max_{j} |\boldsymbol{A}_{jk}| \implies \\
|\boldsymbol{A}_{11}| = \max_{j} |\boldsymbol{A}_{j1}| \implies \\
1 = \underset{j}{\operatorname{argmax}} |\boldsymbol{A}_{j1}|,
\end{aligned}
$$

   and so we choose $j_1 = 1$. In other words, no rows are swapped during this first step, and $\boldsymbol{P}^1 = \boldsymbol{I}$.

2. Show that $\boldsymbol{A}^1$ is also strictly diagonally dominant.

   For brevity, let $\hat{\jmath} \equiv j + 1$ and $\hat{k} = k + 1$ for $j, k \in \mathbb{N}$, and let $\alpha_k \equiv \dfrac{\boldsymbol{A}_{1k}}{\boldsymbol{A}_{11}}$. Then, after the first factorization step, the kth column of $\boldsymbol{A}^1$ can be defined in terms of $\boldsymbol{A}$ as:

$$
\begin{bmatrix} \boldsymbol{A}^1_{1k} \\ \boldsymbol{A}^1_{2k} \\ \cdots \\ \boldsymbol{A}^1_{kk} \\ \cdots \\ \boldsymbol{A}^1_{(m-1),k} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{2\hat{k}} - \dfrac{\boldsymbol{A}_{21}}{\boldsymbol{A}_{11}} \boldsymbol{A}_{1\hat{k}} \\ \boldsymbol{A}_{3\hat{k}} - \dfrac{\boldsymbol{A}_{31}}{\boldsymbol{A}_{11}} \boldsymbol{A}_{1\hat{k}} \\ \cdots \\ \boldsymbol{A}_{\hat{k}\hat{k}} - \dfrac{\boldsymbol{A}_{\hat{k}1}}{\boldsymbol{A}_{11}} \boldsymbol{A}_{1\hat{k}} \\ \cdots \\ \boldsymbol{A}_{m\hat{k}} - \dfrac{\boldsymbol{A}_{m1}}{\boldsymbol{A}_{11}} \boldsymbol{A}_{1\hat{k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{2\hat{k}} - \alpha_{\hat{k}} \boldsymbol{A}_{21} \\ \boldsymbol{A}_{3\hat{k}} - \alpha_{\hat{k}} \boldsymbol{A}_{31} \\ \cdots \\ \boldsymbol{A}_{\hat{k}\hat{k}} - \alpha_{\hat{k}} \boldsymbol{A}_{\hat{k}1} \\ \cdots \\ \boldsymbol{A}_{m\hat{k}} - \alpha_{\hat{k}} \boldsymbol{A}_{m1} \end{bmatrix}
$$

   So we can express any element of $\boldsymbol{A}^1$ in terms of $\boldsymbol{A}$ as

$$
\boldsymbol{A}^1_{jk} = \boldsymbol{A}_{(j+1),(k+1)} + \frac{\boldsymbol{A}_{1,(k+1)}}{\boldsymbol{A}_{11}} \boldsymbol{A}_{(j+1),1} \equiv \boldsymbol{A}_{\hat{\jmath}\hat{k}} - \alpha_{\hat{k}} \boldsymbol{A}_{j1}.
$$

Thus, for $\boldsymbol{A}^1$ to be strictly diagonally dominant, we must have

$$\left|\boldsymbol{A}^1_{kk}\right| > \sum_{j\neq k}^{m-1} \left|\boldsymbol{A}^1_{jk}\right|$$

$$\left|\boldsymbol{A}_{\hat{k}\hat{k}} - \alpha_{\hat{k}}\boldsymbol{A}_{\hat{k}1}\right| > \sum_{j\neq k}^{m-1} \left|\boldsymbol{A}_{\hat{j}\hat{k}} - \alpha_{\hat{k}}\boldsymbol{A}_{\hat{j}1}\right|$$

$$|\boldsymbol{A}_{kk} - \alpha_k\boldsymbol{A}_{k1}| > \sum_{j>1,j\neq k}^{m} |\boldsymbol{A}_{jk} - \alpha_k\boldsymbol{A}_{j1}|, k > 1.$$

