Mathematics for Data Science — Lesson 6

SVD, Norms, Convergence, Calculus

Course Slides

Class 6 — 17 Oct 2024

Contents

Lesson Objectives

- Distinguish operator vs spectral norm and interpret maximal stretch.
- Use SVD to form rank-k approximations and reason about information retention.
- Decide convergence of sequences and compare growth with $\mathcal{O}$, o, Θ.
- Apply first-order linearization (Taylor) in one and several variables.
- Compute gradients and Jacobians; apply the chain rule with correct dimensions.
- Connect these tools to optimization stability, PCA, and backpropagation.

This Section

Operator Norm — Definition (Subordinate)

Definition (subordinate/induced)

Fix a vector norm $\|\cdot\|$ on $\mathbb{R}^n$. For $A \in \mathbb{R}^{m\times n}$, the associated operator norm is

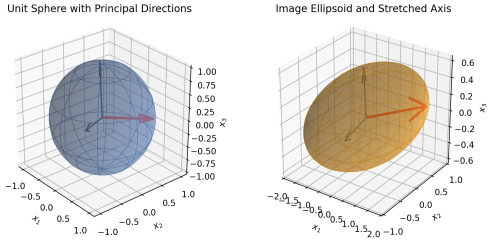
$$\|A\|_{\text{op}} = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

It is subordinate to $\|\cdot\|$ in that $\|Ax\| \leq \|A\|_{\text{op}}\,\|x\|$ for all x.

Defines a norm:

- Nonnegativity/definiteness:
  $\|A\|_{\text{op}} \geq 0$ and $\|A\|_{\text{op}} = 0 \iff A = 0$.
- Homogeneity:
  $\|\alpha A\|_{\text{op}} = |\alpha|\,\|A\|_{\text{op}}$ for any scalar α.
- Triangle inequality:
  $\|A + B\|_{\text{op}} \leq \|A\|_{\text{op}} + \|B\|_{\text{op}}$.

The matrix norm is **induced** by the vector norm.



Maximal stretch: the longest radius of A(unit ball).

Operator Norm — Core Properties

Unit-sphere reduction.

- For any nonzero x, write $x = \|x\|\, u$ with $\|u\| = 1$. Then

$$\frac{\|Ax\|}{\|x\|} = \frac{\|A(\|x\|u)\|}{\|x\|} = \|Au\|.$$

  Maximising over all $x \neq 0$ is therefore equivalent to maximising $\|Au\|$ over the unit sphere.
- The entire behaviour of A is encoded in the image of the unit sphere $A(\mathbb{S}^{n-1})$: scaling any nonzero point back onto the sphere collapses the search to that boundary. (For linear algebraists, one could equally look at the images of basis vectors, but the sphere offers clearer geometric intuition.)

Examples.

- I_n: $A(\mathbb{B}) = \mathbb{B}$, so $\|I_n\|_{\text{op}} = 1$.
- Scalar multiple αI_n: stretches every radius uniformly, giving $\|\alpha I_n\|_{\text{op}} = |\alpha|$.
- Diagonal $D = \text{diag}(d_1, \ldots, d_n)$ (for any absolute/ℓ_p norm): the unit sphere is rescaled axis-wise, so $\|D\|_{\text{op}} = \max_i |d_i|$.
- Orthogonal (Euclidean case): Q rotates/reflects the ball without changing shape, hence $\|Q\|_2 = 1$.

Operator Norm — Lipschitz and Composition

Lipschitz constant viewpoint.

- For any x, y, $\|Ax - Ay\| = \|A(x-y)\| \le \|A\|_{\mathrm{op}} \|x - y\|$. Thus $\|A\|_{\mathrm{op}}$ is the smallest global Lipschitz constant of the linear map $x \mapsto Ax$ with respect to $\|\cdot\|$.
- Stability: a perturbation δx cannot be amplified by more than $\|A\|_{\mathrm{op}}$ at the output. The longest radius of the deformed unit sphere visualises this constant.

Composition $\to$ submultiplicativity.

- If F and G are Lipschitz with constants L_F and L_G, then $F \circ G$ is Lipschitz with constant $L_F L_G$.
- Specialising to linear maps A and B yields $\|AB\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}}$.
- Proof sketch: for any $x \neq 0$,

$$\frac{\|ABx\|}{\|x\|} = \frac{\|A(Bx)\|}{\|Bx\|} \cdot \frac{\|Bx\|}{\|x\|} \le \|A\|_{\mathrm{op}} \|B\|_{\mathrm{op}},$$

  hence taking the supremum over $\|x\| = 1$ gives the claim.

Visual example: first deform the unit sphere by B, then by A; the total reach is bounded by the product of individual reaches. This mirrors how function composition multiplies Lipschitz constants.

Operator Norm — Worked Example (Subordinate)

- Let $A = \begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix}$. For the 1-norm on $\mathbb{R}^2$, the subordinate matrix norm is the maximum absolute column sum:

$$\|A\|_1 = \max\{\, |2| + |1|,\ |-1| + |3| \,\} = \max\{3, 4\} = 4.$$

- For the infinity-norm, the subordinate matrix norm is the maximum absolute row sum:

$$\|A\|_\infty = \max\{\, |2| + |-1|,\ |1| + |3| \,\} = \max\{3, 4\} = 4.$$

- Hence $\|A\|_1 = \|A\|_\infty = 4$, yielding Lipschitz bounds $\|Ax\|_1 \le 4\,\|x\|_1$ and $\|Ax\|_\infty \le 4\,\|x\|_\infty$ for all x.

