CS 450: Numerical Analysis – Lecture 3

Vector and Matrix Norms, SVD, and Matrix Conditioning

Edgar Solomonik

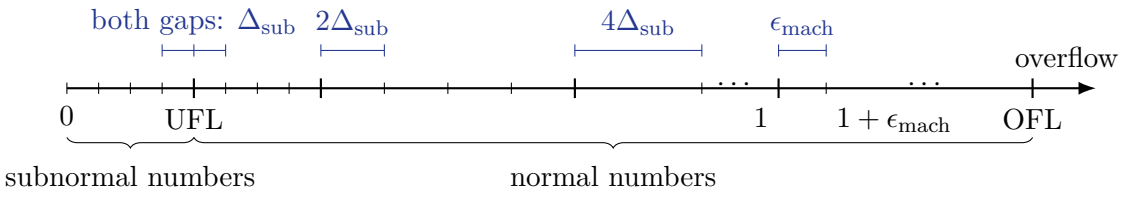
September 1, 2026

Today

- Brief review of the floating-point number line
- Vector and matrix norms
- Conditioning of a linear map
- Singular value decomposition and the matrix condition number

Brief floating-point review

For positive binary floating-point numbers, subnormal values are equally spaced. Let Δ_{sub} denote this spacing. The smallest positive normal number is the underflow level UFL. Normalized values begin with the same spacing Δ_{sub}; the gap doubles whenever the exponent increases by one. The overflow level OFL is the largest finite positive number. Negative values occur symmetrically.



With rounding to nearest, this increasing absolute spacing gives a uniform relative-error bound in the normal range. If a nonzero exact result z is rounded to a normal value and no overflow occurs, then

$$\frac{|\text{fl}(z)-z|}{|z|} \leq u, \qquad \text{equivalently} \qquad \text{fl}(z) = z(1+\delta), \quad |\delta| \leq u.$$

In the subnormal range, no bound of this form is uniform. The corresponding uniform statement is instead the absolute-error bound

$$|\text{fl}(z)-z| \leq \frac{1}{2}\Delta_{\text{sub}}.$$

Vector norms

A *vector norm* is a function

$$\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}$$

such that, for all $x, y \in \mathbb{R}^n$ and every scalar $\alpha \in \mathbb{R}$,

$$\begin{aligned} &\|x\| \geq 0, && \|x\| = 0 \iff x = 0, \\ &\|\alpha x\| = |\alpha| \, \|x\|, && \|x+y\| \leq \|x\| + \|y\|. \end{aligned}$$

The last property is the *triangle inequality*.

p-norms. For $1 \leq p < \infty$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

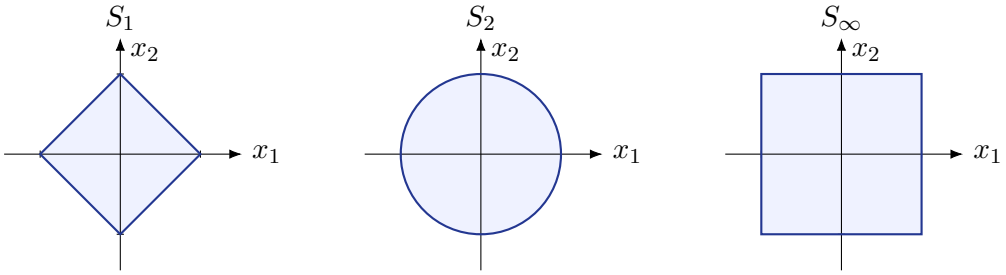
The most common cases are

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}, \qquad \|x\|_\infty = \max_i |x_i|.$$

Unit p-spheres and p-balls. For the p-norm, write

$$S_p = \{x \in \mathbb{R}^n : \|x\|_p = 1\}$$

for the unit p-sphere. In the plots below, the dark outline is S_p and the shaded points satisfy $\|x\|_p < 1$.



The unit p-sphere together with the open unit p-ball is the closed unit p-ball,

$$S_p \cup \{x : \|x\|_p < 1\} = \{x : \|x\|_p \le 1\}.$$

This set is convex. If $\|x\|_p \le 1$, $\|y\|_p \le 1$, and $0 \le t \le 1$, then

$$\|tx + (1-t)y\|_p \le t\,\|x\|_p + (1-t)\,\|y\|_p \le 1.$$

The open unit p-ball is also convex: if the two initial inequalities are strict, so is the final one. The sphere S_p alone is generally not convex: if $x \in S_p$, then $-x \in S_p$, but their midpoint 0 lies in the open ball.

Matrix norms

A norm on the vector space $\mathbb{R}^{m\times n}$ satisfies the same four norm properties. One example is the *Frobenius norm*,

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2} = \|\mathrm{vec}(A)\|_2\,,$$

where $\mathrm{vec}(A)$ stacks the columns of A into one vector.

