1 Lectures 5 and 6

1.1 Quiz Lecture 5

Problem 1.

Floating Point Cancellation For which of the following reasons is cancellation in floating-point computation usually bad?

Choice* The digits lost are the least significant. The digits lost are the most significant. The result is usually not exactly representable. Subsequent operations are likely to underflow or overflow.

Solution 1.1. The digits lost during catastrophic cancellation are the most significant ones (ie the left most ones). Since we're losing digits in the resultant answer, our answer is often exactly representable (remember that we have about $-\log(2^{-52}) \approx 16 = -\log(10^{-16})$ digits of relative accuracy).

Problem 2.

Cancellation and Rounding Which of the following statements is true regarding the relationship between rounding and cancellation?

Choice* In computation that is done with exact (not rounded) values, cancellation cannot occur. Catastrophic cancellation occurs when cancellation amplifies rounding errors in the input. Subtracting two rounded numbers will always amplify the error. Cancellation from subtracting two rounded inputs of similar magnitude and sign can be reduced by first converting the rounded inputs to higher precision.

Solution 1.2. Even with exact values, we can find that leading digits are chopped off, if the two numbers agree to a high number of decimal places. Yes, catastrophic cancellation is the phenemonon that takes places when we subtract so much from a number that whatever is shifted upwards to replace the nullified digits is inaccurate, rounding error. This does not mean that subtracting any two numbers will always amplify rounding error, however. Even if you convert numbers to higher precision, this will not change the fact that cancellation will take place – several initial numbers will still match; it will reduce the chances of catastrophic cancellation taking place, however.

Problem 3.

Cancellation and Relative Error Suppose a and b are stored in single precision and agree to four decimal digits. Assume a is known to seven decimal digits and b is known to five decimal digits.

Let c=b−a.

Accounting only for cancellation error, how many decimal digits of accuracy are in c?

How many decimal digits are in c after accounting for the relative error in a and b?

Solution 1.3. In single precision, we have $-\log(2^{-22}) \approx 7$ digits of accuracy. Thus, a and b are (ignoring their relative accuracies initially) known really to 7 digits, so that if subtract them and nullify their first 4 digist, then there are only 3 digits that remain. If we account for relative error in the subtraction, however, then there are only $\min 7-4, 5-4 = 1$ digits that are accurate.

Problem 4.

Changing the RHS You just solved a linear system Ax=b. Unfortunately, the RHS b that you solved it with was wrong.

Worried, you compute ‖Δb‖‖b‖10—12. The condition number of your matrix is about 10000.

What could your worst-case relative error in the solution x be due to your use of the wrong RHS?

Solution 1.4. Just multiply condition number by relative input error.

Problem 5.

Distance to Singularity Which of the following is a good indicator that a matrix is nearly (as measured by the matrix norm) singular?

Choice* Its norm is small. Its determinant is small. Its condition number is large. Its norm is large.

Solution 1.5. $k(\gamma A) = k(A)$. That is, no scaling of a matrix can change its condition number; all the other quantities can change very much, however.

Problem 6.

What is the 2 norm of a diagonal matrix:

Solution 1.6. It is the maximum of teh diagonal entries. The condition number is the ratio of the maximum diagonal entry in absolute value to the minimum one in absolute value.

1.2 Quiz for Lecture 6

Relative Residual Consider a matrix $A = \begin{bmatrix} -9 & -5 \\ 8 & 3 \end{bmatrix}$ and right-hand side vector b=[3—2]. Using the infinity norm, calculate the relative residual if elements of the solution vector xˆ are rounded to one significant digit. Include at least three significant digits in your answer.

Solution 1.7. $\|A\| = 14$. Let $\hat{x} = A^{-1}b = \begin{bmatrix} -0.08 \\ 0.5 \end{bmatrix}$ if you round to 1 digit. Then $r = A\hat{x} - b = \begin{bmatrix} -0.22 \\ 0.14 \end{bmatrix}$

Meaning that $\|r\| = 0.22$. And $\|x\| = 0.5$

$$\frac{\|r\|}{\|x\|\|A\|} = \frac{0.22}{0.5 * 15}$$

which is correct.

Problem 2.

A stupid definition question.

Problem 3.

Gaussian Elimination In the following questions, consider Gaussian elimination with the prescribed pivoting strategy to generate the lower (L) and upper (U) triangular factors for the following matrix,

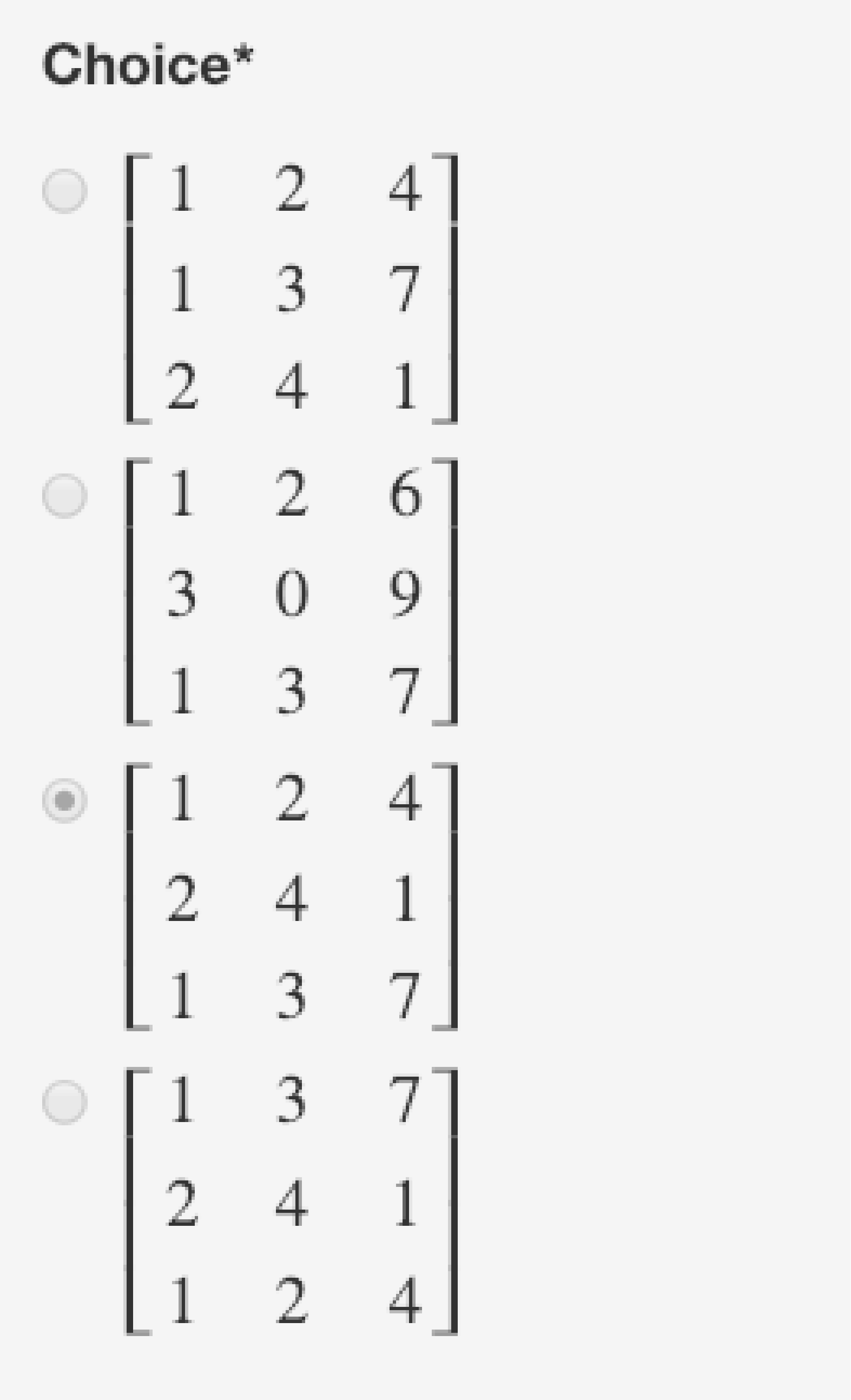
$$A = \begin{bmatrix} 2 & 1 & 3 \\ 2 & 4 & 8 \\ 4 & -7 & 4 \end{bmatrix}$$

Know that with pivoting, we must first swap rows 1 and 3 and then perform gaussian elimination.

Problem 4.

Existence of LU Decomposition with no Pivoting For which of the following matrices does a LU factorization without pivoting not exist?

