#### **Table of Contents**

About assignments

#### **System of Linear Equations**

Schetch of solving the linear equation

Solving tridiagonal linear equation

Algorithm 2.1: Forward-Substitution for Lower Triangular System

LU Factorization

The Gaussian elimination process

The elementary elimination matrix

Properties of elimination matrices

Coding: Properties of elimination matrices

Algorithm 2.3: LU Factorization by Gaussian Elimination

Complexity Analysis

Issue: how to handle small diagonal entries?

Pivoting technique

Gaussian Elimination process with Partial Pivoting

Algorithm 2.4 LU Factoriazation by Gaussian Elimination with Partial Pivoting

Complete pivoting

#### **Sensitivity Analysis**

Issue: An III Conditioned Matrix

Forward Error and Backward Error

Absolute Error and Relative Error

Coding: Floating point numbers have "constant" relative error

(Relative) Condition Number

Quantify Error of a Scalar Operator

Why we should avoid using floating point numbers being too big/small?

Measuring the size of a vector: Norms

Measure the size of a matrix

Coding: Vector and Matrix Norms

Condition Number of a Linear Operator

Coding: Numeric experiment on condition number

#### **Computing Matrix Inverse**

Gauss Jordan Elimination Matrix

Computing the inverse

#### **Special Matrices**

Positive definite symmetric matrix

Cholesky decomposition

Algorithm 2.7 Cholesky Factorization

Banded matrices

Sparse matrices and linear operators

Rank 1 update: Sherman-Morrison formula

#### **Assignments**

present

## **About assignments**

- 1. You should submit your homework through Github pull request. Xuanzhao Gao will comment on your PR, then you should resolve the issues to get your PR merged. You should submit different PRs for different assignments. <a href="Open repo">Open repo</a>
- 2. We will grade based on your merged PRs. You will not be failed if your submition is complete.
- 3. You may check the answers of other students, but you must credit him in your PR description, e.g.

### Reference:

- \* PR #3
- \* PR #4

## System of Linear Equations

Solving:  $\mathbf{A}\mathbf{x} = \mathbf{b}$ 

Quiz: In a cage, chickens and rabbits add up to 35 heads and 94 feet. Please count the number of chickens and rabbits.

#### Table of contents

- Gaussian elimination algorithm
- Pivoting technique
- Sensitivity analysis and condition number
- Getting matrix inverse with Gauss-Jordan elimination
- Solving linear equations for special matrices (optional)

## Schetch of solving the linear equation

$$A = LU$$
$$x = A^{-1}b = U^{-1}L^{-1}b$$

## Solving tridiagonal linear equation

$$Lx = b$$

$$L = egin{pmatrix} l_{11} & 0 & 0 \ l_{21} & l_{22} & 0 \ l_{31} & l_{32} & l_{33} \end{pmatrix}$$

# Algorithm 2.1: Forward-Substitution for Lower Triangular System

$$x_1 = b_1/l_{11}, \;\;\; x_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij} x_j
ight)/l_{ii}, \;\; i=2,\dots,n$$

@assert size(l) == (n, n) "size mismatch"

1 function back\_substitution!(l::AbstractMatrix, b::AbstractVector)

```
4
       x = zero(b)
 5
       # loop over columns
 6
       for j = 1:n
           # stop if matrix is singular
 8
           if iszero(l[j, j])
                error("The lower triangular matrix is singular!")
 9
10
           end
11
           # compute solution component
           x[j] = b[j] / l[j, j]
13
           for i = j+1:n
                # update right hand side
14
15
               b[i] = b[i] - l[i, j] * x[j]
16
           end
17
       end
18
       return x
19 end
l = 4×4 Matrix{Float64}:
      0.547284
                            0.0
                                        0.0
                 0.0
                 0.759989
     -1.068
                            0.0
                                        0.0
     -0.197166 -0.345128
                           -0.81142
                                        0.0
      0.735737 -0.971137
                           -0.300388 -0.0411604
 1 l = tril(randn(4, 4))
b = [-0.930708, -0.623271, -0.471425, -0.917524]
 1 b = randn(4)
 [-1.7006, -3.20992, 2.35952, 50.4085]
 1 back_substitution!(l, copy(b))
```