Operator Norm — ML Application

- Step-size selection: $\|A\|_{\text{op}}$ is a Lipschitz constant for $x \mapsto Ax$. Picking $\eta < 1/\|A\|_{\text{op}}$ avoids unstable jumps in simple gradient-like updates.
- Regularization and stability: bounding $\|A\|_{\text{op}}$ (e.g., spectral norm regularization) curbs worst-case amplification and can reduce exploding activations.
- Preprocessing calibration: when scaling or whitening features with a matrix A, the bound $\|Ax\| \leq \|A\|_{\text{op}} \|x\|$ certifies that magnitudes stay within predictable ranges.
- Geometry-aware modelling: the operator norm inherits the geometry of the chosen vector norm. In audio, for instance, one can define a perceptual norm on a time–frequency representation and regularise network layers via the induced operator norm to respect perceptual fidelity while optimising.
- Functional viewpoint: $\|\cdot\|_{\text{op}}$ is a norm on linear maps, so it extends to continuous operators on infinite-dimensional Banach spaces (though some linear maps may fail to be bounded in that setting).

Spectral Norm — Definition and Context

Definition

For any $A \in \mathbb{R}^{m \times n}$, the spectral norm is defined by the largest singular value

$$\|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}.$$

- Through the SVD $A = U\Sigma V^\top$, each right singular vector v_i (columns of V) is sent to the left singular vector u_i (columns of U) with magnitude scaled by σ_i: $Av_i = \sigma_i u_i$. Because $\{v_i\}$ forms an orthonormal basis, the singular values quantify the stretch of orthonormal directions.
- The largest singular value $\sigma_{\max}$ therefore measures the maximal amplification of any unit vector; it is attained on the right singular vector $v_{\max}$ and points along $u_{\max}$ after applying A.
- For symmetric (or more generally normal) matrices, singular values equal absolute eigenvalues, so $\|A\|_2 = \max_i |\lambda_i|$ and eigenvectors give the principal axes.
- The definition extends directly to bounded linear operators on Hilbert spaces, replacing matrices by their singular-value spectrum.

Spectral Norm — Properties and Intuition

- **Orthogonal invariance:** $\|QAR\|_2 = \|A\|_2$ for any orthogonal Q, R; the norm depends only on singular values.
- **Ellipsoid viewpoint:** A maps the unit sphere to an ellipsoid whose principal axes are the singular vectors $\{u_i\}$; the longest semi-axis has radius $\sigma_{\max}(A)$, showing $\|A\|_2$ is the maximal stretch.
- **Rayleigh quotient (symmetric/normal case):** when A is normal, singular vectors coincide with eigenvectors and $\|A\|_2 = \max_{\|x\|_2=1} |x^\top A x| = \sqrt{\lambda_{\max}(A^\top A)} = \sqrt{\|A^\top A\|_2}$.
- **Bounds:** $\|A\|_2 \le \|A\|_F \le \sqrt{\operatorname{rank}(A)}\, \|A\|_2$; combined with submultiplicativity, this yields tight control in analyses.

Spectral Norm — Worked Example

- Consider $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Compute $A^\top A = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$.
- The eigenvalues of $A^\top A$ solve $\det(A^\top A - \lambda I) = 0$: $\lambda_{1,2} = 5 \pm 4$, i.e. $\lambda_1 = 9$ and $\lambda_2 = 1$.
- Singular values are $\sigma_1 = \sqrt{9} = 3$ and $\sigma_2 = \sqrt{1} = 1$. Thus $\|A\|_2 = \sigma_1 = 3$.
- The eigenvector corresponding to λ_1 (or σ_1) is $v = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Indeed, $\|Av\|_2 = 3$ and $\|v\|_2 = 1$.
- Since A is symmetric, v is also the eigenvector of A with eigenvalue 3, confirming the geometric interpretation.

Spectral Norm — Operator Norm Derivation

- Start from the Euclidean operator norm: $\|A\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \|Ax\|_2$.
- Use the SVD $A = U\Sigma V^\top$ and write $x = Vy$. Orthogonality gives $\|Ax\|_2 = \|\Sigma y\|_2$ with $\|y\|_2 = 1$, so the supremum is $\sigma_{\max}(A)$.
- Conversely, $\sigma_{\max}(A)$ is attained at the first right singular vector, confirming $\|A\|_{\mathrm{op}} = \|A\|_2$.
- This argument extends verbatim to bounded operators between Hilbert spaces, replacing matrices by linear maps and the SVD by the singular-value decomposition of compact operators.

Spectral Norm — ML Application

- **Hessian analysis:** $\|\nabla^2 f(x)\|_2$ bounds curvature, guiding second-order optimization methods.
- **Covariance matrices:** $\|C\|_2$ indicates the strongest variance direction in PCA and explains dominant modes.
- **Stability:** bounding $\|A\|_2$ for update matrices prevents exploding gradients in recurrent networks.
- **Infinite-dimensional learning:** in kernel methods or physics-informed models, $\|T\|_2$ for a bounded operator T on a Hilbert space controls the amplification of functions just as in the finite-dimensional case.

SVD — Definition and Context

Definition

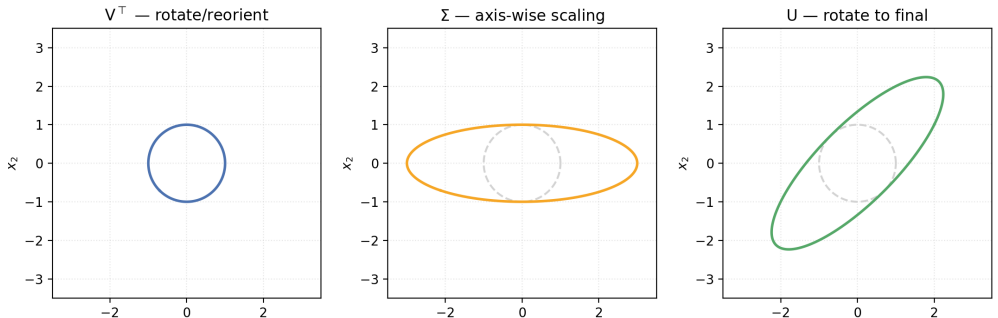
Every $A \in \mathbb{R}^{m \times n}$ admits a singular value decomposition $A = U\Sigma V^\top$ where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and Σ is diagonal with non-negative entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

- Columns $\{u_i\}$ of U and $\{v_i\}$ of V form orthonormal bases of output and input directions; $Av_i = \sigma_i u_i$.
- The decomposition rewrites A as a sum of structured rank-one terms: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$.
- Singular values quantify how A stretches orthonormal directions; σ_1 matches the spectral/Euclidean operator norm.