Starting with the LHS, with $k > 1$:

$$\begin{aligned}
|\boldsymbol{A}_{kk} - \alpha_k\boldsymbol{A}_{k1}| &\geq |\boldsymbol{A}_{kk}| - |\alpha_k\boldsymbol{A}_{k1}| && (\boldsymbol{A} \text{ SDD} \rightarrow \boldsymbol{A}_{kk}\boldsymbol{A}_{11} \geq \boldsymbol{A}_{k1}\boldsymbol{A}_{1k}) \\
&> \sum_{j\neq k} |\boldsymbol{A}_{jk}| - |\alpha_k\boldsymbol{A}_{k1}| && (\boldsymbol{A} \text{ is SDD}) \\
&= \sum_{j>1,j\neq k} |\boldsymbol{A}_{jk}| + |\boldsymbol{A}_{1k}| - |\alpha_k\boldsymbol{A}_{k1}| && (\text{move } j = 1 \text{ out of sum}) \\
&= \sum_{j>1,j\neq k} |\boldsymbol{A}_{jk}| + |\alpha_k\boldsymbol{A}_{11}| - |\alpha_k\boldsymbol{A}_{k1}| && (\text{sub for } \boldsymbol{A}_{1k} \text{ using def. of } \alpha) \\
&> \sum_{j>1,j\neq k} |\boldsymbol{A}_{jk}| + \sum_{j>1} |\alpha_k\boldsymbol{A}_{j1}| - |\alpha_k\boldsymbol{A}_{k1}| && (\boldsymbol{A} \text{ is SDD}) \\
&= \sum_{j>1,j\neq k} |\boldsymbol{A}_{jk}| + \sum_{j>1,j\neq k} |\alpha_k\boldsymbol{A}_{j1}| && (\text{combine right two terms}) \\
&= \sum_{j>1,j\neq k} (|\boldsymbol{A}_{jk}| + |\alpha_k\boldsymbol{A}_{j1}|) && (\text{combine sums}) \\
&\geq \sum_{j>1,j\neq k} |\boldsymbol{A}_{jk} - \alpha_k\boldsymbol{A}_{j1}| && (\text{triangle inequality. QED})
\end{aligned}$$

4 Pivoting [25 pts]

(a) [5 pts]

Go to section (a) of the file `HW3_your_code.jl` and implement a function that takes in a matrix `LU` $\in \mathbb{K}^{m\times m}$ containing the upper triangular part of $\boldsymbol{U}$ as well as the strict lower triangular part of $\boldsymbol{L}$, as well as an array `P` $\in \{1,\dots m\}^m$ that encodes the permutation matrix $\boldsymbol{P}$ by `P`$[j] = i \Leftrightarrow \boldsymbol{P_{ij}} = 1$. Your function should not allocate any memory.

(b) [5 pts]

Go to section (b) of the file `HW3_your_code.jl` and implement the unpivoted LU factorization. Check your code by ensuring that the assertions in section a + b of `HW3_driver.jl` do not produce any errors.

(c) [5 pts]

Generate families of random $m \times m$ matrices and vectors of length m. Plot as a function of the size m, the relative error of the solution obtained from your code in parts (a,b) and the growth

factor introduced in problem 2. Report the floating point type used by your program. You can use the code provided in the second homework as a starting point for creating and saving plots.