Induced or operator norms. A vector p-norm induces the matrix norm

$$\|A\|_p = \max_{x\in S_p} \|Ax\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \tag{3.1}$$

The maximum is attained because S_p is compact and $x \mapsto \|Ax\|_p$ is continuous. Thus, for every A, there is a vector $x_\star \in S_p$ such that

$$\|Ax_\star\|_p = \|A\|_p\,.$$

For every x and all conforming matrices A, B,

$$\|Ax\|_p \le \|A\|_p\,\|x\|_p\,, \qquad \|AB\|_p \le \|A\|_p\,\|B\|_p\,.$$

Two induced norms can be read directly from the entries:

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \quad \text{(maximum column sum)}, \qquad \|A\|_\infty = \max_i \sum_j |a_{ij}| \quad \text{(maximum row sum)}.$$

The Frobenius norm is generally different from the induced 2-norm.

Conditioning of a vector-valued function

Let

$$F : \mathbb{R}^n \longrightarrow \mathbb{R}^m.$$

At an input $x \neq 0$ with $F(x) \neq 0$, the local relative condition number in the 2-norm is the worst first-order amplification of a relative input perturbation:

$$\kappa_{\mathrm{rel}}(F, x) = \lim_{\epsilon \to 0^+} \sup_{0 < \|\delta x\|_2 \leq \epsilon \|x\|_2} \frac{\|F(x + \delta x) - F(x)\|_2 \,/\, \|F(x)\|_2}{\|\delta x\|_2 \,/\, \|x\|_2}. \tag{3.2}$$

The Jacobian. The Jacobian is the matrix of first derivatives,

$$J_F(x) = \begin{bmatrix} \dfrac{\partial F_1}{\partial x_1}(x) & \cdots & \dfrac{\partial F_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial F_m}{\partial x_1}(x) & \cdots & \dfrac{\partial F_m}{\partial x_n}(x) \end{bmatrix}.$$

If F is differentiable, then

$$F(x + \delta x) - F(x) = J_F(x)\delta x + o(\|\delta x\|_2).$$

The corresponding local absolute condition number is $\|J_F(x)\|_2$. Scaling input and output errors relatively, the definition in (3.2) gives

$$\kappa_{\mathrm{rel}}(F, x) = \|J_F(x)\|_2 \, \frac{\|x\|_2}{\|F(x)\|_2}. \tag{3.3}$$

This is the vector analogue of $|f'(x)|\,|x| \,/\, |f(x)|$ for a scalar function.

Specialization to a linear map. Let $A \in \mathbb{R}^{m \times n}$ and $F(x) = Ax$. Since

$$F_i(x) = \sum_{j=1}^{n} a_{ij} x_j, \qquad \frac{\partial F_i}{\partial x_j} = a_{ij},$$

we have $J_F(x) = A$ for every x. Therefore,

$$\kappa_{\mathrm{rel}}(F, x) = \|A\|_2 \, \frac{\|x\|_2}{\|Ax\|_2}. \tag{3.4}$$

For a linear map this expression is exact: the output perturbation is $A\delta x$, with no higher-order term.

The worst case over nonzero inputs is

$$\sup_{x \neq 0} \kappa_{\mathrm{rel}}(F, x) = \|A\|_2 \left(\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \right)^{-1}. \tag{3.5}$$

If A is rank deficient, the minimum gain is zero and the worst-case relative condition number is infinite. If A is square and invertible, set $x = A^{-1}y$ to obtain

$$\begin{aligned} \min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \min_{y \neq 0} \frac{\|y\|_2}{\|A^{-1}y\|_2} \\ &= \left(\max_{y \neq 0} \frac{\|A^{-1}y\|_2}{\|y\|_2} \right)^{-1} = \frac{1}{\|A^{-1}\|_2}. \end{aligned}$$

Thus, for an invertible square matrix,

$$\kappa_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2. \tag{3.6}$$

We next use the singular value decomposition to understand the maximum and minimum gains geometrically.

Orthogonal matrices

A square matrix $Q \in \mathbb{R}^{n \times n}$ is *orthogonal* when

$$Q^T Q = Q Q^T = I, \qquad \text{equivalently} \qquad Q^{-1} = Q^T.$$

Writing $Q = [q_1 \ \cdots \ q_n]$, its columns form an orthonormal basis:

$$q_i^T q_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Orthogonal transformations preserve the Euclidean norm,

$$\|Qx\|_2^2 = x^T Q^T Q x = \|x\|_2^2 .$$

Singular value decomposition

For every $A \in \mathbb{R}^{n\times n}$, there are orthogonal matrices $U = [u_1 \ \cdots \ u_n]$ and $V = [v_1 \ \cdots \ v_n]$, and a diagonal, nonnegative matrix

$$S = \operatorname{diag}(\sigma_1, \ldots, \sigma_n), \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0,$$

such that

$$A = USV^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T . \tag{3.7}$$

The σ_i are the *singular values*, and

$$Av_i = \sigma_i u_i .$$

Thus V^T expresses an input in the right singular-vector basis, S scales those coordinates, and U expresses the result in the left singular-vector basis.