The Julia's version

2

3

n = length(b)

```
1 using LinearAlgebra
[-1.7006, -3.20992, 2.35952, 50.4085]
1 LowerTriangular(l) \ b
```

#### LU Factorization

### The Gaussian elimination process

$$A = egin{pmatrix} a_{11} & a_{12} & a_{13} \ a_{21} & a_{22} & a_{23} \ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

## The elementary elimination matrix

$$M_1A=egin{pmatrix} a_{11}&a_{12}&\dots\ 0&a'_{22}&\dots\ dots&dots&\ddots \end{pmatrix}$$

$$M_k = egin{pmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \ dots & \ddots & dots & dots & \ddots & dots \ 0 & \dots & 1 & 0 & 0 & \dots & 0 \ 0 & \dots & 0 & 1 & 0 & \dots & 0 \ 0 & \dots & 0 & -m_{k+1} & 1 & \dots & 0 \ dots & \ddots & dots & dots & \ddots & dots \ 0 & \dots & 0 & -m_n & 0 & \dots & 1 \end{pmatrix}$$

where  $m_i = a_i/a_k$ 

Then

$$L = M_1^{-1} M_2^{-1} \dots M_{n-1}^{-1}$$

## Properties of elimination matrices

$$M_k^{-1} = 2I - M_k$$
  $M_k M_{k'>k} = M_k + M_{k'} - I$ 

### Coding: Properties of elimination matrices

```
A3 = 3 \times 3  Matrix{Int64}:
      4 4 2
      4 6 4
 1 \quad A3 = [1 \ 2 \ 2; \ 4 \ 4 \ 2; \ 4 \ 6 \ 4]
elementary_elimination_matrix (generic function with 1 method)
 1 function elementary_elimination_matrix(A::AbstractMatrix{T}, k::Int) where T
       n = size(A, 1)
 3
       Qassert size(A, 2) == n
       # create Elementary Elimination Matrices
       M = Matrix{Float64}(I, n, n)
       for i=k+1:n
            M[i, k] = -A[i, k] ./ A[k, k]
 8
       end
 9
       return M
10 end
The elimination
3×3 Matrix{Float64}:
 1.0 0.0 0.0
 -4.0 1.0 0.0
 -4.0 0.0 1.0
 1 elementary_elimination_matrix(A3, 1)
3×3 Matrix{Float64}:
1.0
     2.0
           2.0
0.0 - 4.0 - 6.0
0.0 - 2.0 - 4.0
 1 elementary_elimination_matrix(A3, 1) * A3
The inverse
3×3 Matrix{Float64}:
1.0 0.0 0.0
4.0 1.0 0.0
4.0 0.0 1.0
 1 inv(elementary_elimination_matrix(A3, 1))
The multiplication
3×3 Matrix{Float64}:
1.0 0.0 0.0
0.0 1.0 0.0
0.0 -1.5 1.0
```

1 elementary\_elimination\_matrix(A3, 2)

```
3×3 Matrix{Float64}:
1.0 0.0 0.0
4.0 1.0 0.0
4.0 1.5 1.0

1 inv(elementary_elimination_matrix(A3, 1)) * inv(elementary_elimination_matrix(A3, 2))
```

# **Algorithm 2.3: LU Factorization by Gaussian Elimination**

```
lufact_naive! (generic function with 1 method)
```

```
(3×3 Matrix{Float64}:, 3×3 Matrix{Int64}:)
1.0 0.0 0.0 1 2 2

1 lufact_naive!(copy(A3))
```

```
1 using Test
```