Solution 1.8. I have



Choice*

○ $\begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 7 \\ 2 & 4 & 1 \end{bmatrix}$

○ $\begin{bmatrix} 1 & 2 & 6 \\ 3 & 0 & 9 \\ 1 & 3 & 7 \end{bmatrix}$

● $\begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 1 \\ 1 & 3 & 7 \end{bmatrix}$

○ $\begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 1 \\ 1 & 2 & 4 \end{bmatrix}$

Sufficient Condition: If a pivot is 0 (assuming that we don't pivot the matrix when performing gaussian elimination), then the matrix will fail to provide an

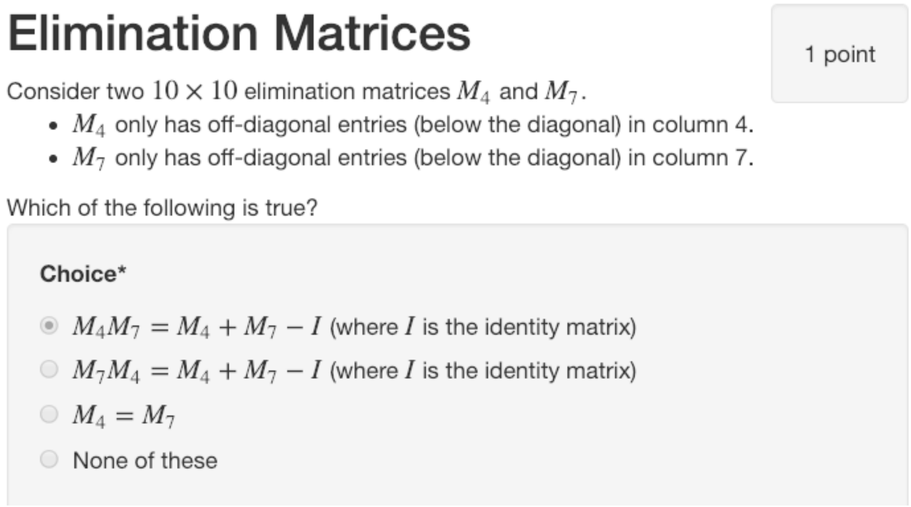
LU factorization.

Problem 5.

Just apply

$$\text{rel error} \leq k(A)\frac{\|r\|}{\|A\|\|x\|}$$

Problem 6.



Elimination Matrices

1 point

Consider two 10×10 elimination matrices M_4 and M_7.

- M_4 only has off-diagonal entries (below the diagonal) in column 4.
- M_7 only has off-diagonal entries (below the diagonal) in column 7.

Which of the following is true?

Choice*

● $M_4M_7 = M_4 + M_7 - I$ (where I is the identity matrix)

○ $M_7M_4 = M_4 + M_7 - I$ (where I is the identity matrix)

○ $M_4 = M_7$

○ None of these

Note that elimination matrices must progress from left to right, meaning that elimination matrices M_1 and M_2 must be such that the off diagonal entry of M_1 is left of the off diagonal entry of M_2, if we want M_1M_2 to merge.

The following is the definition of the residual.

$$r = b - A\hat{x}.$$

Remark 1.9. The residual itself does not reveal much. Suppose we calculate $r = b - Ax$. Now solve for $kAx = kb$ and the residual required to solve that is k times as great. This is why we define the relative residual:

$$\frac{\|r\|}{\|A\| \cdot \|\hat{x}\|}$$

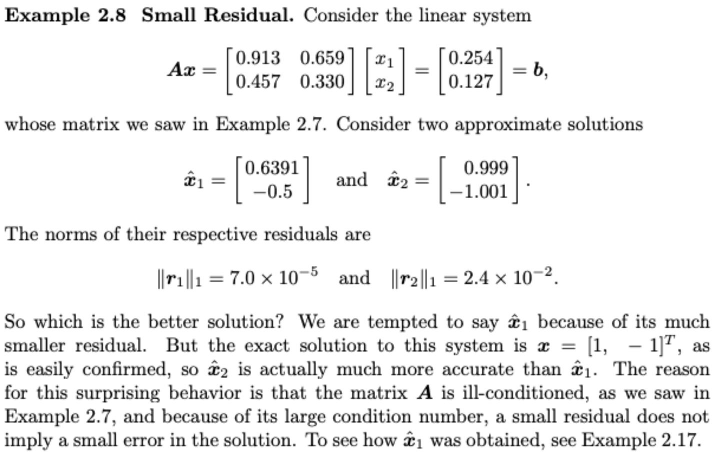
We can obtain a bound on the relative forward error required to solve $Ax = b$ in terms of r.

$$\|\Delta \boldsymbol{x}\| = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| = \|\boldsymbol{A}^{-1}(\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b})\| = \| - \boldsymbol{A}^{-1}\boldsymbol{r}\| \leq \|\boldsymbol{A}^{-1}\| \cdot \|\boldsymbol{r}\|.$$

Dividing both sides by $\|\hat{\boldsymbol{x}}\|$ and using the definition of $\operatorname{cond}(\boldsymbol{A})$, we then have

$$\frac{\|\Delta \boldsymbol{x}\|}{\|\hat{\boldsymbol{x}}\|} \leq \operatorname{cond}(\boldsymbol{A}) \frac{\|\boldsymbol{r}\|}{\|\boldsymbol{A}\| \cdot \|\hat{\boldsymbol{x}}\|}.$$

Remark 1.10. This bound tells us that if the residual is small and the matrix and well conditioned, then the relative error is low.



Example 2.8 Small Residual. Consider the linear system

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 0.913 & 0.659 \\ 0.457 & 0.330 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.254 \\ 0.127 \end{bmatrix} = \boldsymbol{b},$$

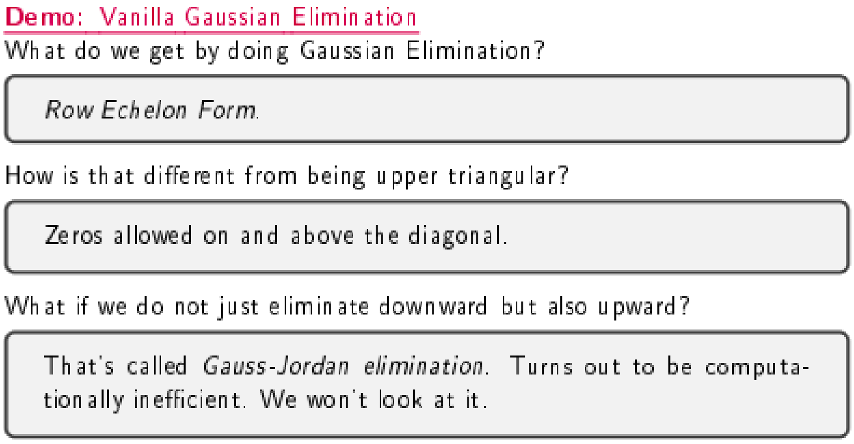
whose matrix we saw in Example 2.7. Consider two approximate solutions

$$\hat{\boldsymbol{x}}_1 = \begin{bmatrix} 0.6391 \\ -0.5 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{x}}_2 = \begin{bmatrix} 0.999 \\ -1.001 \end{bmatrix}.$$

The norms of their respective residuals are

$$\|\boldsymbol{r}_1\|_1 = 7.0 \times 10^{-5} \quad \text{and} \quad \|\boldsymbol{r}_2\|_1 = 2.4 \times 10^{-2}.$$

So which is the better solution? We are tempted to say $\hat{\boldsymbol{x}}_1$ because of its much smaller residual. But the exact solution to this system is $\boldsymbol{x} = [1, \quad -1]^T$, as is easily confirmed, so $\hat{\boldsymbol{x}}_2$ is actually much more accurate than $\hat{\boldsymbol{x}}_1$. The reason for this surprising behavior is that the matrix $\boldsymbol{A}$ is ill-conditioned, as we saw in Example 2.7, and because of its large condition number, a small residual does not imply a small error in the solution. To see how $\hat{\boldsymbol{x}}_1$ was obtained, see Example 2.17.



Demo: Vanilla Gaussian Elimination

What do we get by doing Gaussian Elimination?

Row Echelon Form.

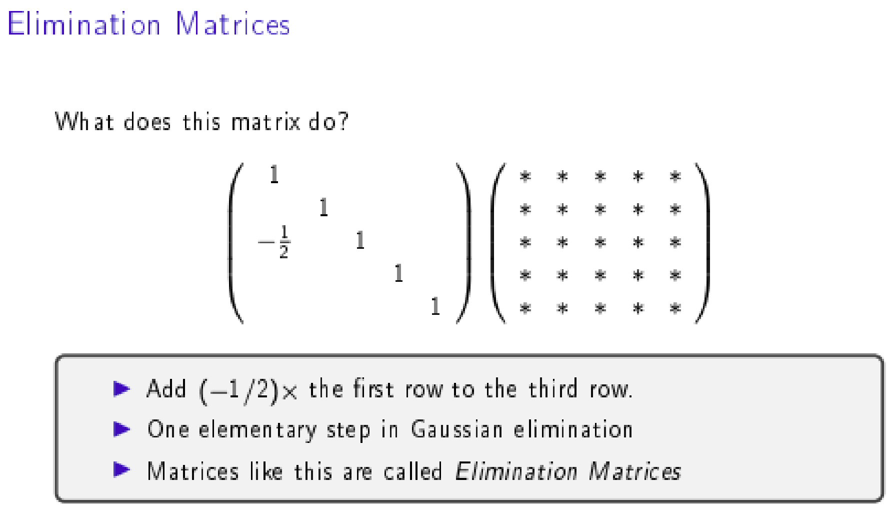
How is that different from being upper triangular?