SVD — Properties and Geometry

- **Orthogonal invariance:** Pre/post-multiplying by orthogonal matrices does not change singular values; only geometry matters.
- **Spectral norm link:** $\sigma_1 = \|A\|_2$ and σ_i^2 are eigenvalues of $A^\top A$ (or $AA^\top$).
- **Rank and energy:** The number of non-zero σ_i equals rank(A), and $\|A\|_F^2 = \sum_i \sigma_i^2$.
- **Eckart–Young:** Truncating after k terms gives the best rank-k approximation in both $\|\cdot\|_2$ and $\|\cdot\|_F$.

2D Geometry: below, the unit circle undergoes $V^\top$ (reorientation), Σ (axis-wise scaling), and U (final orientation):

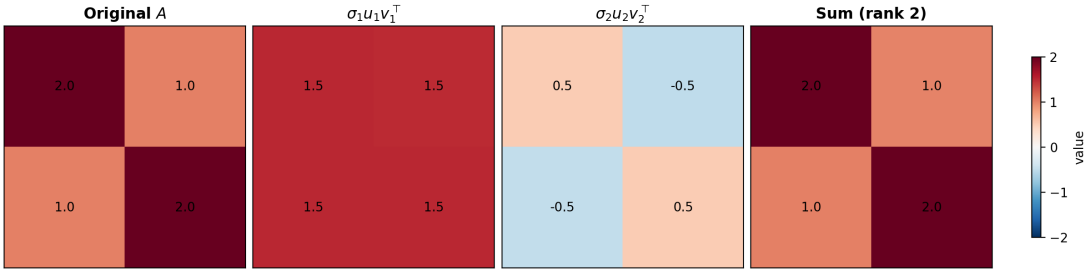


SVD — Worked Example

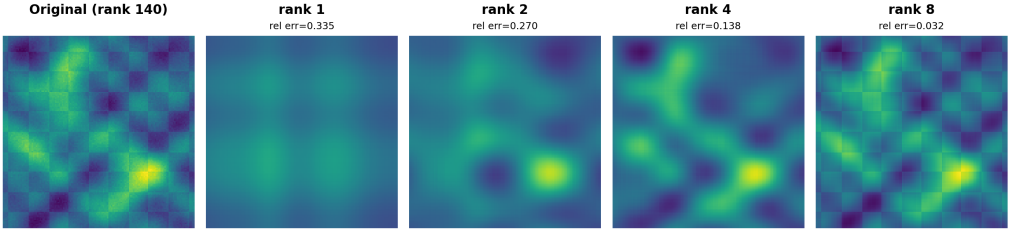
- Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Compute $A^\top A = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$.
- Eigenvalues of $A^\top A$ are 9 and 1, so singular values are $\sigma_1 = 3$ and $\sigma_2 = 1$.
- A corresponding orthonormal pair of right singular vectors is $v_1 = \frac{1}{\sqrt{2}}[1\ 1]^\top$ and $v_2 = \frac{1}{\sqrt{2}}[1\ -1]^\top$; the left singular vectors equal these because A is symmetric.
- Rank-one approximation keeps $\sigma_1 u_1 v_1^\top$, yielding $A_1 = 3\, v_1 v_1^\top = \frac{3}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$; $\|A - A_1\|_F = \sqrt{2}$ and $\|A - A_1\|_2 = 1$.

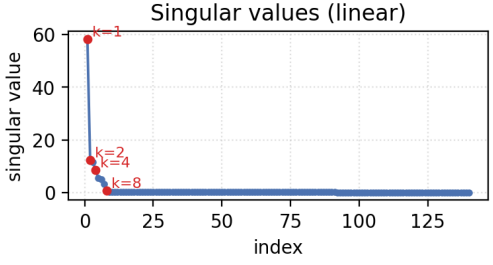
SVD — Sum of Rank-1 Matrices

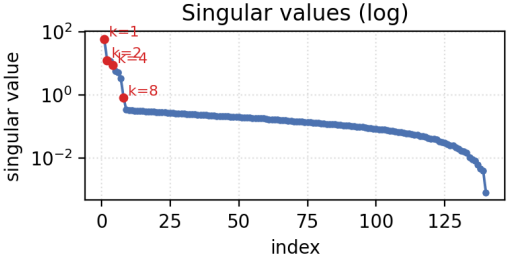
- The SVD writes A as a sum of r rank-1 outer products: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$.
- Each term is interpretable: v_i picks a direction in input space, u_i gives the corresponding output direction, and σ_i sets the strength.
- Truncating the sum at k keeps the top-k directions and gives the best rank-k approximation.