For any x, write

$$x = \sum_{i=1}^{n} \alpha_i v_i, \qquad \alpha_i = v_i^T x.$$

Since the v_i and u_i are orthonormal,

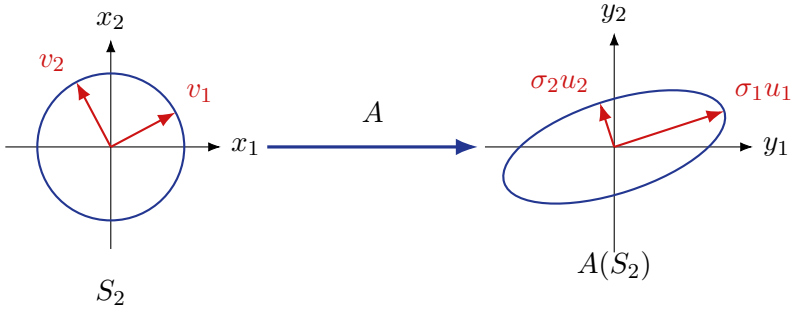
$$\|x\|_2^2 = \sum_{i=1}^{n} \alpha_i^2, \qquad Ax = \sum_{i=1}^{n} \sigma_i \alpha_i u_i, \qquad \|Ax\|_2^2 = \sum_{i=1}^{n} \sigma_i^2 \alpha_i^2 . \tag{3.8}$$

For $\|x\|_2 = 1$, equation (3.8) gives

$$\sigma_n \leq \|Ax\|_2 \leq \sigma_1 .$$

The upper bound is attained at $x = v_1$, and the lower bound at $x = v_n$. Hence

$$\|A\|_2 = \sigma_1, \qquad \min_{\|x\|_2 = 1} \|Ax\|_2 = \sigma_n . \tag{3.9}$$



The matrix condition number

Combining (3.6) and (3.9) gives the final form

$$\boxed{\kappa_2(A) = \begin{cases} \|A\|_2 \, \|A^{-1}\|_2 = \dfrac{\sigma_1(A)}{\sigma_n(A)}, & A \text{ nonsingular}, \\ +\infty, & A \text{ singular}. \end{cases}}$$

Thus $\kappa_2(A)$ is the ratio of the largest to the smallest principal stretch. It is at least 1, with equality precisely when A is a nonzero scalar multiple of an orthogonal matrix.

CS 450: Numerical Analysis – Lecture 4

SVD, Conditioning of Linear Systems, and LU Factorization

Edgar Solomonik

September 3, 2026

Today

- Review: the matrix condition number
- Singular value decomposition
- Conditioning of linear systems
- Algorithms: triangular solve and LU factorization

Review: matrix condition number

Definition. Given $A \in \mathbb{R}^{n\times n}$ with $\text{rank}(A) = n$,

$$\kappa(A) = \|A\|_2 \left\|A^{-1}\right\|_2.$$

Lemma. For all $x, \delta x \in \mathbb{R}^n$ with $x \neq 0$, let y and δy be defined by

$$y = Ax, \qquad y + \delta y = A(x + \delta x).$$

Then

$$\frac{\|\delta y\|_2}{\|y\|_2} \leq \kappa(A)\, \frac{\|\delta x\|_2}{\|x\|_2}, \tag{4.1}$$

and moreover there exist $x, \delta x$ such that equality holds.

Proof. By linearity, $\delta y = A\delta x$ and $y = Ax$, so

$$\frac{\|\delta y\|_2}{\|y\|_2} = \frac{\|A\delta x\|_2}{\|Ax\|_2} \leq \|A\|_2\, \frac{\|\delta x\|_2}{\|Ax\|_2}.$$

How small can $\|Ax\|_2$ be? Since $y = Ax$ gives $x = A^{-1}y$,

$$\|x\|_2 = \left\|A^{-1}y\right\|_2 \leq \left\|A^{-1}\right\|_2 \|y\|_2, \qquad \text{so} \qquad \|Ax\|_2 = \|y\|_2 \geq \frac{\|x\|_2}{\|A^{-1}\|_2}.$$

Combining the two bounds,

$$\frac{\|\delta y\|_2}{\|y\|_2} \leq \|A\|_2 \left\|A^{-1}\right\|_2 \frac{\|\delta x\|_2}{\|x\|_2} = \kappa(A)\, \frac{\|\delta x\|_2}{\|x\|_2}. \qquad \square$$

Equality can hold because both inequalities in the proof are attainable: there exists $\delta x \in \mathbb{R}^n$ with $\|A\delta x\|_2 = \|A\|_2 \|\delta x\|_2$ (the induced-norm maximum is attained), and similarly for A^{-1} there exists $\tilde{y}$ with $\left\|A^{-1}\tilde{y}\right\|_2 = \left\|A^{-1}\right\|_2 \|\tilde{y}\|_2$, so that $x = A^{-1}\tilde{y}$ satisfies $\|Ax\|_2 = \|x\|_2 / \left\|A^{-1}\right\|_2$. The two choices are independent of each other, so both bounds are tight simultaneously. *(Demo on matrix conditioning here.)*

The maximum- and minimum-stretch directions of A are apparent from the singular value decomposition.