Zeros allowed on and above the diagonal.

What if we do not just eliminate downward but also upward?

That's called *Gauss-Jordan elimination*. Turns out to be computationally inefficient. We won't look at it.

Remark 1.11. Also note that a matrix is in row echelon form if the first non-zero entry of each row (what was the pivot during gaussian elimination) is to the right of the first non-zero entry of any preceding row; moreover, entries in rows above the pivot (but in the same column) must be 0.



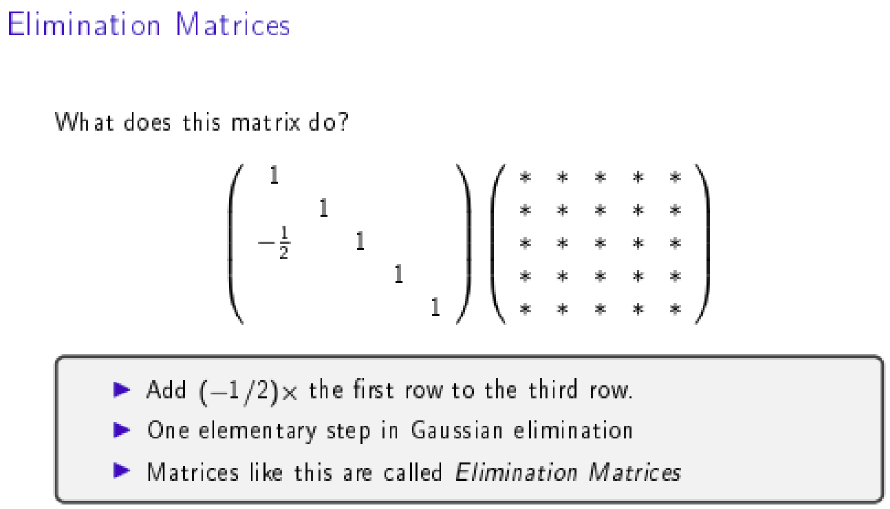
Elimination Matrices

What does this matrix do?

$$\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ -\frac{1}{2} & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

- Add $(-1/2)\times$ the first row to the third row.
- One elementary step in Gaussian elimination
- Matrices like this are called *Elimination Matrices*

Remark 1.12. If we add k to the identity matrix at entry i, j, and left multiply the resultant matrix C by some matrix of interest A, then the result is to take the j th row of A multiply it by k and then add it to i. We can undo this process by using the same matrix but, in place of k, using $-k$. This second matrix is the inverse to the elimination matrix C.



Elimination Matrices

What does this matrix do?

$$\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ -\frac{1}{2} & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

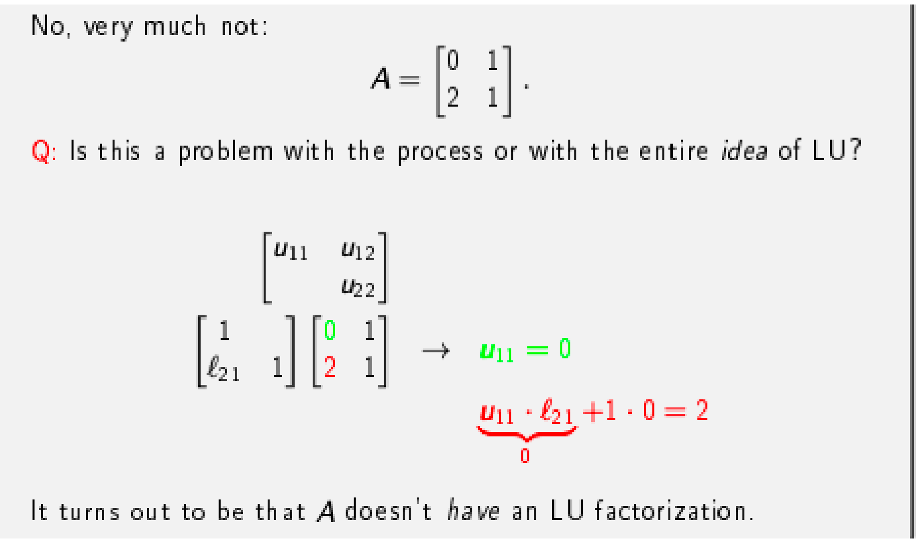
- Add $(-1/2)\times$ the first row to the third row.
- One elementary step in Gaussian elimination
- Matrices like this are called *Elimination Matrices*

Remark 1.13. Suppose that we multiply A by an elimination matrix M_1, then by M_2 up to M_l, where M_l is the last matrix required to turn A into Row Echelon Form. Eventually, we will have

$$(M_l \dots M_1)A = U \implies A = (M_l \dots M_1)^{-1}U$$

At first glance, this is okay, because it turns out that left multiplication of an elimination matrix X by Y such that X has a non-zero off diagonal at column i and Y has a non-zero off diagonal at column j where $i < j$ results in an elimination matrix that just merges X and Y[1]

For whatever reason, pivoting foils this attempt:



No, very much not:

$$A = \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}.$$

Q: Is this a problem with the process or with the entire *idea* of LU?

$$\begin{bmatrix} u_{11} & u_{12} \\ & u_{22} \end{bmatrix}$$

$$\begin{bmatrix} 1 & \\ \ell_{21} & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix} \to \quad u_{11} = 0$$

$$\underbrace{u_{11} \cdot \ell_{21}}_{0} + 1 \cdot 0 = 2$$

It turns out to be that A doesn't *have* an LU factorization.

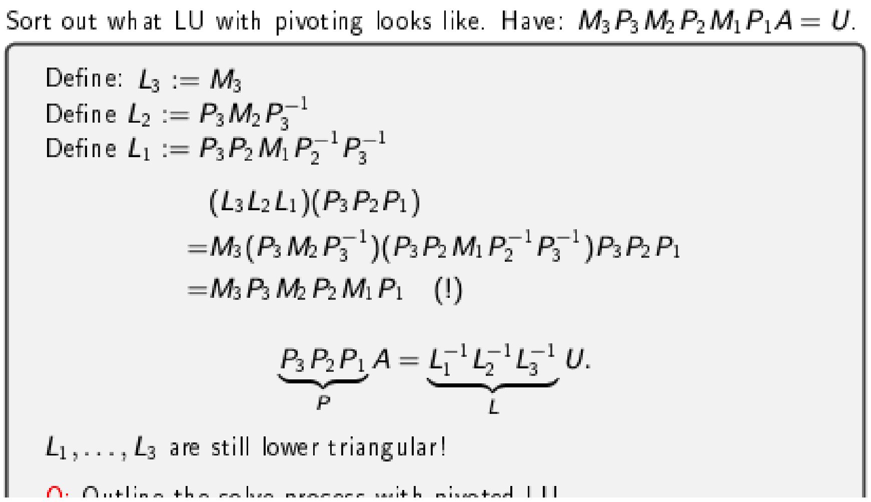
The solution is to repeatedly apply permutations to A (in the form of permutation matrices) so that the pivot is the largest element in terms of absolute value in its column.

Thus, we now have

$$(M_l P_l \dots M_1 P_1)A = U \implies A = (M_l P_l \dots M_1 P_1)^{-1} U$$

However, what should be L above is not always left triangular. It can be shown that a factorization of $(M_l P_l \dots M_1 P_1)^{-1}$ does, however, give us a lower triangular system.

[1] Note that merging also takes place if we multiply two elimination matrices that have their off diagonal non-zero entry in the same column as each other.



Sort out what LU with pivoting looks like. Have: $M_3P_3M_2P_2M_1P_1A = U$.

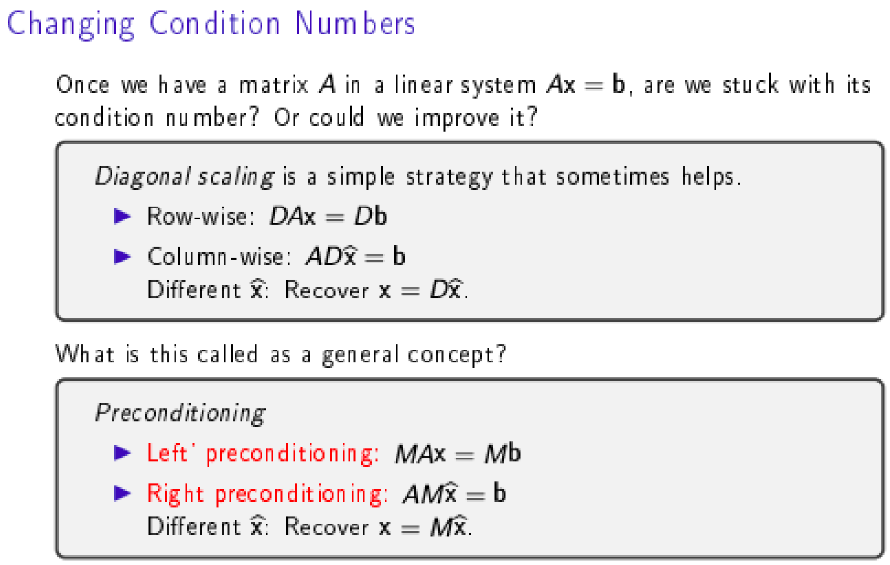
Define: $L_3 := M_3$
Define $L_2 := P_3M_2P_3^{-1}$
Define $L_1 := P_3P_2M_1P_2^{-1}P_3^{-1}$

$$\begin{aligned}&(L_3L_2L_1)(P_3P_2P_1)\\ =&M_3(P_3M_2P_3^{-1})(P_3P_2M_1P_2^{-1}P_3^{-1})P_3P_2P_1\\ =&M_3P_3M_2P_2M_1P_1 \quad (!)\end{aligned}$$

$$\underbrace{P_3P_2P_1}_{P}A = \underbrace{L_1^{-1}L_2^{-1}L_3^{-1}}_{L}U.$$

$L_1,\dots,L_3$ are still lower triangular!