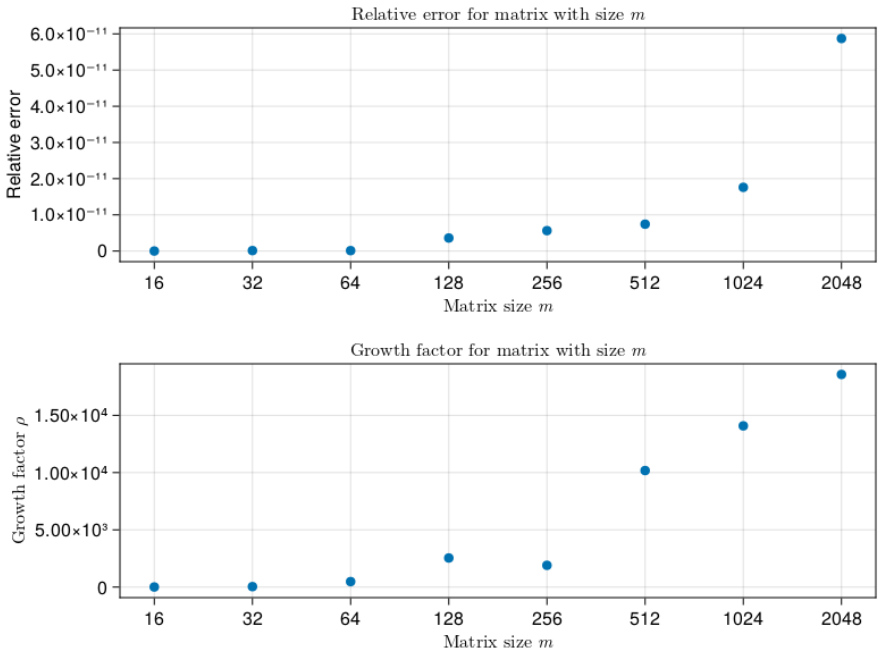

Figure 1 Relative errors and growth factors, without pivoting

See Fig. 1. The floating point type used by my program is `Float64`. All tested matrices are normally distributed with a mean of 0 and a standard deviation of 1, with a unit diagonal offset (to help unpivoted factorization not get stuck), and no row permutations. Relative error was computed as

$$\frac{\|\tilde{\boldsymbol{b}} - \boldsymbol{b}\|_2}{\|\boldsymbol{b}\|_2}.$$

(d) [5 pts]

Go to section (d) of the file `HW3_your_code.jl` and implement the unpivoted LU factorization. Your code should pass the assertions in section (c) of `HW3_driver.jl`. Your function should take the matrix A as an input to modify in place, and return an integer array P according to the specifications of (a). Repeat the experiment of (c) using the pivoted LU factorization.

See Fig. 2. Note that, as expected, row pivoting results in a dramatic reduction in the growth factor, bringing it down from a max of $\approx 2 \cdot 10^4$ to a max of ≈ 30! Also, the relative errors are consistently reduced by 2-4 decimal places (varying across experiments).

Note that the two experiements were run with different random matrices, so we can't make an exact 1:1 comparison.

(e) [5 pts]

Go to section (e) of the file `HW3_driver.jl` and implement a function that takes an integer m as an input and returns an $m \times m$ matrix as introduced in problem 2 (d). Plot the error of the