Singular value decomposition

Definition (SVD). For every $A \in \mathbb{R}^{n\times n}$ there exist orthogonal matrices $U = [u_1 \ \cdots \ u_n] \in \mathbb{R}^{n\times n}$ and $V = [v_1 \ \cdots \ v_n] \in \mathbb{R}^{n\times n}$ and a diagonal, nonnegative matrix

$$S = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}, \qquad \sigma_i \geq \sigma_{i+1} \geq 0,$$

such that

$$A = USV^T = \sum_{i=1}^{n} \sigma_i \, u_i v_i^T, \qquad Av_i = \sigma_i u_i. \tag{4.2}$$

Remark: non-uniqueness of the SVD. The singular values of A are unique, but the singular vectors are not. First, signs can be pushed between u_i and v_i: replacing (u_i, v_i) by $(-u_i, -v_i)$ leaves $\sigma_i u_i v_i^T$ unchanged. Second, if a singular value repeats, $\sigma_i = \cdots = \sigma_j$, the corresponding singular vectors are determined only up to a common orthogonal change of basis: replacing $[v_i \ \cdots \ v_j]$ by $[v_i \ \cdots \ v_j]Q$ and $[u_i \ \cdots \ u_j]$ by $[u_i \ \cdots \ u_j]Q$, for any orthogonal Q of the block size, gives another valid SVD, so any orthonormal basis of the associated subspace supplies suitable singular vectors. For example, $A = I$ has $A = VIV^T$ for every orthogonal V.

Action of A in the singular bases. Consider Ax for any $x \in \mathbb{R}^n$. Expanding x in the orthonormal basis $v_1, \ldots, v_n$,

$$x = \sum_{i=1}^{n} v_i \, (v_i^T x), \qquad Ax = \sum_{i=1}^{n} u_i \, \sigma_i \, \underbrace{(v_i^T x)}_{\gamma_i}.$$

Writing $\gamma = V^T x$, the product $Ax = USV^T x$ reads, from right to left: $\gamma = V^T x$ collects the coefficients of x in the V basis; $z = S\gamma$ scales them; and $Ax = Uz$ writes the output as a linear combination of the columns of U.

Extreme stretches and the condition number. Since U is orthogonal, $\|Ax\|_2 = \|Uz\|_2 = \|z\|_2$, with

$$\|z\|_2^2 = \sum_{i=1}^{n} (\sigma_i \gamma_i)^2 = \sum_{i=1}^{n} \sigma_i^2 \gamma_i^2.$$

Since V is orthogonal, $\|\gamma\|_2 = \left\|V^T x\right\|_2 = \|x\|_2$. So for $\|x\|_2 = 1$, i.e., $\sum_i \gamma_i^2 = 1$, the maximum of $\|Ax\|_2$ is attained by choosing $\gamma_1 = 1$ and $\gamma_2 = \cdots = \gamma_n = 0$. Hence the maximizer x of $\|Ax\|_2$ is the singular vector v_1 associated with the largest singular value, and $\|A\|_2 = \sigma_1$. Likewise the minimum over $\|x\|_2 = 1$ is σ_n, attained at $x = v_n$, so

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \frac{\sigma_1}{\sigma_n}. \tag{4.3}$$

SVD of the inverse. When A is nonsingular ($\sigma_n > 0$),

$$A^{-1} = (USV^T)^{-1} = V^{-T} S^{-1} U^{-1} = V S^{-1} U^T, \qquad S^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{bmatrix}.$$

Up to reordering the diagonal into decreasing order ($1/\sigma_n \geq \cdots \geq 1/\sigma_1$), this is an SVD of A^{-1}. Its largest singular value is $1/\sigma_n$, so $\left\|A^{-1}\right\|_2 = 1/\sigma_n$, consistent with (4.3).

SVD existence (sketch)

Existence of the SVD can be proven by iterative maximization. Find

$$v_1 = \operatorname*{argmax}_{\|v\|_2 = 1} \|Av\|_2, \qquad \sigma_1 = \|Av_1\|_2, \qquad u_1 \sigma_1 = Av_1.$$

The maximum is attained because the unit sphere is compact; if $\sigma_1 = 0$, then $A = 0$ and the construction is immediate. The vector v_1 serves as the first right singular vector for some valid SVD; it is essentially unique (up to sign) exactly when $\sigma_1 > \sigma_2$, echoing the non-uniqueness remark above.

Now remove this rank-one action and consider

$$A_1 = A - u_1 \sigma_1 v_1^T,$$

which satisfies $A_1 v_1 = 0$. In terms of the SVD whose existence is being established (used here only as motivation, not in the construction), $A_1 = \sum_{i=2}^{n} u_i \sigma_i v_i^T$: a matrix of rank at most $n-1$ whose largest singular value is σ_2. Choose

$$v_2 = \operatorname*{argmax}_{\|v\|_2 = 1} \|A_1 v\|_2 .$$

Any component of v_2 along v_1 is annihilated by A_1, so dropping it and renormalizing cannot decrease the attained stretch; a maximizer may therefore be taken with $v_2 \perp v_1$. Set $\sigma_2 = \|A_1 v_2\|_2$ and $u_2 = A_1 v_2 / \sigma_2$ (one can also show $u_2 \perp u_1$), remove the next rank-one term, and continue on the remaining orthogonal directions. This yields orthonormal vectors v_i and u_i and values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ with

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = U S V^T.$$

If a residual becomes zero, assign the remaining singular values zero and complete the two orthonormal bases arbitrarily.