Q: Outline the solve process with pivoted LU



Changing Condition Numbers

Once we have a matrix A in a linear system $A\mathbf{x} = \mathbf{b}$, are we stuck with its condition number? Or could we improve it?

Diagonal scaling is a simple strategy that sometimes helps.

- Row-wise: $DA\mathbf{x} = D\mathbf{b}$
- Column-wise: $AD\widehat{\mathbf{x}} = \mathbf{b}$
  Different $\widehat{\mathbf{x}}$: Recover $\mathbf{x} = D\widehat{\mathbf{x}}$.

What is this called as a general concept?

Preconditioning

- Left' preconditioning: $MA\mathbf{x} = M\mathbf{b}$
- Right preconditioning: $AM\widehat{\mathbf{x}} = \mathbf{b}$
  Different $\widehat{\mathbf{x}}$: Recover $\mathbf{x} = M\widehat{\mathbf{x}}$.

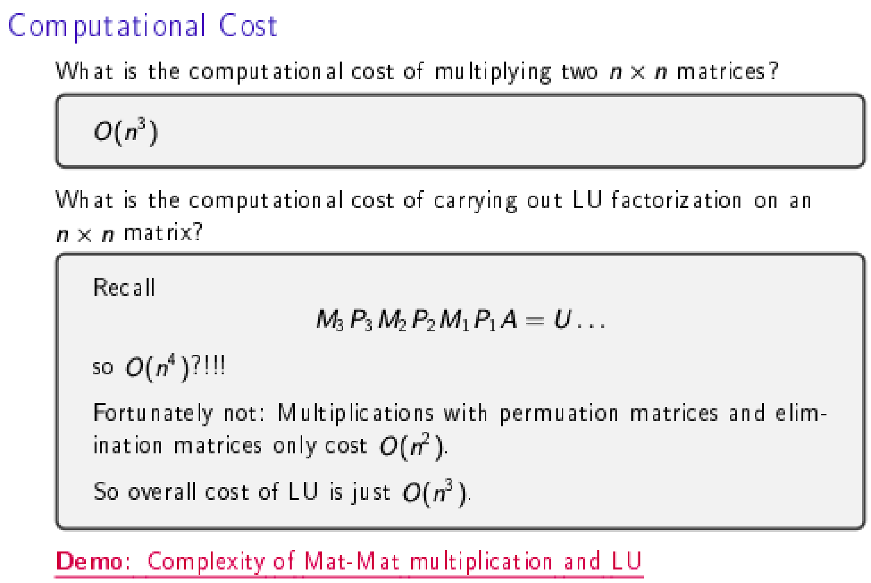
Remark 1.14. Suppose that D above satisfies $k(D) \approx 1$. Then

$$k(DA) = \|DA\|\|(DA)^{-1}\| \leq \|D\|\|A\|\|A^{-1}\|\|D^{-1}\| \leq k(A)$$

so that the condition number of $K(DA)$ is no greater than the condition number of A.

Assuming that D is invertible, then the set of x satisfying $Ax = b$ is precisely the set of x satisfying $Ax = b$. Left multiplication by D of A is called, understandably, left preconditioning and scales A in a row-wise manner; right multiplication by D of A is called right preconditioning.

Remark 1.15.



Computational Cost

What is the computational cost of multiplying two $n \times n$ matrices?

$O(n^3)$

What is the computational cost of carrying out LU factorization on an $n \times n$ matrix?

Recall

$$M_3P_3M_2P_2M_1P_1A = U \dots$$

so $O(n^4)$?!!!

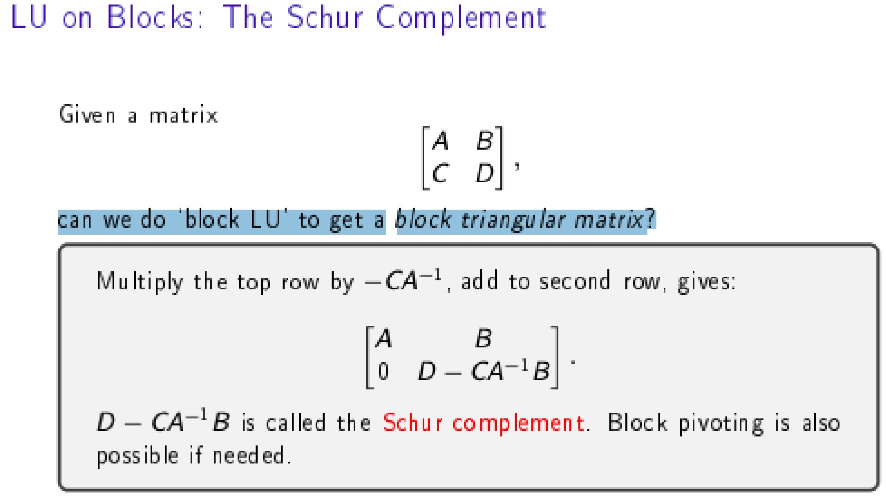
Fortunately not: Multiplications with permuation matrices and elimination matrices only cost $O(n^2)$.

So overall cost of LU is just $O(n^3)$.

Demo: Complexity of Mat-Mat multiplication and LU

Multiplication by a permutation matrix is only an n operation, since it involves switching rows. Multiplication by an elimination matrix simply involves scaling one row and mulitplying it by another, and this process is done at most n times for any one elimination matrix (making it $O(n^2)$ as well). Since these transformations are applied at most n times, the process of getting a matrix into LU form is only $O(n^3)$.

Remark 1.16.



LU on Blocks: The Schur Complement

Given a matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

can we do 'block LU' to get a *block triangular matrix*?

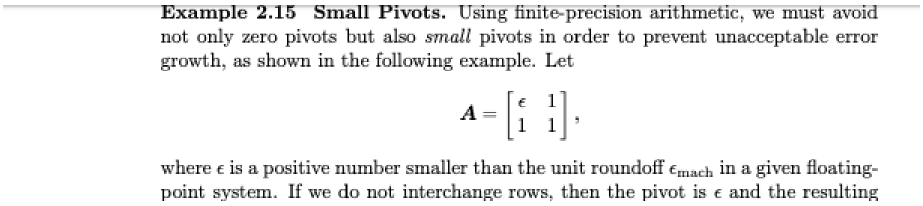
Multiply the top row by $-CA^{-1}$, add to second row, gives:

$$\begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix}.$$

$D - CA^{-1}B$ is called the Schur complement. Block pivoting is also possible if needed.

Not sure why this is significant.

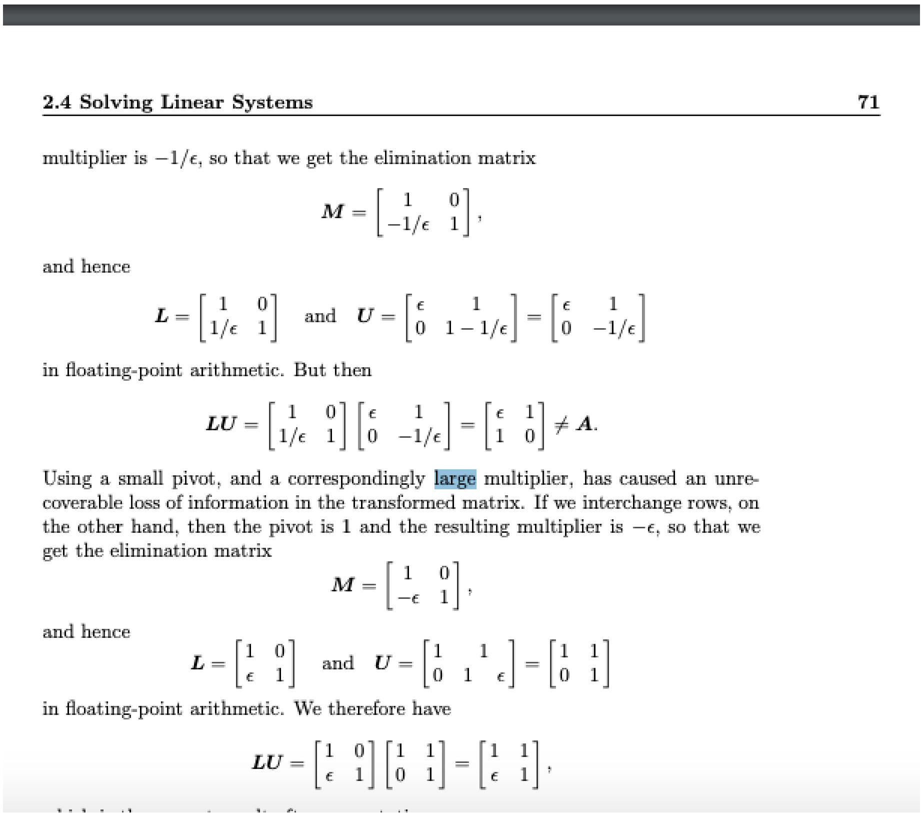
Remark 1.17. Unresolved



Example 2.15 Small Pivots. Using finite-precision arithmetic, we must avoid not only zero pivots but also *small* pivots in order to prevent unacceptable error growth, as shown in the following example. Let

$$\boldsymbol{A} = \begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix},$$

where ϵ is a positive number smaller than the unit roundoff ϵ_{mach} in a given floating-point system. If we do not interchange rows, then the pivot is ϵ and the resulting



