Numerical Linear Algebra for Computational Science and Information Engineering

Extra notes

by Nicolas Venkovic

School of Computation, Information and Technology (CIT), Technical University of Munich, Germany

Summer 2025

Each extra note refers to a specific lecture which we covered in class. Those lectures are listed as follows:

- Lecture 01 – Essentials of linear algebra
- Lecture 02 – Essentials of the Julia language
- Lecture 03 – Floating-point arithmetic and error analysis
- Lecture 04 – Direct methods for dense linear systems
- Lecture 05 – Sparse data structures and basic linear algebra subprograms
- Lecture 06 – Introduction to direct methods for sparse linear systems
- Lecture 07 – Orthogonalization and least-squares problems
- Lecture 08 – Basic iterative methods for linear systems
- Lecture 09 – Basic iterative methods for eigenvalue problems
- Lecture 10 – Locally optimal block preconditioned conjugate gradient
- Lecture 11 – Arnoldi and Lanczos procedures
- Lecture 12 – Krylov subspace methods for linear systems
- Lecture 13 – Multigrid methods
- Lecture 14 – Preconditioned iterative methods for linear systems
- Lecture 15 – Restarted Krylov subspace methods
- Lecture 16 – Elements of randomized numerical linear algebra
- Lecture 17 – Introduction to communication-avoiding algorithms
- Lecture 18 – Matrix function evaluation

Lecture 01 – Essentials of linear algebra

Hints for Pb. 3 – $\|xy^T\|_2 = \|x\|_2\|y\|_2$

- As per slide #47, we have $\|A\|_2 = \sigma_{max}(A) = \lambda_{max}(A^TA)^{1/2}$.
- Use this to express $\|xy^T\|_2$ in terms of $\|x\|_2$ and yy^T.
- Since yy^T is symmetric, its Rayleigh quotient is maximized by the eigenvector with largest eigenvalue.
- Use this fact, with the Cauchy-Schwarz inequality, to finish-up the proof.

Hints for Pb. 5 – SVD, pseudo-inverse, and associated projections

- First, from its low-rank nature, the matrix admits a decomposition $A = U\Sigma V^T$. Then, as per slide #20, the pseudo-inverse, which is unique, must satisfy 4 identities, namely $AA^\dagger A = A$, $A^\dagger AA^\dagger = A^\dagger$, $(AA^\dagger)^T = AA^\dagger$ and $(A^\dagger A)^T = A^\dagger A$.

- To prove that a matrix P is an orthogonal projector onto $\mathcal{M}$, there are three things to do:

  1. Show that P is a projector by verifying that $P^2 = P$.
  2. Show that P projects onto $\mathcal{M}$, typically by verifying $\text{range}(P) \subseteq \mathcal{M}$ and $\mathcal{M} \subseteq \text{range}(P)$.
  3. Show that P is an orthogonal projector.

     In general, this is done by showing $x - Px \perp \text{range}(P)$ for all x.

     Alternatively, when orthogonality is stated with respect to the dot product, an orthogonal projector is Hermitian, as can be checked from the general formulation of orthogonal projectors we derived together on slide #28. Actually, a stronger statement holds, namely, if orthogonality is stated through the dot product, then a projector is orthogonal if and only if it is Hermitian. Therefore, here, it suffices to show $P^T = P$.

More on orthogonal projectors

- We defined orthogonal matrices as $Q \in \mathbb{R}^{m\times n}$ such that $Q^TQ = I_n$.

- Let P be a projector, orthogonal with respect to the dot product, onto a proper subspace of $\mathbb{R}^n$, i.e., $\text{range}(P) \subset \mathbb{R}^n$. Is P an orthogonal matrix? No.