Linear systems of equations

Given $A \in \mathbb{R}^{n \times n}$ with $\operatorname{rank}(A) = n$ and $b \in \mathbb{R}^n$, solve for x in

$$Ax = b.$$

Full rank is needed for well-posedness, and then $x = A^{-1} b$. *(In-class exercise: formulating a linear system.)*

Conditioning of linear systems

The solution map $b \mapsto x = A^{-1} b$ is itself a linear map, so the lemma from the start of the lecture applies with A^{-1} in place of A:

$$\kappa_{\text{lin-sys}}(A) = \kappa(A^{-1}) = \left\|A^{-1}\right\|_2 \|A\|_2 = \kappa(A).$$

From first principles. Perturb the input b by δb and let δx be the resulting output perturbation:

$$A(x + \delta x) = b + \delta b, \qquad \text{so} \qquad \delta x = A^{-1} \delta b, \quad x = A^{-1} b.$$

We want the minimal k so that $\|\delta x\|_2 / \|x\|_2 \leq k \, \|\delta b\|_2 / \|b\|_2$. Directly,

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|A^{-1}\delta b\|_2}{\|A^{-1} b\|_2} \leq \left\|A^{-1}\right\|_2 \frac{\|\delta b\|_2}{\|A^{-1} b\|_2} = \underbrace{\left\|A^{-1}\right\|_2 \frac{\|b\|_2}{\|A^{-1} b\|_2}}_{\text{good bound for a given } b} \cdot \frac{\|\delta b\|_2}{\|b\|_2}. \tag{4.4}$$

For a particular right-hand side b, the underbraced factor in (4.4) may be far smaller than $\kappa(A)$. For the worst case over b, we need an upper bound on $\|b\|_2 / \|A^{-1} b\|_2$:

$$\left\|A^{-1} b\right\|_2 \geq \sigma_{\min}(A^{-1}) \, \|b\|_2 = \frac{1}{\sigma_{\max}(A)} \, \|b\|_2 = \frac{\|b\|_2}{\|A\|_2},$$

hence $\|b\|_2 / \|A^{-1} b\|_2 \leq \|A\|_2$ and

$$\boxed{\frac{\|\delta x\|_2}{\|x\|_2} \leq \|A\|_2 \left\|A^{-1}\right\|_2 \frac{\|\delta b\|_2}{\|b\|_2} = \kappa(A) \frac{\|\delta b\|_2}{\|b\|_2}.} \tag{4.5}$$

Residual and solution error. In practice, an algorithm produces an approximate solution $\widehat{x} \approx x$. The error $\widehat{x} - x$ cannot be evaluated without knowing x, but the *residual*

$$r = b - A\widehat{x}$$

can always be computed. The residual says that $\widehat{x}$ solves a nearby linear system exactly:

$$A\widehat{x} = b - r,$$

which is the perturbed system above with $\delta b = -r$ and $\delta x = \widehat{x} - x$. The bound (4.5) therefore controls the solution error in terms of the residual magnitude,

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \le \kappa(A)\, \frac{\|r\|_2}{\|b\|_2}.$$

A matching lower bound follows from $r = A(x - \widehat{x})$ and $x = A^{-1}b$, which give

$$\|r\|_2 \le \|A\|_2 \, \|x - \widehat{x}\|_2 \,, \qquad \|x\|_2 \le \left\|A^{-1}\right\|_2 \|b\|_2 \,,$$

so that

$$\boxed{\frac{1}{\kappa(A)}\, \frac{\|r\|_2}{\|b\|_2} \;\le\; \frac{\|\widehat{x} - x\|_2}{\|x\|_2} \;\le\; \kappa(A)\, \frac{\|r\|_2}{\|b\|_2}.} \tag{4.6}$$

The relative residual is thus a computable proxy for the relative error, reliable up to a factor of $\kappa(A)$ in either direction. If A is well conditioned, a small residual certifies a small error; if A is ill conditioned, it does not. The algorithms discussed next are judged by whether they produce small residuals; how much solution accuracy a small residual buys is then governed by $\kappa(A)$.

Algorithms for solving linear systems

Special case: diagonal systems. For $Dx = b$ with D diagonal,

$$x_i = b_i / d_{ii}.$$

Cost: n operations.

Special case: triangular systems. For $Lx = b$ with L lower triangular (and similarly for upper triangular), the n equations can be solved in order by forward substitution:

$$\begin{aligned}
x_1 &= b_1/\ell_{11}, \\
x_2 &= (b_2 - \ell_{21}x_1)/\ell_{22}, \\
x_3 &= (b_3 - \ell_{31}x_1 - \ell_{32}x_2)/\ell_{33}, \\
&\;\;\vdots \\
x_i &= \Big(b_i - \sum_{j=1}^{i-1} \ell_{ij}x_j\Big)\Big/\ell_{ii}.
\end{aligned}$$

Row i involves $i - 1$ multiply–add pairs, i.e., $n/2$ terms on average with one addition and one multiplication per term, so the cost is

$$\frac{n^2}{2} \cdot 2 = n^2 \quad \text{operations (to leading order).}$$

For an upper-triangular system $Ux = b$, start from the last equation (backward substitution), or note the equivalence with the lower-triangular case obtained by reversing the order of the entries of x and b.