SVD — Visualization







SVD — Least Squares via Pseudoinverse

- Problem: given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, solve $x^* = \arg\min_x \|Ax - b\|_2$.
- With the SVD $A = U\Sigma V^\top$ and pseudoinverse $A^+ = V\Sigma^+ U^\top$ (invert nonzero singular values), the solution is $x^* = A^+ b$.
- Geometry: $\hat{b} = Ax^* = U_r U_r^\top b$ is the orthogonal projection of b onto range(A); the residual lies in range$(A)^\perp$.
- Minimal norm: among all least-squares solutions, x^* has minimal Euclidean norm (important for underdetermined systems).
- Conditioning/regularisation: $\kappa_2(A) = \sigma_1/\sigma_r$ measures sensitivity. Truncated SVD (TSVD) $x_k = V_k \Sigma_k^{-1} U_k^\top b$ filters small singular values and stabilises the fit.

SVD — ML Application

- **Principal Component Analysis:** project centred data onto top singular vectors to denoise and reduce dimension.
- **Latent-factor models:** approximate user–item or word–context matrices with low rank to recover hidden structure.
- **Model compression:** replace dense layers by low-rank factors to lower parameter counts and accelerate inference.
- **LoRA (Low-Rank Adaptation):** fine-tune large language models by learning low-rank updates to pretrained weights, reducing memory and compute.
- **Preprocessing reminder:** centre (and often scale) features before applying SVD/PCA; otherwise the first singular direction reflects the mean.

This Section

Sequences — Definition and Context

Definitions in $(\mathbb{R}^d, \|\cdot\|)$

$$
\begin{aligned}
(x_n) \to x^\star &\iff \forall \varepsilon > 0,\ \exists N :\ n \geq N \Rightarrow \|x_n - x^\star\| < \varepsilon, \\
(x_n) \to +\infty &\iff \forall A > 0,\ \exists N :\ n \geq N \Rightarrow x_n \geq A, \\
(x_n) \text{ Cauchy} &\iff \forall \varepsilon > 0,\ \exists N :\ m, n \geq N \Rightarrow \|x_n - x_m\| < \varepsilon.
\end{aligned}
$$

- In finite dimensions, convergence $\Leftrightarrow$ Cauchy because $(\mathbb{R}^d, \|\cdot\|)$ is complete.
- Coordinate-wise limit agreement: $(x_n) \to x^\star$ iff each coordinate sequence converges.
- Any norm on $\mathbb{R}^d$ yields the same notion of convergence thanks to norm equivalence.
- Divergence to $-\infty$ mirrors the $+\infty$ definition with the inequality reversed.

Sequences — Properties and Intuition

- Monotone and bounded real sequences converge; limit is the supremum/infimum.
- Algebra of limits (finite d): sums, scalar multiples, inner products, and norms of convergent sequences remain convergent.
- Equivalent norms $\|\cdot\|_a \sim \|\cdot\|_b$ give uniform bounds $\alpha\|x\|_a \le \|x\|_b \le \beta\|x\|_a$, so limits do not depend on the chosen norm.
- ε–N view: choose N so the tail of the sequence sits inside a shrinking ball $B(x^\star, \varepsilon)$; geometric intuition is "once inside, stays inside".

Sequences — Worked ε–N Examples

$(-1)^n/n \to 0$

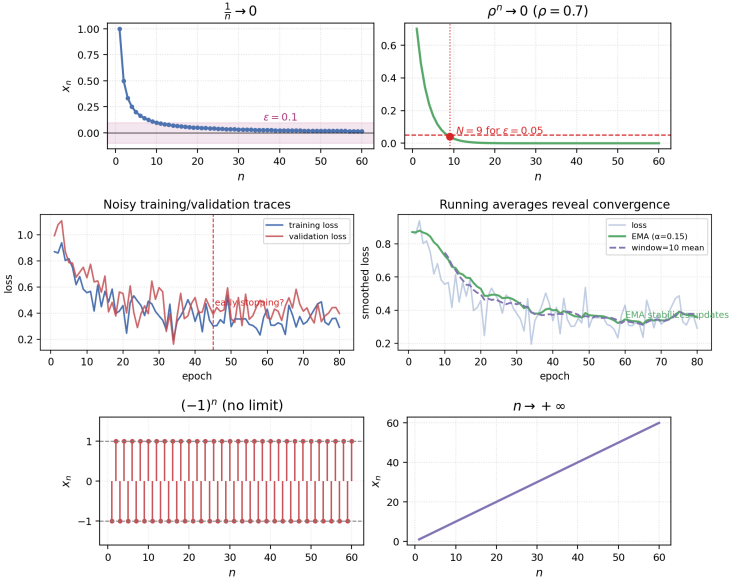
- Pick $N = \left\lceil \frac{1}{\varepsilon} \right\rceil$ so that $n \geq N \Rightarrow |(-1)^n/n| \leq 1/N < \varepsilon$.
- Numerical check: $\varepsilon = 0.05 \Rightarrow N = 20$; from then on every term sits in $(-0.05, 0.05)$.

$\rho^n \to 0$ for $0 < \rho < 1$

- Solve $\rho^N < \varepsilon$ using $\rho^n = e^{n \log \rho}$ with $\log \rho < 0$ to get $N > \frac{\log \varepsilon}{\log \rho}$.
- Example: $\rho = 0.7$, $\varepsilon = 0.05$ gives $N = \left\lceil \frac{\log 0.05}{\log 0.7} \right\rceil = 23$; tail iterates remain inside $B(0, \varepsilon)$.

Takeaway: Constructing explicit $N(\varepsilon)$ witnesses convergence and surfaces decay rates.

Sequences — Visualization



Sequences — ML Application

- **Training diagnostics:** Track $\{f(x_n)\}$ or validation metrics to detect stagnation, divergence, or oscillatory regimes before overfitting.
- **Adaptive steps:** Running averages $\hat{g}_n = \alpha g_n + (1-\alpha)\hat{g}_{n-1}$ smooth stochastic gradients for Adam/RMSProp.
- **Stopping rules:** Use convergence of residuals $\|x_{n+1} - x_n\|$ or moving windows to decide when to halt iterative solvers.