#### Test Passed

Better implementation

```
lufact! (generic function with 1 method)
 1 function lufact!(a::AbstractMatrix)
 2
        n = size(a, 1)
 3
        @assert size(a, 2) == n "size mismatch"
 4
        m = zero(a)
 5
        m[1:n+1:end] .+= 1
 6
        # loop over columns
        for k=1:n-1
 7
 8
            # stop if pivot is zero
 9
            if iszero(a[k, k])
10
                error("Gaussian elimination fails!")
11
            end
12
            # compute multipliers for current column
13
            for i=k+1:n
14
                m[i, k] = a[i, k] / a[k, k]
15
            end
16
            # apply transformation to remaining sub-matrix
17
            for j=k+1:n
                for i=k+1:n
18
                    a[i,j] -= m[i,k] * a[k, j]
19
20
                end
21
            end
22
        end
23
        return m, triu!(a)
24 end
lufact (generic function with 1 method)
 1 lufact(a::AbstractMatrix) = lufact!(copy(a))
 DefaultTestSet("LU", [], 3, false, false, true, 1.677652817066939e9, 1.677652817066971e9)
 1 @testset "LU" begin
 2
       a = randn(4, 4)
 3
       L, U = lufact(a)
 4
       @test istril(L)
 5
       @test istriu(U)
        @test L * U ≈ a
 6
 7 end
                     Pass
    Test Summary:
                           Total
                                                                                      ②
                                  0.0s
A4 = 4×4 Matrix{Float64}:
      -1.13722
                  0.32243
                              1.17589
                                        0.309568
                  0.502538
      -0.499461
                             -0.212333
                                        0.325516
       0.291973
                  0.860287
                             -1.58332
                                        0.191323
       0.79411
                 -0.58919
                              0.293682
                                        0.449737
 1 A4 = randn(4, 4)
  (4×4 Matrix{Float64}:
                                       , 4×4 Matrix{Float64}:
                         0.0
                                    0.0 -1.13722 0.32243
                                                                           0.309568
    1.0
                0.0
                                                                1.17589
```

1 lufact(A4)

```
julia_lures = LU{Float64, Matrix{Float64}, Vector{Int64}}
              L factor:
              4×4 Matrix{Float64}:
                                    0.0
                                              0.0
                1.0
                           0.0
                0.439196
                           1.0
                                    0.0
                                              0.0
               -0.256744
                           2.6129
                                    1.0
                                              0.0
               -0.698293 -1.00862 0.609723 1.0
              U factor:
              4×4 Matrix{Float64}:
               -1.13722 0.32243
                                    1.17589
                                               0.309568
                         0.360928 -0.728778
                0.0
                                               0.189556
                0.0
                         0.0
                                    0.622801
                                              -0.224488
                0.0
                         0.0
                                               0.993971
                                    0.0
 1 julia_lures = lu(A4, NoPivot()) # the version we implemented above has no pivot
4×4 Matrix{Float64}:
                                 0.309568
 -1.13722 0.32243
                      1.17589
 0.0
           0.360928
                     -0.728778
                                 0.189556
 0.0
           0.0
                      0.622801
                               -0.224488
 0.0
           0.0
                                 0.993971
 1 julia_lures.U
LU{Float64, Matrix{Float64}, Vector{Int64}}
 1 typeof(julia_lures)
 (:factors, :ipiv, :info)
 1 fieldnames(julia_lures |> typeof)
```

## **Complexity Analysis**

 $O(n^3)$ 

## Issue: how to handle small diagonal entries?

```
(2×2 Matrix{Float64}:, 2×2 Matrix{Float64}:)
    1.0     0.0     1.0     1.0

1     lufact(small_diagonal_matrix[end:-1:1, :])
```

## Pivoting technique

$$PA = LU$$

```
2×2 Matrix{Float64}:
1.0    1.0
1.0e-20    1.0
1 [10^(e) 1; 1 1][end:-1:1, :]
```

# Gaussian Elimination process with Partial Pivoting

$$A = egin{pmatrix} a_{11} & a_{12} & a_{13} \ a_{21} & a_{22} & a_{23} \ a_{31} & a_{32} & a_{33} \end{pmatrix} \ \dots M_2 P_2 M_1 P_1 A = U$$

NOTE:  $P_{k+1}$  and  $M_k$  commute

# Algorithm 2.4 LU Factoriazation by Gaussian Elimination with Partial Pivoting

```
1 function lufact_pivot!(a::AbstractMatrix)
 2
       n = size(a, 1)
 3
       @assert size(a, 2) == n "size mismatch"
 4
       m = zero(a)
       P = collect(1:n)
 5
 6
       # loop over columns
 7
       @inbounds for k=1:n-1
 8
           # search for pivot in current column
 9
           val, p = findmax(x->abs(a[x, k]), k:n)
           p += k-1
10
            # find index p such that |a_{pk}| \ge |a_{ik}| for k \le i \le n
11
12
           if p != k
13
                # swap rows k and p of matrix A
                for col = 1:n
14
                    a[k, col], a[p, col] = a[p, col], a[k, col]
15
16
                end
                \# swap rows k and p of matrix M
17
18
               for col = 1:k-1
                    m[k, col], m[p, col] = m[p, col], m[k, col]
19
20
                end
21
               P[k], P[p] = P[p], P[k]
22
           end
           if iszero(a[k, k])
23
24
                # skip current column if it's already zero
25
                continue
26
           end
27
           # compute multipliers for current column
28
           m[k, k] = 1
29
           for i=k+1:n
               m[i, k] = a[i, k] / a[k, k]
30
31
           end
32
           # apply transformation to remaining sub-matrix
33
           for j=k+1:n
34
               akj = a[k, j]
35
                for i=k+1:n
36
                    a[i,j] = m[i,k] * akj
37
                end
           end
38
39
       end
40
       m[n, n] = 1
41
       return m, triu!(a), P
42 end
```