2.4 Solving Linear Systems **71**

multiplier is $-1/\epsilon$, so that we get the elimination matrix

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ -1/\epsilon & 1 \end{bmatrix},$$

and hence

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{U} = \begin{bmatrix} \epsilon & 1 \\ 0 & 1 - 1/\epsilon \end{bmatrix} = \begin{bmatrix} \epsilon & 1 \\ 0 & -1/\epsilon \end{bmatrix}$$

in floating-point arithmetic. But then

$$\boldsymbol{LU} = \begin{bmatrix} 1 & 0 \\ 1/\epsilon & 1 \end{bmatrix} \begin{bmatrix} \epsilon & 1 \\ 0 & -1/\epsilon \end{bmatrix} = \begin{bmatrix} \epsilon & 1 \\ 1 & 0 \end{bmatrix} \neq \boldsymbol{A}.$$

Using a small pivot, and a correspondingly large multiplier, has caused an unrecoverable loss of information in the transformed matrix. If we interchange rows, on the other hand, then the pivot is 1 and the resulting multiplier is $-\epsilon$, so that we get the elimination matrix

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ -\epsilon & 1 \end{bmatrix},$$

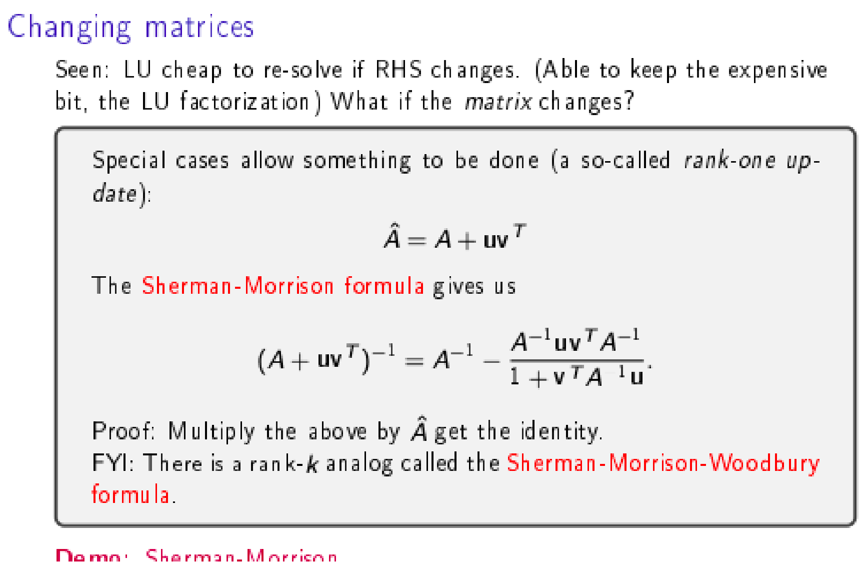
and hence

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \quad \epsilon \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

in floating-point arithmetic. We therefore have

$$\boldsymbol{LU} = \begin{bmatrix} 1 & 0 \\ \epsilon & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix},$$

Remark 1.18. Notice that if we already have an LU factorization, then computing a rank 1 update is just an $O(n^2)$ operation.



Changing matrices

Seen: LU cheap to re-solve if RHS changes. (Able to keep the expensive bit, the LU factorization) What if the *matrix* changes?

Special cases allow something to be done (a so-called *rank-one update*):

$$\hat{A} = A + \mathbf{u}\mathbf{v}^T$$

The Sherman-Morrison formula gives us

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{\ 1}\mathbf{u}}.$$

Proof: Multiply the above by $\hat{A}$ get the identity.
FYI: There is a rank-k analog called the Sherman-Morrison-Woodbury formula.

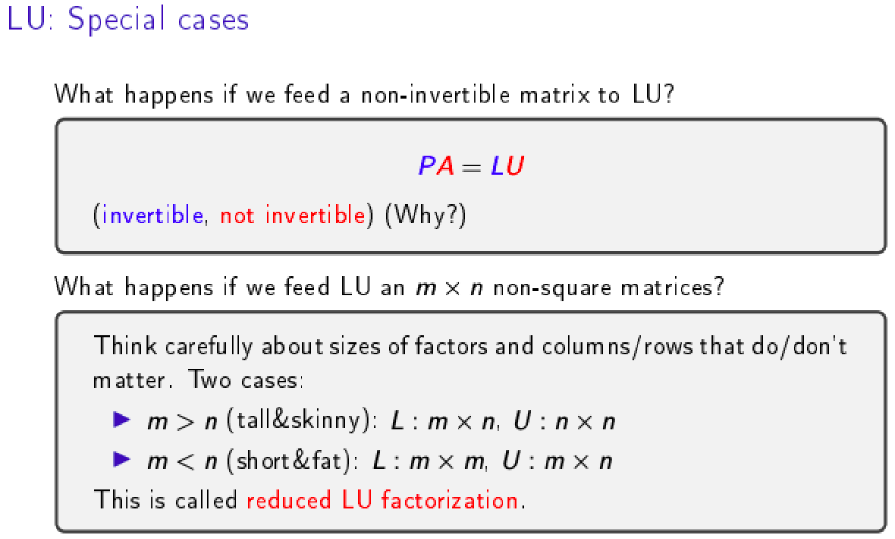
Demo: Sherman-Morrison

For

$$\left(A + uv^T\right)^{-1} b = A^{-1}b - \frac{\left(A^{-1}u\right) v^T A^{-1} b}{1 + v^T A^{-1} u}$$

And $A^{-1}x$ for any x is an $O(n^2)$ operation. The only other operation in this formula is to compute a dot product.

Remark 1.19.



LU: Special cases

What happens if we feed a non-invertible matrix to LU?

$$PA = LU$$

(invertible, not invertible) (Why?)

What happens if we feed LU an $m \times n$ non-square matrices?

Think carefully about sizes of factors and columns/rows that do/don't matter. Two cases:

- $m > n$ (tall&skinny): $L : m \times n$, $U : n \times n$
- $m < n$ (short&fat): $L : m \times m$, $U : m \times n$

This is called reduced LU factorization.

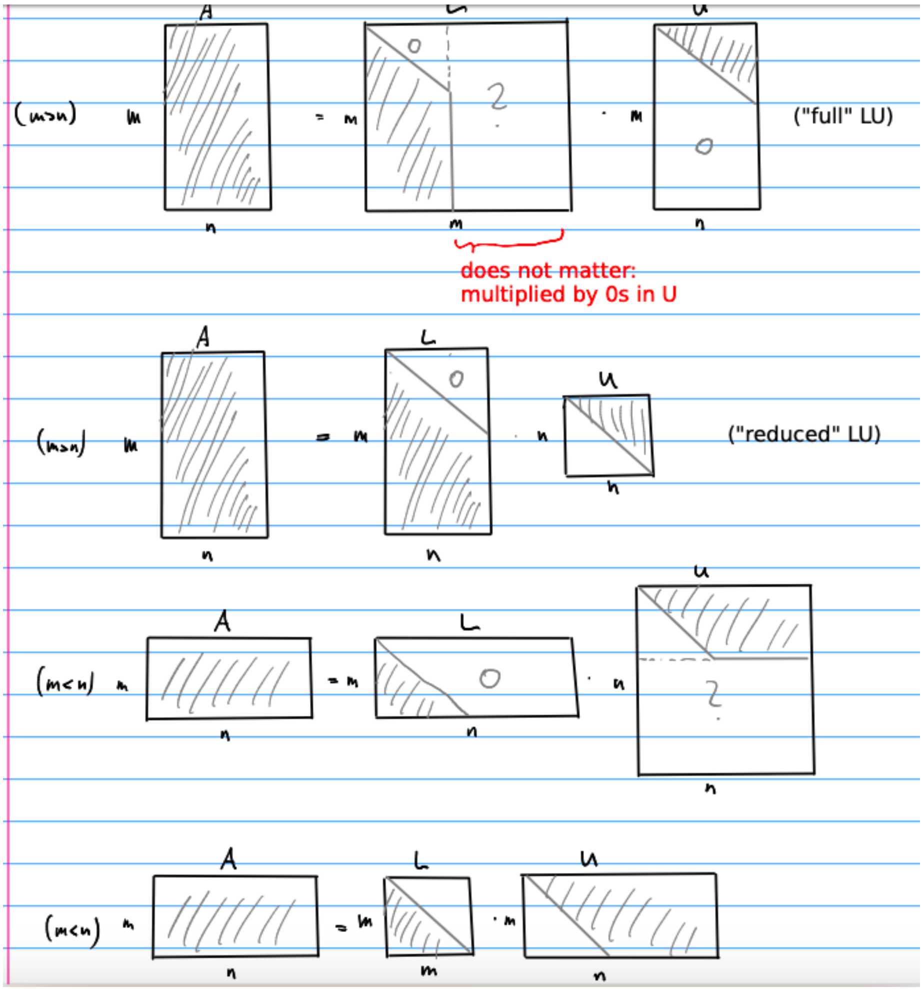
A matrix A always admits an LU factorization, even if A is singular. First, observe that every column of A must contain at least one non-zero number – or else, why would the column be part of A. Thus, if a pivot entry does not contain a non-zero value, we can rotate rows so that the pivot entry does have a non-zero value. We then can apply elimination matrices, as needed, until all row values in the pivot's column are 0.