General case: LU factorization. If $A = LU$ with L lower triangular and U upper triangular, then $Ax = b$ can be solved by two triangular solves (forward and backward substitution):

$$Ly = b, \qquad Ux = y.$$

To compute the factorization, normalize L to have unit diagonal and partition, with $a_{12}, u_{12} \in \mathbb{R}^{1\times(n-1)}$ row vectors, $a_{21}, \ell_{21} \in \mathbb{R}^{(n-1)\times 1}$ column vectors, and $A_{22}, L_{22}, U_{22} \in \mathbb{R}^{(n-1)\times(n-1)}$:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \ell_{21} & L_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & U_{22} \end{bmatrix}.$$

Equating blocks row by row,

$$\begin{aligned} \begin{bmatrix} u_{11} & u_{12} \end{bmatrix} &= \begin{bmatrix} a_{11} & a_{12} \end{bmatrix}, \\ \ell_{21} &= a_{21}/u_{11}, \\ A_{22} &= \ell_{21}u_{12} + L_{22}U_{22}, \end{aligned}$$

so that

$$\underbrace{A_{22} - \ell_{21}u_{12}}_{\text{Schur complement}} = \underbrace{L_{22}U_{22}}_{\text{LU factorization of } (n-1)\times(n-1) \text{ matrix}}.$$

The first row of U and first column of L are available directly, and the remaining factors form the LU factorization of the $(n-1)\times(n-1)$ Schur complement, which is computed by the same procedure.

The need for pivoting. If $u_{11} = a_{11} = 0$, then $\ell_{21} = a_{21}/u_{11}$ does not exist. A similar issue arises whenever some $u_{ii} = 0$; note that $u_{ii} \neq a_{ii}$ in general for $i > 1$, so the issue cannot be detected from the diagonal of A alone. Solution: pivoting—permute the rows of A so that each $u_{ii} \neq 0$.

Row pivoting. For every $A \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(A) = n$, there exist a unit-diagonal lower-triangular L, an upper-triangular U, and a permutation matrix P such that

$$PA = LU.$$

Partial pivoting. The permutation is assembled from row swaps interleaved with the elimination steps, $P = P_{n-1}\cdots P_1$ (with P_1 applied first), where each P_i is a swap matrix exchanging row i with some row $j \geq i$, e.g.,

$$P_3 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & 1 \\ & & 1 & \end{bmatrix}.$$

Each swap also reorders the previously computed multiplier columns of L. Pick each P_i so that the pivot magnitude $|u_{ii}|$ is maximal among the available rows. *(A worked example is deferred to the next lecture.)*

CS 450: Numerical Analysis – Lecture 5

LU Factorization: Existence, Partial Pivoting, and Structured Systems

Edgar Solomonik

September 8, 2026

Today

- Review: non-pivoted LU factorization
- Existence and uniqueness criteria
- Partial pivoting
- Linear systems with structure

LU factorization

Given $A \in \mathbb{R}^{n\times n}$, an LU factorization is defined by $L, U \in \mathbb{R}^{n\times n}$, with L lower triangular and unit-diagonal and U upper triangular, such that

$$A = LU.$$

Why is this useful? It reduces a general linear solve to triangular solves, and triangular matrices have convenient properties. Given $L \in \mathbb{R}^{n\times n}$ lower triangular:

- L is full rank if and only if its diagonal entries are all nonzero. If they are, consider $y = Lx$ for $x \neq 0$: if the first nonzero entry of x is x_i, then $y_i = \ell_{ii}x_i \neq 0$ (all earlier entries of x vanish), so $Lx \neq 0$. Conversely, $\det(L) = \prod_i \ell_{ii}$ vanishes if any diagonal entry does.
- L^{-1}, when it exists, has a simple triangular structure:

$$\begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}^{-1} = \begin{bmatrix} L_{11}^{-1} & 0 \\ -L_{22}^{-1}L_{21}L_{11}^{-1} & L_{22}^{-1} \end{bmatrix},$$

  as verified by multiplying out; by induction, the inverse of a lower-triangular matrix is lower triangular.
- $Lx = b$ can be solved at $O(n^2)$ cost: blockwise, $x_1 = L_{11}^{-1}b_1$ (solve), then $x_2 = L_{22}^{-1}(b_2 - L_{21}x_1)$ (substitute and solve)—forward substitution from Lecture 4.

Hence, given a factorization $A = LU$:

- $\operatorname{rank}(L) = n$, since L is unit-diagonal;
- $\operatorname{rank}(A) = \operatorname{rank}(U)$;
- $A^{-1} = U^{-1}L^{-1}$;
- $Ax = b$ can be solved at $O(n^2)$ cost via $L(Ux) = b$: first solve for y in $Ly = b$, then solve for x in $Ux = y$.

When does a non-pivoted LU exist?

Theorem. $A \in \mathbb{R}^{n\times n}$ admits a full-rank non-pivoted LU factorization if and only if every leading principal submatrix $A_{11} \in \mathbb{R}^{m\times m}$, $m = 1, \ldots, n$, is full rank, $\det(A_{11}) \neq 0$. When it exists, the factorization is unique: there is exactly one pair of a unit-diagonal lower-triangular L and an upper-triangular U with $A = LU$.