1 using BenchmarkTools

```
1 @bind benchmark_lu CheckBox()
```

```
1 if benchmark_lu let
2    n = 200
3    A = randn(n, n)
4    Qbenchmark lufact_pivot!($A)
5 end end
```

```
if benchmark_lu let
n = 200
A = randn(n, n)
depenchmark lu($A)
end end
```

## Complete pivoting

PAQ = LU

# **Sensitivity Analysis**

### Issue: An Ill Conditioned Matrix

$$A = egin{pmatrix} 0.913 & 0.659 \ 0.457 & 0.330 \end{pmatrix}$$

#### Forward Error and Backward Error

Forward error:

$$\operatorname{dist}(f(x),\hat{f}(x))$$

Backward error

$$\operatorname{dist}(x, f^{-1}(\hat{f}(x)))$$

### **Absolute Error and Relative Error**

Absolute error:  $\| oldsymbol{x} - \hat{oldsymbol{x}} \|$ 

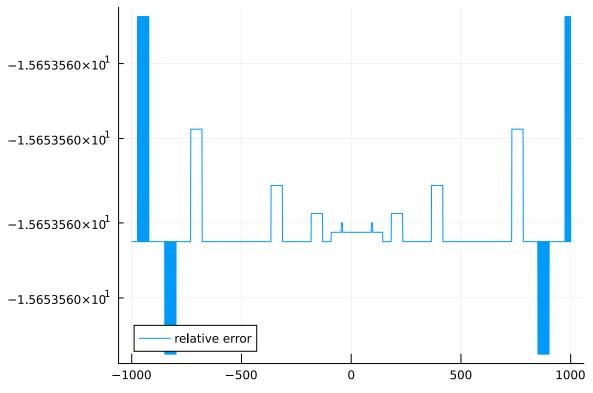
Relative error:  $\frac{\|x - \hat{x}\|}{\|x\|}$ 

where  $\|\cdot\|$  is a measure of size.

# Coding: Floating point numbers have "constant" relative error

2.220446049250313e-16

1 eps(Float64)



GKS: Possible loss of precision in routine SET\_WINDOW ①

```
2.220446049250313e-16
```

```
1 eps(1.0)/1.0
```

#### 2.220446049250313e-16

```
1 eps(2.0)/2.0
```

#### 1.570092458683775e-16

```
1 eps(sqrt(2))/sqrt(2)
```

## (Relative) Condition Number

$$\lim_{arepsilon o 0^+} \sup_{\|\delta x\| \ \le arepsilon} rac{\|\delta f(x)\|/\|f(x)\|}{\|\delta x\|/\|x\|}$$