The foregoing tells us that there still exists a sequence of permutations and elimination matrices that bring A into upper triangular form. Since these matrices are each invertible, it follows that L is still invertible. Since $|PA| = 0$, we must have, therefore, $|U| = 0$, which means that 0 must occupy some diagonal entry of U.

Why? A matrix fails to be invertible iff 0 is an eigenvalue. 0 is an eigenvalue iff some entry on the diagonal is 0, because the eigenvalues of a triangular matrix are precisely its diagonal entries.

Why are permutation matrices invertible? Group theory promises us that some power of a permutation brings it back to the identity. That is $P^k = I$ for some I. Thus $P^{-1} = P^{k-1}$.

Remark 1.20. Why can we take an LU decomposition and then reduce it, as shown below?



If $m > n$, applying elimination matrices that use row $n+1$ is a no-op, since the use of row n has (along with prior use of rows $1 \ldots n-1$) already made row $n+1$ be entirely 0. or if $n > m$. As a consequence, in the unreduced LU factorization (the first and third pictures above), we see that every column in $\{n+1 \dots m\}$ is really just 0 except at a diagonal entry (where it is 1). As the picture points out, however, determinining what exists in the unreduced L is not useful, since $A = LU$ where $L = [QB]$ and $U = \begin{bmatrix} Q' \\ B' \end{bmatrix}$ where $Q = m \times n$, $B = m \times m - n$, $Q' = n \times n$ and $B' = m - n \times n$. Since $BB' = \mathbb{0}$, there is no need to store B or B'.

Using similar reasoning, we can understand the case that $n > m$.

2 Lecture 7

Least Squares

2.1 Quiz

Problem 1.

Gaussian Elimination with Partial Pivoting Under what conditions will Gaussian elimination with partial pivoting succeed in computing the LU factorization of an n×n matrix A?

Solution 2.1. Always.

Problem 2.

Basic Linear Algebra Subprograms Which of the following is true?

Select all that apply: Most BLAS level 3 functions can be composed out of multiple lower level functions BLAS level 1 functions do not involve matrices BLAS level 2 functions generally do more arithmetic operations per matrix/vector entry than level 3 functions BLAS level 2 functions do not involve vectors

Solution 2.2. Recall that level 1 is vector-vector operations; level 2 is matrix vector operations and level 3 is matrix matrix operations.

Problem 4.

Rank One Change If an n×n linear system Ax=b has already been solved by LU factorization, and then the matrix is changed by adding a matrix of rank 1, how much work is required to solve the new linear system with the same right-hand side?

Solution 2.3. The Sherman Morrison Formula tells us that this cost is $O(n^2)$.

Problem 5.

If we reduce a 9×24 matrix, then the shapes work out to be:

Solution 2.4.

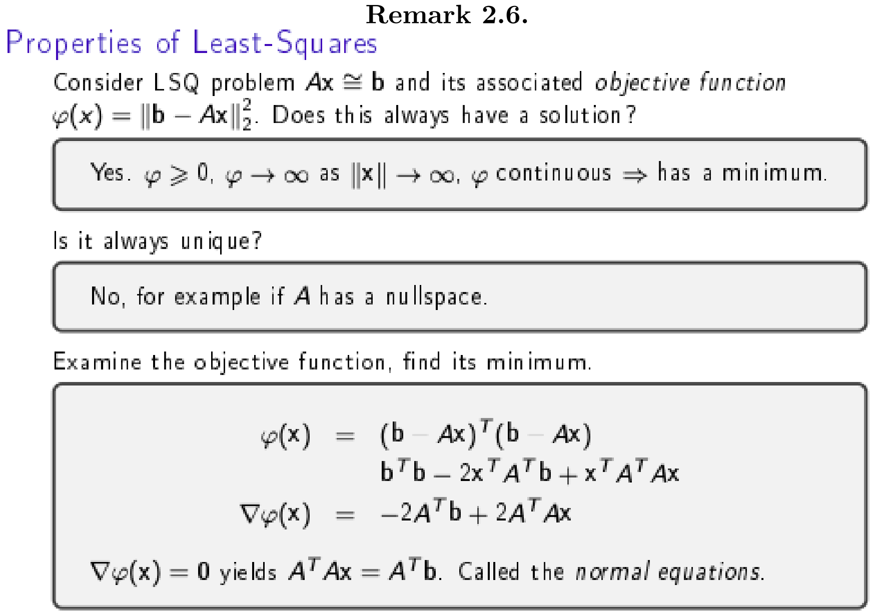
$$(9 \times 9) \times (9 \times 24)$$

Problem 6.

Which problem requires the largest amount of work:
kkkkkk

Remark 2.5. We assume that we work with tall, skinny matrices that have full column rank.

Remark 2.6.



Properties of Least-Squares

Consider LSQ problem $Ax \cong \mathbf{b}$ and its associated *objective function* $\varphi(\mathbf{x}) = \|\mathbf{b} - A\mathbf{x}\|_2^2$. Does this always have a solution?

Yes. $\varphi \geqslant 0$, $\varphi \to \infty$ as $\|\mathbf{x}\| \to \infty$, φ continuous $\Rightarrow$ has a minimum.

Is it always unique?

No, for example if A has a nullspace.

Examine the objective function, find its minimum.

$$\begin{aligned} \varphi(\mathbf{x}) &= (\mathbf{b} - A\mathbf{x})^T(\mathbf{b} - A\mathbf{x}) \\ & \quad \mathbf{b}^T\mathbf{b} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{x}^T A^T A\mathbf{x} \\ \nabla\varphi(\mathbf{x}) &= -2A^T\mathbf{b} + 2A^T A\mathbf{x} \end{aligned}$$

$\nabla\varphi(\mathbf{x}) = \mathbf{0}$ yields $A^T A\mathbf{x} = A^T\mathbf{b}$. Called the *normal equations*.

The textbook proves that there is always a unique vector $y \in \text{Span}(A)$ such that $\phi(y) = \|b - y\|^2$ is minimal; as a consequence, there is at least one vector $x \in \mathbb{R}^m$ where A is $\mathbb{R}^{n \times m}$ that minimizes $Ax \approx b$. This vector x is unique iff A is full rank.

Definition 2.7. A matrix P is a projection if $P^2 = P$. A matrix is an orthogonal projection if $P^2 = P$ and $P^T = P$.

Proposition 2.8. If P is an orthogonal projection, then the span of $P_\perp = (I - P)$ is orthogonal to the span of P.

Proof. Given $x, y \in \mathbb{R}^n$, we see that

$$\begin{aligned} &\langle Px, (I - P)y\rangle \\ &= \langle Px, y - Py\rangle \\ &= \langle Px, y\rangle - \langle Px, Py\rangle \\ &= \langle Px, y\rangle - \langle x, Py\rangle \\ &= 0 \end{aligned}$$

□

Corollary 2.9. Given an orthogonal projection P, any vector x can be expressed as $x = Px + P_\perp x$.

Proposition 2.10. The vector x satisfying $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$ is precisely the x such that $Ax = Pb$ where P is a projection onto A.

Proof. Note that in what follows, all norms refer to the 2 norm.

$$\|Ax - b\| = \|P(Ax - b) + P_{\perp}(Ax - b)\|$$

Since P and $P_{\perp}$ map to orthogonal subspaces, we can apply the Pythagorean theorem

$$\begin{aligned} &= \|P(Ax - b)\| + \|P_{\perp}(Ax - b)\| \\ &= \|P(Ax - b)\| + \|-P_{\perp}b\| \\ &= \|P(Ax - b)\| + \|P_{\perp}b\| \\ &= \|(Ax - Pb)\| + \|P_{\perp}b\| \end{aligned}$$

The RHS is fixed, so we can only minimize the LHS

□

Corollary 2.11. The x aforementioned is $(A^TA)^{-1}A^Tb$

Proof.

$$\begin{aligned} Ax &= Pb \\ \iff A^TAx &= A^TPb \\ \iff A^TAx &= (PA)^Tb \\ \iff A^TAx &= (A)^Tb \\ \iff x &= (A^TA)^{-1}(A)^Tb \end{aligned}$$

□

Proposition 2.12. $P = (A^TA)^{-1}A^T$ is an orthogonal projection, assuming that A has full column rank.