Sequences — Pitfalls and Checks

- Oscillation does not preclude convergence: need shrinking amplitude (e.g. $(-1)^n/n$) rather than raw sign flips.
- Cauchy test is simpler than limit guessing for abstract sequences; divergence can follow from finding two subsequences with different limits.
- Norm choice in $\mathbb{R}^d$ cannot change convergence verdicts; if doubts arise, check coordinates.
- Divergence must be certified: monotone yet unbounded sequences head to $\pm\infty$, but bounded oscillations might fail to converge.

Asymptotic Orders — Definition and Context

Comparing magnitudes as $n \to \infty$ (or $x \to a$)

- $x_n = \mathcal{O}(y_n)$ if $\exists C > 0, N$ with $|x_n| \leq C|y_n|$ for $n \geq N$.
- $x_n = o(y_n)$ if $\forall \varepsilon > 0$, $\exists N$ with $|x_n| \leq \varepsilon |y_n|$ for $n \geq N$ (equivalently $\frac{x_n}{y_n} \to 0$).
- $x_n = \Theta(y_n)$ if $\exists c_1, c_2 > 0$ and N such that $c_1|y_n| \leq |x_n| \leq c_2|y_n|$ for $n \geq N$.

- Same definitions transfer to functions $f(x)$, $g(x)$ as $x \to a$ or $x \to \infty$.
- Captures leading-order terms in series, residuals, and computational costs.

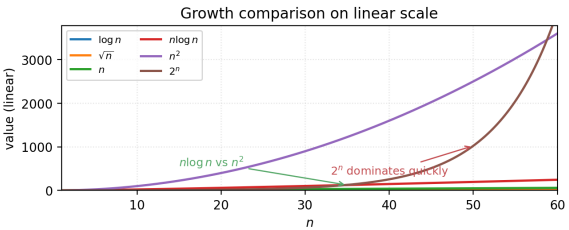
Asymptotic Orders — Properties and Intuition

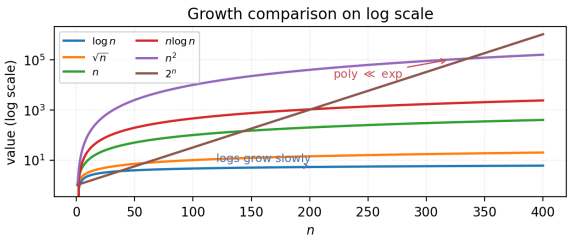
- Dominance ladder: $\log n \ll n^{\alpha} \ll \beta^{n}$ for any $\alpha > 0$, $\beta > 1$.
- Algebra: limits respect addition/multiplication (e.g. $o(y_n) + o(y_n) = o(y_n)$, $O(a_n)O(b_n) = O(a_n b_n)$).
- Transitivity: $x_n = O(y_n)$ and $y_n = O(z_n)$ imply $x_n = O(z_n)$ (similarly for o).
- Taylor language: $f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|)$; the remainder becomes negligible relative to $\|h\|$.

Asymptotic Orders — Worked Examples

- $\log n = o(n^{\alpha})$ for every $\alpha > 0$ since $\frac{\log n}{n^{\alpha}} \to 0$ (by L'Hôpital or domination).
- $n \log n$ and n^2: $\frac{n \log n}{n^2} = \frac{\log n}{n} \to 0$ so $n \log n = o(n^2)$, but $n^2 \neq o(n \log n)$.
- Harmonic sum $H_n = \sum_{k=1}^{n} \frac{1}{k}$ satisfies $\log(n+1) \leq H_n \leq 1 + \log n$, hence $H_n = \Theta(\log n)$.
- Exponential beats polynomial: $n^{\alpha} = o(\beta^n)$ for any $\alpha > 0$, $\beta > 1$ because $\frac{n^{\alpha}}{\beta^n} \to 0$.

Asymptotic Orders — Visualization





- Linear scale highlights absolute separation (exponential leaves the frame quickly).
- Log scale linearises multiplicative factors and exposes crossover points between polynomials.

Asymptotic Orders — ML Application

- **Algorithmic complexity:** SGD iterations cost $\mathcal{O}(nd)$ per pass; quasi-Newton updates can jump to $\mathcal{O}(nd^2)$.
- **Convergence rates:** Linear convergence $\|x_n - x^\star\| = \mathcal{O}(\rho^n)$ vs. sublinear $\mathcal{O}(1/n)$ informs optimizer selection.
- **Variance reduction:** Step-size schedules with $\sum \eta_t = \infty$ yet $\sum \eta_t^2 < \infty$ ensure stochastic approximations remain Cauchy.

Asymptotic Orders — Pitfalls and Checks

- Big-$\mathcal{O}$ hides constants: $\mathcal{O}(n)$ could be $1000n$; compare real coefficients when judging practicality.
- Little-o requires the ratio to vanish; checking $\lim \frac{x_n}{y_n}$ is the fastest route.
- Two-sided Θ demands both upper and lower bounds—missing either breaks equivalence.
- When studying vector sequences, pick any norm for estimates, but remember results must hold uniformly for large n.

This Section

Limits and Continuity — Definition and Context

Definition (ε–η in $\mathbb{R}$)

For $f : \mathbb{R} \to \mathbb{R}$, $\lim_{x \to a} f(x) = \ell$ if $\forall\, \varepsilon > 0\, \exists\, \eta > 0 :\ |x - a| < \eta \Rightarrow |f(x) - \ell| < \varepsilon$. Continuity at a means $\lim_{x \to a} f(x) = f(a)$.