## **Quantify Error of a Scalar Operator**

$$y = \exp(x)$$

# Why we should avoid using floating point numbers being too big/small?

$$a+b$$

## Measuring the size of a vector: Norms

$$\|v\|_p = \left(\sum_i |v_i|^p
ight)^{1/p}$$

#### Measure the size of a matrix

$$\|A\|=\max_{x\neq 0}\frac{\|Ax\|}{\|x\|}$$

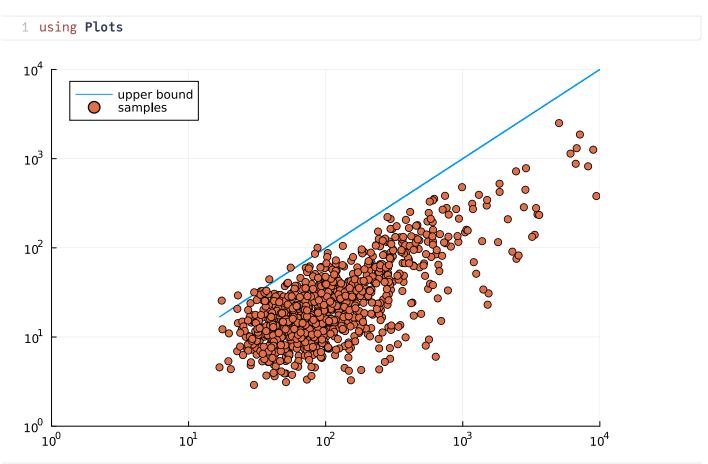
## **Coding: Vector and Matrix Norms**

```
7.0
 1 norm([3, 4], 1)
4.0
 1 norm([3, 4], Inf)
2.0
 1 # 10 norm is not a true norm
 2 norm([3, 4], 0)
1.0
 1 norm([3, 0], 0)
mat = 2×2 Matrix{Float64}:
       -0.284666
                   -1.24104
       -0.0321259
                    2.3657
 1 mat = randn(2, 2)
3.6067408469604665
 1 opnorm(mat, 1)
2.3978261092617075
 1 opnorm(mat, Inf)
2.673499761890822
 1 opnorm(mat, 2)
ArgumentError: invalid p-norm p=0. Valid: 1, 2, Inf
  1. opnorm(::Matrix{Float64}, ::Int64) @ generic.jl:740
  2. top-level scope @ Local: 1 [inlined]
 1 opnorm(mat, 0)
10.020429724445629
 1 cond(mat)
```

## **Condition Number of a Linear Operator**

 ${\rm cond}(A) = \|A\| \|A^{-1}\|$ 

# Coding: Numeric experiment on condition number



```
1 let
                                         n = 1000
                                         p = 2
                                          errors = zeros(n)
                                         conds = zeros(n)
     6
                                         for k = 1:n
                                                                 A = rand(10, 10)
                                                                 b = rand(10)
     9
                                                                 dx = A \setminus b
                                                                 sx = Float32.(A) \setminus Float32.(b)
 10
                                                                 errors[k] = (norm(sx - dx, p)/norm(dx, p)) / (norm(b-Float32.(b), p)/norm(b, p)/norm(b, p)) / (norm(b-Float32.(b), p)/norm(b, p)/norm
11
                                                                 p))
                                                                 conds[k] = cond(A, p)
12
13
                                          end
                                          plt = plot(conds, conds; label="upper bound", xlim=(1, 10000), ylim=(1, 10000),
14
                                          xscale=:log10, yscale=:log10)
                                          scatter!(plt, conds, errors; label="samples")
15
16 end
 17
```

## **Computing Matrix Inverse**

### Gauss Jordan Elimination Matrix

$$N_k = egin{pmatrix} 1 & \dots & 0 & -m_1 & 0 & \dots & 0 \ dots & \ddots & dots & dots & \ddots & dots \ 0 & \dots & 1 & -m_{k-1} & 0 & \dots & 0 \ 0 & \dots & 0 & 1 & 0 & \dots & 0 \ 0 & \dots & 0 & -m_{k+1} & 1 & \dots & 0 \ dots & \ddots & dots & dots & \ddots & dots \ 0 & \dots & 0 & -m_n & 0 & \dots & 1 \end{pmatrix}$$

where  $m_i = a_i/a_k$ 

## Computing the inverse

$$SN_nN_{n-1}\dots N_1A=I$$

Then

$$A^{-1} = SN_nN_{n-1}\dots N_1$$

## **Special Matrices**

## Positive definite symmetric matrix

- (Real) Symmetric:  $A = A^T$ ,
- Positive definite:  $x^T A x > 0$  for all  $x \neq 0$ .