Proof. Verify to yourself that it is symmetric and $P^2 = P$. Also verify that $\text{span}(P) = \text{span}(A)$. □

Corollary 2.13. The x aformentioned is orthogonal to $b - Ax$, the residual.

Proof.

$$b = Pb + P_{\perp}b$$

Substitute the definition of x and P found above

$$b = Ax + (b - Ax)$$

□

Proposition 2.14. Suppose we know that the columns of $Q \in \mathbb{R}^{m \times n}$ form an orthonormal basis for span(A). Then QQ^T is an orthogonal projector for A.

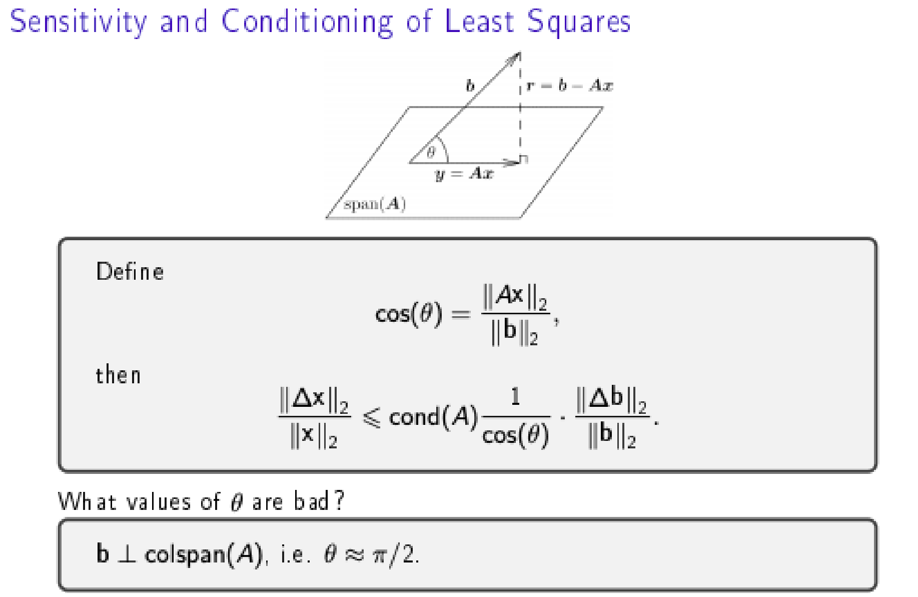
Proof. $(QQ^T)(QQ^T) = QQ^T$. Thus, this matrix is a projection; is it also clearly symmetric; finally, note that its span is precisely the span of A. □

Corollary 2.15. With Q as above and $P = QQ^T$, then the optimal x satisfying $Ax \approx b$ is $Ax = Pb$. Leftmulitply both sides by Q^T to obtain

$$Q^T Ax = Q^T b$$

If we do this, we can avoid the hassle of using the normal equations.

Definition 2.16. I take the following as definitions. No time to look into their proofs:



Sensitivity and Conditioning of Least Squares

Define

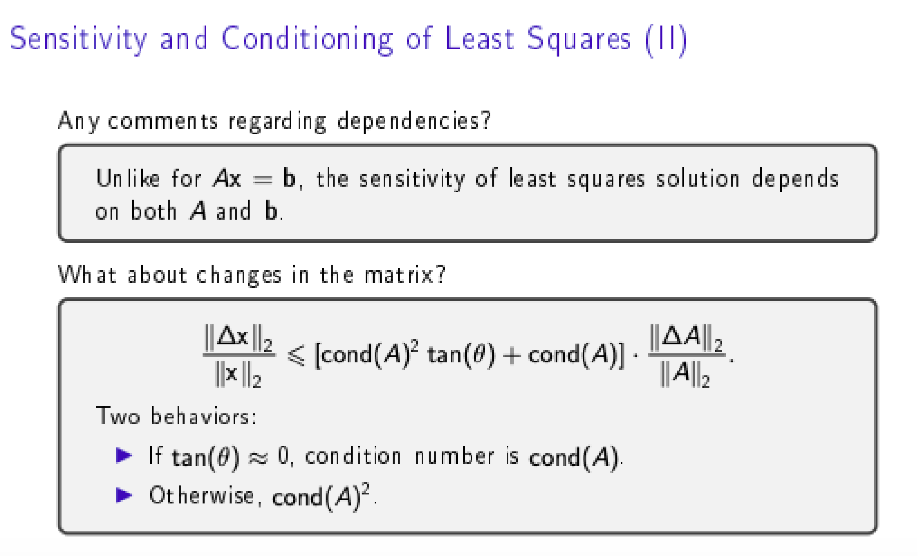
$$\cos(\theta) = \frac{\|Ax\|_2}{\|\mathbf{b}\|_2},$$

then

$$\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leqslant \operatorname{cond}(A)\frac{1}{\cos(\theta)} \cdot \frac{\|\Delta\mathbf{b}\|_2}{\|\mathbf{b}\|_2}.$$

What values of θ are bad?

$\mathbf{b} \perp \operatorname{colspan}(A)$, i.e. $\theta \approx \pi/2$.



Sensitivity and Conditioning of Least Squares (II)

Any comments regarding dependencies?

Unlike for $A\mathbf{x} = \mathbf{b}$, the sensitivity of least squares solution depends on both A and $\mathbf{b}$.

What about changes in the matrix?

$$\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leqslant [\operatorname{cond}(A)^2\tan(\theta) + \operatorname{cond}(A)] \cdot \frac{\|\Delta A\|_2}{\|A\|_2}.$$

Two behaviors:

- If $\tan(\theta) \approx 0$, condition number is $\operatorname{cond}(A)$.
- Otherwise, $\operatorname{cond}(A)^2$.

3 Lecture 8

Problem Transformations

3.1 Quiz

Problem 1. Polynomial Data Fitting If a first-degree polynomial x1+x2t is fit to the three data points (1,1), (2,1), (3,2), by linear least squares, what are the resulting values of the parameters x1 and x2?

Choice* x1=1, x2=0 x1=1/2, x2=1/2 x1=1/3, x2=1/2 x1=−1, x2=1

Solution 3.1. The solution is $(A^TA)^{-1}A^Tb$. Remember that the inverse of a general matrix is x^2.

Proposition 3.2. $k(A^TA) = k(A)^2$.

Proof.

- Recall that $\|A\| = \max_\sigma(A)$.
- $(A^TA)^T(A^TA) = (A^TA)^2$.
  - It follows that $\|A^TA\| = \|A\|^2$.
- Let $\Sigma(A)$ be the set of eigenvalues of A^TA. Strangely, it is difficult to argue that if $\lambda \in \Sigma(A)$, then $\frac{1}{\lambda} \in \Sigma(A^{-1})$. It can be argued, however, that $\lambda \in \Sigma(A^TA) \iff 1/\lambda \in \Sigma(A^TA)^{-1}$. Thus the smallest singular value of A^TA is, when reciprocated, the largest singular value of A^TA^{-1}.
  Not sure why this is significant. Crap

□

Remark 3.3. It seems that using A^TA is also disadvantageous insofar as if matrix looks like say

$$\begin{bmatrix} 1 & 0 \\ \epsilon & 0 \\ 1 & \epsilon \end{bmatrix}$$

and we compute A^TA then we will get a term that may be $1 + \epsilon^2$. Suppose that $\epsilon < \sqrt{\epsilon_{mach}}$. Then the term will end up being 1.

Remark 3.4. If Q is orthogonal, then $\|v\| = \|Qv\|$.

3.2 Householder Transformations

Remark 3.5. Recall that a matrix is orthogonal iff its transpose is its inverse. This means that both the rows and columns of the matrix, say Q, constitute an orthonormal set: the vectors are pairwise orthonormal, and they each have unit length.

Remark 3.6. Suppose that we obtain a system of the form:

$$\begin{bmatrix} R \\ 0 \end{bmatrix} x = Q^Tb := \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

Then we can solve for $Rx = c_1$ if R is invertible but not for c_2; as a consequence, the minimal norm that we can obtain when we ssolve this system is c_2^2.

Definition 3.7. A QR factorization expresses an $m \times n$ matrix A where $m \geq n$ as QR where Q is $m \times m$ and R is $m \times n$. This often better expressed as

$$\begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}$$

where Q_1 is $m \times n$, Q_2 is $m \times (m-n)$; R is $n \times n$ and 0 represents a $(m-n) \times n$ block. The reduced QR factorizaton is defined to be Q_1R.