  - Since the projector is orthogonal, and orthogonality is stated with respect to the dot product, we have $P^T = P$ (as previously stated), so that $P^TP = P^2$.
  - But since P is a projector, we also have $P^2 = P$, so that, for a projector orthogonal with respect to the dot product, we have $P^TP = P$.
  - Thus, if P were an orthogonal matrix, we would have $P^TP = I_n \implies P = I_n$.
  - But since $\text{range}(P)$ is a proper subspace of $\mathbb{R}^n$, the rank of P must be smaller than n, so $P \neq I_n$.

    Therefore, a projector which is orthogonal with respect to the dot product, onto a proper subspace of $\mathbb{R}^n$, cannot be an orthogonal matrix, i.e., $P^TP \neq I_n$.

Lecture 03 – Floating-point arithmetic and error analysis

Slide #12 – Perturbation of linear systems

Here, the bound on $\|\delta x\|$ is obtained by applying the triangular inequality of vector norms, and assuming the matrix norm is consistent with the vector norm.

In L01, we said a matrix norm is consistent with a vector norm if $\|Ax\| \leq \|A\|\|x\|$.

We said that induced (matrix) norms are, by definition, consistent with the vector norm they're induced from, e.g., $\|Ax\|_2 \leq \|A\|_2\|x\|_2$.

We also said that, although the Frobenius norm is not induced by any vector norm, it is consistent with the 2-norm, i.e., $\|Ax\|_2 \leq \|A\|_F\|x\|_2$.

Slide #13 – Perturbation of linear systems, cont'd

- The conditioning number of a linear system is defined as $\kappa(A) := \|A^{-1}\|\|A\|$.

- As we saw in L01, $\|A\|_2 = \sigma_{max}(A)$. But since the singular values of A^{-1} are the inverses of those of A, we have that $\sigma_{max}(A^{-1}) = 1/\sigma_{min}(A)$, so that $\|A^{-1}\|_2 = 1/\sigma_{min}(A)$.

- Therefore, when using the 2-norm, the conditioning number of a linear system is given by $\kappa(A) = \sigma_{max}(A)/\sigma_{min}(A)$.

- For the case of normal matrices, i.e., for matrices $A \in \mathbb{F}^{n\times n}$ such that $A^HA = AA^H$, from being diagonalizable, and assuming the eigenvalues are ordered such that $|\lambda_1(A)| \geq \cdots \geq |\lambda_n(A)|$, we have $\kappa(A) = |\lambda_1(A)|/|\lambda_n(A)|$.

Slide #14 – Backward error of linear systems

Here, we used the reverse triangular inequality $\|x+y\| \geq \|x\| - \|y\|$, which holds for all norms, and can be derived as follows from the standard triangular inequality:

$$\begin{aligned} \|(x+y)-y\| &\leq \|x+y\| + \|-y\| \\ \|x\| &\leq \|x+y\| + \|y\| \\ \|x\| - \|y\| &\leq \|x+y\|. \end{aligned}$$

Slide #18 – Backward error of an eigenpair

- In practice, it is common to monitor the convergence of an approximate eigenpair $(\tilde{\lambda}, \tilde{u})$ using the criterion $\dfrac{\|r\|}{|\tilde{\lambda}| \cdot \|\tilde{u}\|} \leq \varepsilon$ instead of the backward error $\dfrac{\|r\|}{\|A\| \cdot \|\tilde{u}\|}$.
- Since, when using 2-norms, we have $|\lambda_1(A)| \leq \|A\|_2$, and often $|\lambda_n(A)| \ll |\lambda_1(A)|$, using the aforementioned criterion can be too conservative, and result into failure to properly diagnose convergence when trying to solve for eigenvalues with smaller magnitude.
- However, as we saw from the equivalence of matrix norms in L01, we have $\|A\|_2 \leq \|A\|_F$, so that using a criterion $\dfrac{\|r\|_2}{\|A\|_F \cdot \|\tilde{u}\|_2} \leq \varepsilon$ can be a less conservative alternative, as it often is easier to approximate $\|A\|_F$ than $\|A\|_2$.