### Cholesky decomposition

$$A = LL^T$$

### **Algorithm 2.7 Cholesky Factorization**

```
chol! (generic function with 1 method)
 1 function chol!(a::AbstractMatrix)
        n = size(a, 1)
        Qassert size(a, 2) == n
       for k=1:n
            a[k, k] = sqrt(a[k, k])
            for i=k+1:n
                a[i, k] = a[i, k] / a[k, k]
 8
            for j=k+1:n
10
                for i=k+1:n
11
                    a[i,j] = a[i,j] - a[i, k] * a[j, k]
12
13
            end
14
        end
15
        return a
16 end
10×10 Matrix{Float64}:
 -5.55112e-17 -2.77556e-17
                             -2.77556e-16 ...
                                                1.38778e-17
                                                             -2.77556e-17
                               2.77556e-17
  0.0
               -1.11022e-16
                                                5.89806e-17
                                                              -1.11022e-16
  0.0
                0.0
                               0.0
                                               -1.11022e-16
                                                              -6.93889e-17
                               2.77556e-17
  0.0
                0.0
                                                5.55112e-17
                                                              -6.93889e-18
                0.0
  0.0
                               0.0
                                               -1.63389e-16
                                                              4.16334e-17
                1.6263e-19
                               0.0
                                               -2.08167e-17
                                                              1.71738e-16
  0.0
  0.0
               -2.77556e-17
                               1.38778e-17
                                               -5.55112e-17
                                                               1.249e-16
  0.0
                0.0
                              -6.93889e-18
                                                0.0
                                                              -8.32667e-17
                               2.77556e-17
                                                5.55112e-17
                                                               2.77556e-17
  0.0
                0.0
 -1.38778e-17
                                                               0.0
                0.0
                               0.0
                                                0.0
 1 let
 2
        n = 10
       Q, R = qr(randn(10, 10))
        a = Q * Diagonal(rand(10)) * Q'
       L = chol!(copy(a))
        tril(L) * tril(L)' - a
        # cholesky(a) in Julia
```

#### **Banded matrices**

8 end

## Sparse matrices and linear operators

Accessing every element is not allowed, but matrix-vector multiplication is defined.

You need iterative solvers like GMRES (Package: IterativeSolvers).

## Rank 1 update: Sherman-Morrison formula

$$(A - uv^T)^{-1} = A^{-1} + A^{-1}u(1 - v^TA^{-1}u)^{-1}v^TA^{-1}$$

which requires only  $O(n^2)$  extra work.

## **Assignments**

- 1. Get the relative condition number of division operation a/b.
- 2. Classify each of the following matrices as well-conditioned or ill-conditioned:

(a). 
$$\begin{pmatrix} 10^{10} & 0 \\ 0 & 10^{-10} \end{pmatrix}$$
(b). 
$$\begin{pmatrix} 10^{10} & 0 \\ 0 & 10^{10} \end{pmatrix}$$
(c). 
$$\begin{pmatrix} 10^{-10} & 0 \\ 0 & 10^{-10} \end{pmatrix}$$
(d). 
$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

3. Implement the Gauss-Jordan elimination algorithm to compute matrix inverse. In the following example, we first create an augmented matrix (A|I). Then we apply the Gauss-Jordan elimination matrices on the left. The final result is stored in the augmented matrix as  $(I, A^{-1})$ .

$$\begin{bmatrix} 1 & 0 & 0 \\ -4 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 4 & 4 & 2 & 0 & 1 & 0 \\ 4 & 6 & 4 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -4 & -6 & -4 & 1 & 0 \\ 0 & -2 & -4 & -4 & 0 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & -4 & -6 & -4 & 1 & 0 \\ 0 & -2 & -4 & -4 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 & -1 & 0.5 & 0 \\ 0 & -4 & -6 & -4 & 1 & 0 \\ 0 & 0 & -1 & -2 & -0.5 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 & -1 & 0.5 & 0 \\ 0 & -4 & -6 & -4 & 1 & 0 \\ 0 & 0 & -1 & -2 & -0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & -1 \\ 0 & -4 & 0 & 8 & 4 & -6 \\ 0 & 0 & -1 & -2 & -0.5 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.25 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & -1 \\ 0 & -4 & 0 & 8 & 4 & -6 \\ 0 & 0 & -1 & -2 & -0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & -1 \\ 0 & 1 & 0 & -2 & -1 & 1.5 \\ 0 & 0 & 1 & 2 & 0.5 & -1 \end{bmatrix},$$

Task: Please implement a function gauss\_jordan that computes the inverse for a matrix at any size. Please also include the following test in your submission.

```
@testset "Gauss Jordan" begin
    n = 10
    A = randn(n, n)
```

```
@test gauss_jordan(A) * A \approx Matrix{Float64}(I, n, n)
end
```