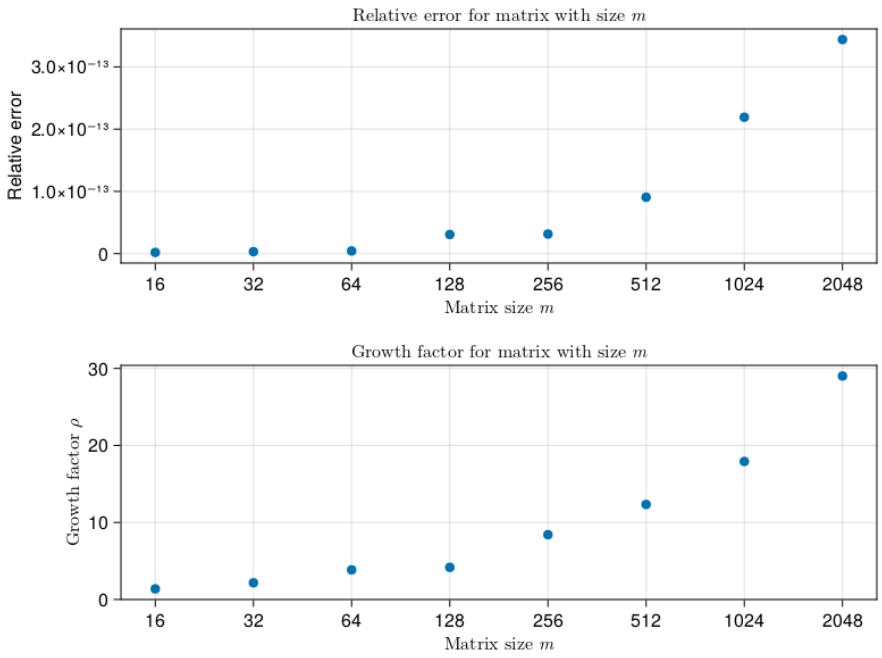

Figure 2 Relative errors and growth factors, with pivoting

solution when solving equations in this matrix as a function of m. Compare the error to the built-in solution (the \ operator). Draw your conclusions from this comparison.

See Fig. 3. For this special kind of maximum-growth matrix, we can see that there is *no difference* between our solution and the built-in \ operator.

When I print out the resulting LU factorization, the results look exactly accurate, so I have convinced myself this is not an error. I suppose this results in part from all numbers in the input matrix and the LU matrices being represented exactly by floating point representations of natural numbers. Beyond this, I am not sure, and I look forward to learning more about this phenomenon! (I have a feeling there's a simple explanation I'm missing...)

References

[1] Jack Sherman and Winifried J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21:124–127, 1950.

[2] Max A. Woodbury. Inverting modified matrices. *Statistical Research Group, Memo.*, Rep. no. 42, 1950.

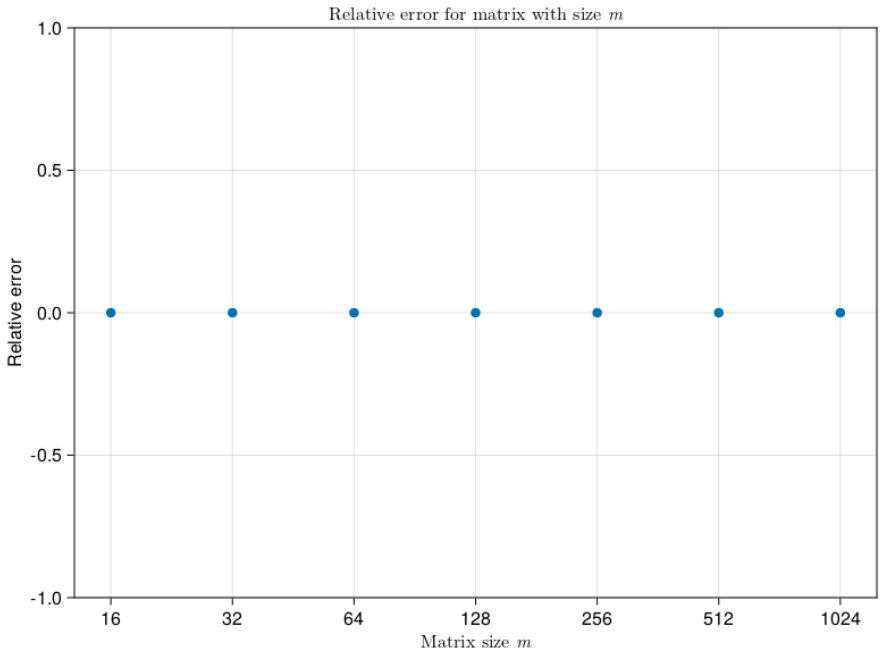

Figure 3 Relative errors and growth factors, with pivoting
Using growth matrix