Remark 3.8. A natural question raised is: what is the purpose of Q_2? Q_2 is the orthogonal complement of Q_1 – it is sometimes helpful to remember this. Since there are many ways to identify the orthogonal complement of Q_1, Q_2 is not unique. It can be shown, however, that $Q_1 R$ is unique if we force the diagonal entries of R to be positive

$Q_1 R$ is also unique up to multiplication of the diagonal entries of R by -1 and corresponding multiplication of a column in Q_1 by -1 (in particular, if $R_{i,j}$ is multiplied by -1, then $R_{i,j'}$ for $j' \geq j$ must be multiplied by -1 and column i of Q_1 must be multiplied by -1.

Remark 3.9. Observe that many methods to compute a QR factorization rely on forming Q multiplying successive orthogonal matrices like

$$Q_n Q_{n-1} \cdots Q_1$$

To obtain Q, we need to use all of the left factor. By this, I mean that if we compute $Q_2 Q_1$, then while we can drop columns of Q_1 past the nth column, we need to use all of Q_2 when performing the multiplication.

<++>

4 Lecture 9

Problem Transformations II – February 5th

Definition 4.1. Given a matrix v, a projection matrix onto span v is $\frac{vv^T}{v^Tv}$. This is unique, I think. In any case, this projection matrix is symmetric and satisfies $P^2 = P$, making it an orthogonal transformation.

Proposition 4.2. A householder transformation finds the normal vector v such that if α is reflected across the plane whose normal is given by v, then the resulting vector is nullified in all but the first k components. The projection matrix is given by

$$I - 2\frac{vv^T}{v^Tv}$$

Proof.

$$P = \frac{vv^T}{v^Tv}$$

is the projection matrix that projects onto span(v). We established that P is an orthogonal transformation; therefore, $(I - P)x$ will project onto the set $\{x | v^T x = 0\}$, which is the plane whose normal is v. Multiplication by $I - P$ amounts to travelling from x and then onto this plane; we now need to travel once more to attain a reflection. Thus, the projection that will allow us to attain the form is

$$I - 2\frac{vv^T}{v^Tv}$$

□

Proposition 4.3. It can be shown that the vector v which allows us to zero the last $m - k$ components of an m vector

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

where a_1 is $k - 1$, a_2 is $m - k + 1$ and $1 \leq k < m$ is

$$\begin{bmatrix} 0 \\ a_2 \end{bmatrix} - \alpha e_k$$

where $\alpha = -sign(a_k)\|a_2\|_2$.

The choice in sign for α reflects our desire to avoid cancellation.

Definition 4.4. The matrix that rotates a vector θ degrees counter clockwise in the $\mathbb{R}^2$ plane is

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

Note that since sin is an odd function ($f(-x) = -f(x)$) and cos and even function ($f(-x) = f(x)$), it follows that the matrix rotating an angle θ degrees clockwise is

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

Remark 4.5. We are interested in finding the angle θ and hence the value of $\cos(\theta)$ and $\sin(\theta)$ that solves the problem:

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sqrt{a_1^2 + a_2^2} \\ 0 \end{bmatrix}$$

We will call the resulting matrix M that performs the transformation.

Note that the answer to this question is

$$c = \frac{a_1}{\sqrt{a_1^2 + a_2^2}}, s = \frac{a_2}{\sqrt{a_1^2 + a_2^2}}$$

If we had instead solved (note the difference in the matrix M).

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sqrt{a_1^2 + a_2^2} \\ 0 \end{bmatrix}$$

the resulting values of c and s would change, but that is okay – so long as we redefine M to be this matrix with the solved values for c and s.

Remark 4.6. Suppose that we wish to rotate the b entry of some column onto the a entry of the same column. Suppose that we have solved for the matrix M that accomplishes this transformation where the b entry is a_2 and the a entry is a_1. Then we can construct a givens transformation:

A givens rotation imbeds M into a matrix A with ones (initially) along the diagonal so that $A_{a,a} = M_{0,0}$, $A_{a,b} = M_{0,1}$, $A_{b,a} = M_{1,0}$, $A_{b,b} = M_{1,1}$.

5 Lecture 10

SVD

Definition 5.1. Every matrix admits a decomposition:

$$A = U\Sigma V^T$$

The reduced SVD is the same decompositon but Σ is reduced to be the small as possible.

Entries of U are called left singular vectors, entries in Σ are singular values. Entries in V^T can

Theorem 5.2. If $\|A\| = \|\Sigma\| = \sigma_1$ where σ_1 is the largest diagonal entry in Σ.

Remark 5.3. If a singular value appearing in Σ is negative, can it be made positive?

Solution 5.4. Yes, flip an appropriate singular vector in sign. UNRESOLVED.

Proposition 5.5. Take it for granted that $\|A\|_2 = \sigma_1$. Given this, $k(A) = \frac{\sigma_1}{\sigma_{\min m,n}}$

Proof. Recall that we have redefined $k(A)$ to now be

$$\|A\|\|A^+\|$$

We have also found that $A^+ = V\Sigma^+U^T$, from which it follows that $\|A^+\|$ is the greatest diagonal entry of Σ^+ which is the reciprocal of the smallest diagonal entry in Σ. □

Proposition 5.6. The null space of V^T is given by the rows of V^T corresponding to the singular values of A (ie the values in Σ) that are 0.

Proof. Let these rows be collected in the set V. We argue that $V \subseteq \mathcal{N}(A)$. Just hit A by each $v_i \in V$.

Recall that $A = U\Sigma V^T$ which is really (assuming that A is an $m \times n$ matrix)

$$[u_1 \dots u_n] \begin{bmatrix} \sigma_1 & \dots & \\ & \sigma_2 \dots & \\ & & \sigma_3 \dots \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}$$

from which it follows that

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

Recall that V_i form an orthogonal basis for $\mathbb{R}^n$. Observe that Av_i where $\sigma_i = 0$ results in 0. On the other hand, Av_i where $\sigma_i \neq 0$ is non-zero. Thus, we have determined which of a basis' vectors result in annihilation. This has determined the null space. □

Proposition 5.7. The rank of A is given by the number of singular values that are not zero.

Proof.

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$$

tells us that to obtain any vector in the column space of A we only need to have access to the u_i such that $\sigma_i \neq 0$. However many u_i exist is the rank of the matrix. □

Remark 5.8. Rank is not robust to rounding error. Suppose we have a rank one matrix; then introduce rounding error (each entry in the matrix has ϵ_{mach} added or subtracted from it. By doing this, it is conceivable that the rank of the matrix is changed significantly. Far better, as a consequence, is to compute the numerical rank, which asks how many singular values fall above a tolerance. That is, for $\sigma \in \Sigma$, determine whether $|\sigma| > \epsilon$.

Theorem 5.9. Eckhart Young Mirsky The best k rank approximation to A is given by

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Remark 5.10. Convince yourself that the summation above is the matrix that one obtains by, alternatively, zeroing all but the first k diagonal entries of Σ and then multiplying out $U\Sigma V^T$.

Proposition 5.11. In the circumstance that A^TA is invertible, the inverse $(A^TA)^{-1}A^{-1}$ agrees with the inverse obtained from the SVD $V\Sigma^+U^T$.

Proof.

$$\begin{aligned} A &= U\Sigma V^T \\ \implies (A^TA)^{-1}A^{-1} &= (V\Sigma U^TU\Sigma V^T)^{-1}V\Sigma^+U \\ &= V\Sigma^+U \end{aligned}$$

□

Example 5.12. To compute the pseudoinverse of a matrix like

$$\begin{bmatrix} a & & \\ & b & \\ & & c \\ 0 & 0 & 0 \end{bmatrix}$$

the foregoing establishes that we need not compute the SVD and then obtain the pseudoinverse by re-arranging the resultant factors. We can compute the SVD by also computing $(^{-1}A^TA)A^T$ which, in this case, gives us

$$\begin{bmatrix} 1/a & & \\ & 1/b & \\ & & 1/c \\ 0 & 0 & 0 \end{bmatrix}$$

Remark 5.13. Using a k rank approximation is not unambiguously good, for computation of the SVD requires $O(n^3)$ time.

Unresolved 5.14. If A_k is the best k rank approximation to A, is it the case that

$$\|A_k - A\|$$

is the norm of the matrix obtained by zeroing out the first k singular values of A.

Theorem 5.15.

Proof.

$$\begin{aligned} U\Sigma V^T x &\cong b \\ \Sigma(V^T x) &\cong U^T b \\ \Sigma y &\cong U^T b \end{aligned}$$

Solve for y

$$\begin{aligned} y_i &= (U^T b)_i/\sigma_i \text{ for } i \in [k] \\ y_i &= 0 \text{ ow} \end{aligned}$$

Then solve for x where $V^T x = y$. Note that y is the optimal of all vectors $\hat{y}$ that solve $\Sigma\hat{y} \cong U^T b$. It follows that x is the minimal of all vectors that solve $Ax \cong b$ since $x = Vy$, meaning that $\|x\|_2 = \|y\|_2$. □

Definition 5.16. The solution to the total least squares problem

is $V\Sigma^+ U^T b$ where Σ^+ is computed by reciprocating singular values in their spots (save those singular values that are 0.

Remark 5.17. Remember the cost of householder for nonsquare and perhaps the cost for all $n \times n$ matrix.