- Plain-language: we can make $f(x)$ as close as we like to ℓ by taking x close enough to a. Here ε is the tolerated output error, and η is how close we need to be in input.
- One-sided limits: $x \to a^-$ (from the left) and $x \to a^+$ (from the right) must both exist and be equal for a two-sided limit.
- Extension to $\mathbb{R}^d$: replace $|\cdot|$ by a distance $\|\cdot\|$; the idea is the same (inputs near $a \Rightarrow$ outputs near ℓ).
- Sequential viewpoint: if $x_n \to a$ then, for continuous f, $f(x_n) \to f(a)$. This turns limit questions into sequence questions.

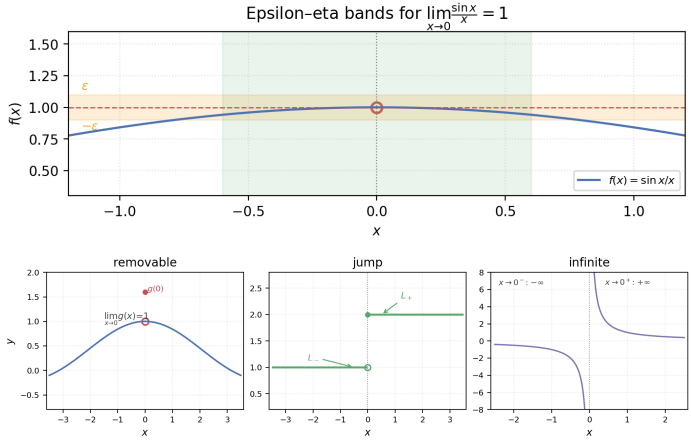
Limits and Continuity — Properties and Intuition

- Algebra (quick rules): add/subtract/multiply limits termwise; divide only if the denominator limit is nonzero.
- Composition: if $g \to b$ and f is continuous at b, then $(f \circ g) \to f(b)$.
- Squeeze (diagnostic): if $g(x) \leq f(x) \leq h(x)$ for x near a, and $\lim g = \lim h = \ell$, then $\lim f = \ell$.
- Intuition: think "bands and zooming" — ε is an output band; pick η so the x-window maps into that band. For piecewise f, check left and right behaviour separately.
- Quick recipe to try first: (1) substitute $x = a$; (2) if you get $0/0$, simplify (factor/rationalise) or try squeeze; (3) check one-sided limits if piecewise.

Limits and Continuity — Worked Example

- Example 1 (squeeze): $\lim_{x \to 0} \dfrac{\sin x}{x} = 1$. Steps: (i) for $x \neq 0$, $\cos x \leq \frac{\sin x}{x} \leq 1$; (ii) take limits: $\cos x \to 1$ and $1 \to 1$; (iii) conclude middle tends to 1.
- Example 2 (jump): $h(x) = \begin{cases} 1, & x < 0 \\ 2, & x \geq 0 \end{cases}$. Compute $\lim_{x \to 0^-} h = 1$, $\lim_{x \to 0^+} h = 2$; different limits $\Rightarrow$ no two-sided limit.
- Example 3 (removable): $g(x) = \frac{\sin x}{x}$ has a hole at 0. Define $g(0) = 1$ to make g continuous at 0.

Limits and Continuity — Visualization



- Top: choose η so that $x \in (a-\eta, a+\eta)$ keeps $f(x)$ in the ε-tube around ℓ.
- Bottom: compare left/right behaviour — removable (fix a point), jump (no fix), infinite (blows up).

Limits and Continuity — ML Application

- **Continuous activations:** small feature changes $\Rightarrow$ small output changes, which stabilises training; gradients exist almost everywhere (ReLU, GELU).
- **Loss stability near minima:** continuity ensures that near a good solution, tiny parameter nudges do not spike the loss.
- **Preprocessing:** patch removable discontinuities (e.g., divide-by-zero, sentinel values) so the model does not see artificial jumps.

Limits and Continuity — Pitfalls/Checks

- $0/0$ is *indeterminate*, not 0 — simplify or use squeeze; $c/0$ with $c \neq 0$ indicates blow-up (no finite limit).
- Two-sided limit exists only if left/right limits both exist and are equal.
- Continuity does not imply differentiability (corners like $|x|$); differentiability does imply continuity.
- For piecewise f, always report where the formula changes and check each side.

Derivative and First-Order Taylor — Definition and Context

Definition

The derivative at x is the limit of the difference quotient $f'(x) = \lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h}$. Differentiability implies continuity, and the first-order Taylor model

$$f(x+h) = f(x) + hf'(x) + o(|h|)$$

has a remainder that becomes negligible relative to $|h|$.

- Plain-language: $f'(x)$ is the instantaneous rate of change (slope) at x; units match "output per input".
- Linear model near x_0: $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ — a straight-line zoom of the curve.

Derivative and First-Order Taylor — Properties and Intuition

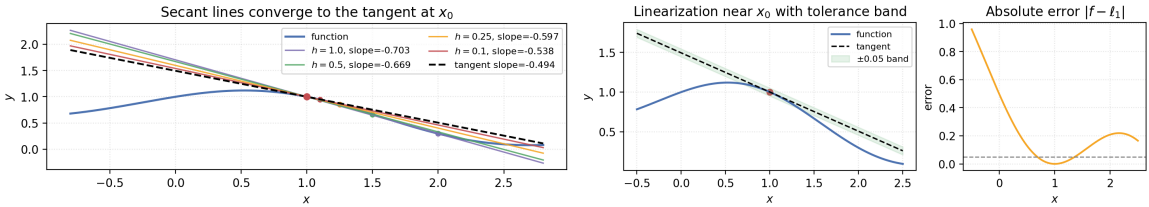
- **Geometric slope:** the secant slope $\dfrac{f(x_0+h)-f(x_0)}{h}$ approaches the tangent slope as $h \to 0$.
- **Units and sign:** $f'(x)$ measures "output per input"; positive means increasing, negative means decreasing.
- **Rules (statements):** linearity, product/quotient (where defined), and chain rule for compositions.
- **Local linearity:** differentiable functions look straight when you zoom in; linearization error grows with $|h|$.
- **Quick recipe:** (1) pick x_0; (2) compute $f(x_0)$ and $f'(x_0)$; (3) approximate with $f(x_0)+f'(x_0)(x-x_0)$; (4) check the approximation error for your h.