Hint for Pb. 7 – Unit roundoff of floating-point number systems with given precision

Note that $\mathrm{fl}(1) = 1$ for all floating-point number systems, irrespective of precision.

Hints for Pb. 9 – Backward error of approximate solutions of linear systems

- In slides #14-15, we defined the backward error $\eta_{A,b}(\tilde{x})$ of an approximation $\tilde{x}$ of x such that $Ax = b$ as the minimal achievable norm of perturbations δA and δb such that $\tilde{x}$ solves exactly the perturbed system. That is, for all δA and δb such that $(A + \delta A)\tilde{x} = b + \delta b$, we have $\|\delta A\| \geq \eta_{A,b}(\tilde{x})\|A\|$ and $\|\delta b\| \geq \eta_{A,b}(\tilde{x})\|b\|$.
- You can use the fact that $\|xy^T\|_2 = \|x\|_2\|y\|_2$ for all $x, y \in \mathbb{R}^n$.

Lecture 04 – Direct methods for dense linear systems

Slide #24 – Proof of existence of Cholesky factorization by inductive construction

On slide #24, we show that the HPD matrix A can be recast into

$$A = \begin{bmatrix} a_{11} & a_1^H \\ a_1 & A_1 \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11} l_1^H \\ l_{11} l_1 & L_1 L_1^H + l_1 l_1^H \end{bmatrix}$$

where, by construction, we impose $l_{11} > 0$. However, to do so, we assume that the Cholesky factorization

$$L_1 L_1^H = A_1 - l_1 l_1^H$$

exists. To prove the existence of such a factorization, we need to show that $A_1 - l_1 l_1^H$ is HPD. For that, we recast A into a form XBX^H, i.e.,

$$A = \begin{bmatrix} a_{11} & a_1^H \\ a_1 & A_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_1/l_{11} & I_{n-1} \end{bmatrix} \begin{bmatrix} l_{11}^2 & 0 \\ 0 & A_1 - l_1 l_1^H \end{bmatrix} \begin{bmatrix} 1 & l_1^H/l_{11} \\ 0 & I_{n-1} \end{bmatrix}$$

where $X = \begin{bmatrix} 1 & 0 \\ l_1/l_{11} & I_{n-1} \end{bmatrix}$ and $B = \begin{bmatrix} l_{11}^2 & 0 \\ 0 & A_1 - l_1 l_1^H \end{bmatrix}$. We then claim that, because X is non-singular ($\Longleftarrow$ X is triangular with non-zero diagonal components) and A is HPD, $A_1 - l_1 l_1^H$ is indeed HPD. However, this statement should perhaps be further detailed in order to be well-understood. Such explanations are

- If $A = XBX^H$ is HPD and X is non-singular, then B is HPD. See Pb. 12.b. Therefore, $\begin{bmatrix} l_{11}^2 & 0 \\ 0 & A_1 - l_1 l_1^H \end{bmatrix}$ is HPD.
- Since $A_1 - l_1 l_1^H$ is a principal sub-matrix of an HPD matrix, it is also HPD. See Pb. 12.a.

Hint for Pb. 12.c – Principal sub-matrices of non-singular matrices

You may disprove the statement simply by providing a counter example.

Lecture 07 – Orthogonalization and least-squares problems

Slide #29 – Least-squares problem with rank-deficient matrix

Let x_0 be the LSQ problem solution with minimal norm. Then, for each $\delta x \in \text{null}(A)$, we have $A(x_0 + \delta x) = Ax_0$ so that $x_0 + \delta x$ is also an LSQ problem solution. Additionaly, by definition of x_0, we need to also have

$$\begin{aligned} \|x_0\|_2 &< \|x_0 + \delta x\|_2 \\ 0 &< \delta x^T \delta x + 2x_0^T \delta x. \end{aligned}$$

For the above expression to be true for all $\delta x \in \text{null}(A)$, we need to have $x_0^T \delta x = 0$, i.e., $x_0 \perp \text{null}(A)$.