Proof. *Necessity.* Let $A = LU$ be a full-rank LU factorization and fix $m \leq n$. Partitioning conformally with $A_{11} \in \mathbb{R}^{m\times m}$,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix},$$

the leading blocks give $A_{11} = L_{11}U_{11}$. Since $\det(L_{11}) = 1$ and the diagonal entries $u_{11}, \ldots, u_{mm}$ of U_{11} are nonzero ($\prod_{i=1}^{n} u_{ii} = \det(U) \neq 0$), $\det(A_{11}) = \prod_{i=1}^{m} u_{ii} \neq 0$.

Sufficiency and uniqueness. Both follow from a single induction on the leading block size. For $m = 1$, the only factorization is $[a_{11}] = [1] \cdot [a_{11}]$, full rank when $a_{11} \neq 0$. Suppose the leading block $A_{11} \in \mathbb{R}^{m\times m}$ has exactly one full-rank LU factorization $L_{11}U_{11}$, and border it with the next row and column of A (a_{12} a column, a_{21} a

row, a_{22} a scalar). Then

$$\begin{bmatrix} A_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ a_{21}U_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} U_{11} & L_{11}^{-1}a_{12} \\ 0 & a_{22} - a_{21}A_{11}^{-1}a_{12} \end{bmatrix},$$

as verified by multiplying out, using $a_{21}U_{11}^{-1}\,L_{11}^{-1}a_{12} = a_{21}A_{11}^{-1}a_{12}$; and any full-rank LU factorization of the bordered block coincides with it, since the block equations force the factors in turn: the leading blocks form a full-rank LU factorization of A_{11} (their diagonals lie on the bordered factors' diagonals), hence equal L_{11} and U_{11} by the induction hypothesis, and the bordered column of U, row of L, and final diagonal entry of U are then each uniquely determined. Since the determinant of the bordered block equals $\det(A_{11})\,(a_{22} - a_{21}A_{11}^{-1}a_{12})$, the new pivot is nonzero exactly when the bordered block is nonsingular, and the factorization is then full rank. □

If instead $a_{22} - a_{21}A_{11}^{-1}a_{12} = 0$ at some step, the construction still produces an LU factorization of the bordered block, but with a zero pivot—a rank-deficient "low-rank" LU.

The Schur complement and the principal pivot transform

Note the role of A_{11} in the LU factorization in general. Equating blocks,

$$A_{11} = L_{11}U_{11}, \qquad A_{21} = L_{21}U_{11}, \qquad A_{12} = L_{11}U_{12}, \qquad A_{22} = L_{21}U_{12} + L_{22}U_{22},$$

so the trailing factors satisfy $L_{22}U_{22} = A_{22} - L_{21}U_{12}$, the Schur complement, and

$$S = A_{22} - L_{21}U_{12} = A_{22} - A_{21}\underbrace{U_{11}^{-1}\,L_{11}^{-1}}_{A_{11}^{-1}}A_{12} = A_{22} - A_{21}A_{11}^{-1}A_{12}. \tag{5.1}$$

For a block system $A\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, S describes the reduced system. In particular:

Lemma (principal pivot transform). If A_{11} is full rank, then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \qquad \Longrightarrow \qquad \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1} & S \end{bmatrix} \begin{bmatrix} b_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ b_2 \end{bmatrix}, \tag{5.2}$$

with $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$: the transform exchanges the input block x_1 with the output block b_1 of the linear map.

Proof. Premultiplying the first block row $A_{11}x_1 + A_{12}x_2 = b_1$ by A_{11}^{-1} gives

$$x_1 + A_{11}^{-1}A_{12}x_2 = A_{11}^{-1}b_1, \qquad \text{i.e.,} \qquad x_1 = A_{11}^{-1}b_1 - A_{11}^{-1}A_{12}x_2,$$

the first block row of the claim. Substituting x_1 into the second block row $A_{21}x_1 + A_{22}x_2 = b_2$ gives

$$A_{21}A_{11}^{-1}b_1 + \underbrace{(A_{22} - A_{21}A_{11}^{-1}A_{12})}_{S}\,x_2 = b_2,$$

the second block row. □

In-class exercise: a formula for A^{-1} via the PPT. Use the principal pivot transform to derive a formula for A^{-1}. *Resolution.* Apply the transform twice: first forward, then backward. The forward application (5.2) exchanges the input block x_1 with the output block b_1; call the resulting matrix T, so that $T\begin{bmatrix} b_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ b_2 \end{bmatrix}$. Now apply the transform to T backward, pivoting on its trailing diagonal block $T_{22} = S$ (the same lemma with the roles of the two blocks exchanged, proved identically), which exchanges x_2 with b_2:

$$\begin{bmatrix} T_{11} - T_{12}S^{-1}T_{21} & T_{12}S^{-1} \\ -S^{-1}T_{21} & S^{-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The matrix on the left maps b to x and hence equals A^{-1}; substituting the blocks of T from (5.2) gives

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}, \tag{5.3}$$

valid whenever A_{11} and S are invertible; given A_{11} invertible, S is invertible exactly when A is, since $\det(A) = \det(A_{11})\det(S)$. Equation (5.3) is the block form of the Gauss–Jordan inversion formula.