Derivative and First-Order Taylor — Worked Example

- **Step 1 — Define:** $f(x) = e^{-0.2x^2} + 0.2\sin(2x)$.
- **Step 2 — Differentiate:** $f'(x) = (-0.4x)e^{-0.2x^2} + 0.4\cos(2x)$ (sum and chain rules).
- **Step 3 — Evaluate at $x_0 = 1$:** $f(1) \approx 1.0006$, $f'(1) \approx -0.494$.
- **Step 4 — Linear model:** $\ell_1(x) = f(1) + f'(1)(x-1) \approx 1.0006 - 0.494\,(x-1)$.
- **Step 5 — Check numerically:** $f(1 \pm 0.1)$ and $\ell_1(1 \pm 0.1)$ are close (small error near x_0); error grows for larger $|h|$.
- **Contrast (non-differentiable):** $|x|$ is continuous at 0 but not differentiable — left slope -1, right slope $+1$.

x	$f(x)$	$\ell_1(x)$	$\lvert f(x) - \ell_1(x)\rvert$
0.9	1.04521	1.04999	0.00477
1.1	0.94676	0.95120	0.00444

Derivative and First-Order Taylor — Visualization



- **Left:** secant lines (different h) converge to the dashed tangent at x_0; their slopes approach $f'(x_0)$.
- **Right:** absolute error $|f - \ell_1|$ stays below a tolerance only near x_0 — visualising the "local" range where the linear model is reliable.
- **Takeaway:** closer to x_0 means smaller $|h|$ and smaller error; farther away increases curvature effects.

Derivative and First-Order Taylor — ML Application

- **Monotone decrease (small steps):** using the linear model, $f(x - \eta f'(x)) \approx f(x) - \eta(f'(x))^2$ — tiny steps reduce f when $\eta > 0$.
- **Learning-rate intuition:** larger $|f'(x)|$ suggests smaller η to stay in the local linear regime and avoid overshoot.
- **Sensitivity/feature effects:** $f'(x_0)$ estimates how small input changes move the output — a local feature effect.
- **Gradient checking:** compare finite-difference slopes to autodiff to detect implementation bugs.

Derivative and First-Order Taylor — Pitfalls/Checks

- Differentiable $\Rightarrow$ continuous; the converse fails when the tangent "breaks" (e.g., $|x|$ at 0).
- Linearizations are local — expect larger errors as you move away from x_0.
- Little-o vs Big-$\mathcal{O}$: $r(h) = o(|h|)$ means $\frac{r(h)}{|h|} \to 0$; $r(h) = \mathcal{O}(|h|)$ only means bounded by a constant multiple.
- One-sided derivatives must agree to have a derivative at a kink; otherwise the derivative is undefined there.

This Section

Gradient — Definition

Definition

For $f : \mathbb{R}^n \to \mathbb{R}$, the partial derivative is

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \to 0} \frac{f(x + h e_i) - f(x)}{h},$$

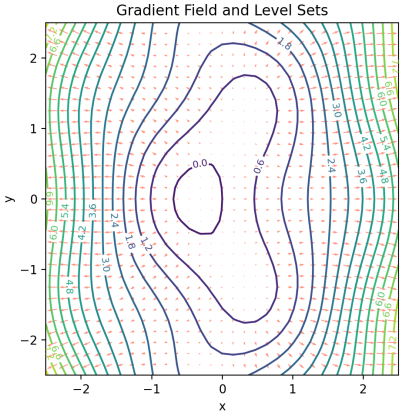
and the gradient is $\nabla f(x) = (\partial f / \partial x_1, \ldots, \partial f / \partial x_n)^\top$.

- If f is differentiable, $f(x + h) = f(x) + \nabla f(x) \cdot h + o(\|h\|)$.
- Directional derivative: $D_u f(x) = \nabla f(x) \cdot u$ for unit vector u.

Gradient — Intuition

- $\nabla f(x)$ generalizes slope to n dimensions: it points toward steepest ascent.
- Connect to Lesson 3 inner products: $D_u f(x)$ is the projection of $\nabla f(x)$ onto u.
- Example: for $f(x, y) = x^2 + \frac{1}{4}y^2 + 0.5\sin(2x)\cos(2y)$, compute partials analytically.

Gradient — Visualization



- Contours show level sets; arrows indicate gradient direction and magnitude.
- Observe orthogonality between gradient vectors and level curves.

Gradient — ML Application

- Parameter updates: $w_{k+1} = w_k - \eta \nabla f(w_k)$ powers most learning algorithms.
- Feature saliency: gradient magnitude highlights influential inputs (e.g., saliency maps).
- Constraint handling: projected gradients use directional derivatives to stay within feasible sets.

Gradient — Worked Example

- For $f(x,y) = x^2 + \frac{1}{4}y^2 + 0.5\sin(2x)\cos(2y)$,

$$\nabla f(x,y) = \big(2x + \cos(2x)\cos(2y),\ \tfrac{1}{2}y - \sin(2x)\sin(2y)\big).$$

- Directional derivative: for unit u, $D_u f(x,y) = \nabla f(x,y) \cdot u$; steepest ascent occurs when u aligns with $\nabla f(x,y)$.
- Example point $(x,y) = (0.6, -0.4)$: plug-in to obtain $\nabla f(0.6,-0.4)$ and verify $D_u f = \nabla f \cdot u$.