Low-rank LU and numerical pitfalls

If $\operatorname{rank}(A) = r < n$, one can hope for a "low-rank" LU, $A = LU$ with $L \in \mathbb{R}^{n\times r}$ lower-trapezoidal (unit diagonal) and $U \in \mathbb{R}^{r\times n}$ upper-trapezoidal. Again, existence requires $\det(A_{11}) \neq 0$ for the leading blocks $A_{11} \in \mathbb{R}^{m\times m}$ with $m \leq \operatorname{rank}(A)$. This condition may not hold in general, and it is hard to predetermine.

Moreover, a *small* $\det(A_{11})$ is also problematic numerically. Returning to the algorithm with the first-row/column partition,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \ell_{21} & L_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & U_{22} \end{bmatrix},$$

we read off $u_{11} = a_{11}$ and $u_{12} = a_{12}$, then

$$\ell_{21} = a_{21}/u_{11},$$

which exists if $u_{11} \neq 0$ but is large in magnitude if $a_{11} \approx 0$, and finally

$$L_{22}U_{22} = A_{22} - \ell_{21}u_{12},$$

into which a large ℓ_{21} propagates—and from there into every later step. Note also that

$$\det(A_{11}) = \det(L_{11})\det(U_{11}) = \prod_{i=1}^{m} u_{ii}$$

for each leading block (unit diagonal of L), so $u_{mm} = \det(A_{11}^{(m)})/\det(A_{11}^{(m-1)})$: the nonzero-minor condition is exactly the nonzero-pivot condition, and small minors manifest as small pivots and large multipliers.

Partial pivoting

The general cure for both the existence and the numerical issues is partial pivoting.

Permutation matrices. $P \in \{0,1\}^{n\times n}$ is a permutation matrix if each column of P is a distinct column of the identity matrix. Properties:

- $P^T = P^{-1}$ (that is, $P^TP = I$);
- $\|P\|_m = 1$ for $m = 1, 2, \infty$, and consequently $\kappa(P) = 1$: permuting is perfectly conditioned;
- Px can be computed in $O(n)$ time—a reordering, with no arithmetic.

A row swap is the special case exchanging entries j and k: $y = Px$ has $y_i = x_i$ except that $y_j = x_k$ and $y_k = x_j$, given explicitly by

$$P = I - (e_j - e_k)(e_j - e_k)^T.$$

Pivoted LU. A row-pivoted LU factorization is

$$PA = LU,$$

with P a permutation matrix. Suitable P, L, U exist for any A with $\operatorname{rank}(A) = n$ (more care—column pivoting—is needed in the low-rank case).

Partial pivoting computes a row-pivoted LU factorization whose permutation is a product of (elementary) swap permutation matrices,

$$P = P_{n-1}\cdots P_1$$

(P_1 applied first), where each P_k is a swap of the form above, exchanging row k with some row $j \geq k$ chosen during the elimination.

The GEPP algorithm. Gaussian elimination with partial pivoting computes $[L, U, P] = \mathrm{GEPP}(A, n)$ recursively:

function GEPP(A, n)
  if $n = 1$ **then**
    return $L = [1],\ U = [a_{11}],\ P = [1]$
  end if
  find i such that $|a_{i1}|$ is maximal
  $P_1 \leftarrow I - (e_1 - e_i)(e_1 - e_i)^T; \quad B \leftarrow P_1 A$
  $u_{11} \leftarrow b_{11}, \quad u_{12} \leftarrow b_{12}, \quad \ell_{21} \leftarrow b_{21}/b_{11}$ ▷ one step of LU on B
  $[L_{22}, U_{22}, P_2] \leftarrow \mathrm{GEPP}(B_{22} - b_{21}b_{11}^{-1}b_{12},\ n-1)$ ▷ recurse on the Schur complement
  return $L = \begin{bmatrix} 1 & 0 \\ P_2\,\ell_{21} & L_{22} \end{bmatrix}, \quad U = \begin{bmatrix} b_{11} & b_{12} \\ 0 & U_{22} \end{bmatrix}, \quad P = \begin{bmatrix} 1 & 0 \\ 0 & P_2 \end{bmatrix} P_1$
end function

Note that the later permutation P_2 is applied to the earlier multiplier column b_{21}/b_{11}. Correctness ($PA = LU$) follows by multiplying out, using $L_{22}U_{22} = P_2(B_{22} - b_{21}b_{11}^{-1}b_{12})$ from the recursive call.

Properties of GEPP:

- GEPP produces an LU factorization for any full-rank A: the pivot search fails only if the current first column is entirely zero, which would make that Schur complement—and hence A—rank-deficient.
- All subdiagonal entries of L have magnitude at most 1, since each multiplier is $\ell_{21} = b_{21}/b_{11}$ with $|b_{11}|$ maximal in its column.

AI-use disclosure. *These notes were prepared with the assistance of Claude and ChatGPT, including transcription from handwritten lecture notes and drafting or editing of prose, examples, and derivations. The author reviewed the final document and assumes responsibility for its content.*