Jacobian — Definition

Definition

For $F : \mathbb{R}^n \to \mathbb{R}^m$, F is differentiable at x if there exists a linear map $DF(x)$ such that

$$F(x + h) = F(x) + DF(x)h + o(\|h\|).$$

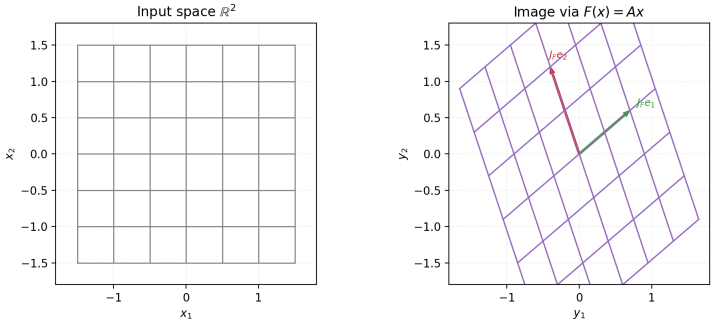
The Jacobian matrix $J_F(x)$ has entries $[J_F(x)]_{ij} = \partial F_i / \partial x_j$ and represents $DF(x)$.

- For scalar-valued f, $DF(x) = \nabla f(x)^\top$.
- The differential transports basis vectors via columns of $J_F(x)$.

Jacobian — Intuition

- Think of $DF(x)$ as the best linear approximation: it maps small displacements to output changes.
- Columns of $J_F(x)$ show how each input direction influences all outputs.
- Example: $F(x, y) = (0.7x - 0.4y,\ 0.6x + 1.2y)$ reuses linear transformations from earlier lessons.

Jacobian — Visualization



- Left: original grid; Right: image under F shows shear + scaling captured by J_F.
- Observe how the Jacobian maps basis vectors e_1, e_2 to the columns of the matrix.

Jacobian — ML Application

- Normalizing flows: Jacobians determine volume change and log-determinant terms.
- Sensitivity of multi-output models (e.g., multi-class logits) relies on $J_F(x)$ structure.
- Auto-diff frameworks compute Jacobians to propagate gradients efficiently.

Jacobian — Shapes and Linearization

- For $F : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian $J_F(x)$ has shape $m \times n$; columns describe images of basis directions.
- Linear approximation: $F(x + h) \approx F(x) + J_F(x)\, h$ for small h.
- Quick check: verify units/dimensions match in any calculation before multiplying matrices.

Chain Rule — Definition

Definition

If $F : \mathbb{R}^m \to \mathbb{R}^p$ and $G : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at $G(x)$ and x, then the composition $F \circ G$ is differentiable with

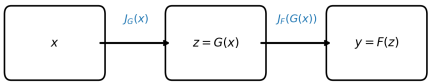
$$D(F \circ G)(x) = DF(G(x)) \circ DG(x), \quad J_{F \circ G}(x) = J_F(G(x))\, J_G(x).$$

- Scalar chain rule is the $p = n = 1$ case: $(f \circ g)'(x) = f'(g(x))g'(x)$.

Chain Rule — Intuition

- Interpret as successive linear approximations: first move through G, then through F.
- The Jacobian product mirrors how outputs feed into subsequent layers.
- Example: $g(x) = (\sin x_1 + x_2^2, x_1 e^{x_2})$, $f(z) = z_1^2 + 3z_2$; compute derivatives step by step.

Chain Rule — Visualization



- Computational graph highlights the flow of information and associated Jacobians.
- Edge labels remind us where each derivative factor originates.

Chain Rule — ML Application

- Backpropagation: gradients in neural networks multiply Jacobians layer by layer.
- Sensitivity analysis: chain rule links perturbations in inputs to outputs through intermediate representations.
- Calibration: understanding derivative chaining helps detect vanishing/exploding gradients.

Chain Rule — Worked Example and Dimensions

- Let $g(x) = (\sin x_1 + x_2^2,\ x_1 e^{x_2}) \in \mathbb{R}^2$ and $f(z) = z_1^2 + 3z_2 \in \mathbb{R}$.
- $J_g(x) \in \mathbb{R}^{2\times 2}$, $\nabla f(z)^\top \in \mathbb{R}^{1\times 2}$, so $J_{f\circ g}(x) = \nabla f(g(x))^\top J_g(x) \in \mathbb{R}^{1\times 2}$.
- Explicitly: $J_g(x) = \begin{bmatrix} \cos x_1 & 2x_2 \\ e^{x_2} & x_1 e^{x_2} \end{bmatrix}$ and $\nabla f(z) = \begin{bmatrix} 2z_1 \\ 3 \end{bmatrix}$.
- Multiply to obtain the gradient of $f \circ g$ with respect to x and verify shapes are compatible.

Key Takeaways

- Operator vs spectral norm: maximal stretch and principal directions.
- SVD as rank-one sum; best rank-k approximation via top singular values.
- Sequence convergence and growth comparison with $\mathcal{O}$, o, Θ.
- First-order linearization in one/many variables; $o(\cdot)$ remainder interpretation.
- Gradients, Jacobians, and the chain rule power optimization/backprop.

Practice

- Prove: if $a_n = \mathcal{O}(1/n)$ and $b_n = \mathcal{O}(1/n)$, then $a_n + b_n = \mathcal{O}(1/n)$; is $a_n b_n = \mathcal{O}(1/n)$?
- Compute the Jacobian of $F(x, y) = (xe^y,\ \sin(x + y))$ at $(0, 0)$ and use it to linearize F near $(0, 0)$.
- Given $A \in \mathbb{R}^{m \times n}$ with SVD $A = U\Sigma V^\top$, write the best rank-1 approximation and its Frobenius error.
- For $f(x) = |x|$, classify continuity and differentiability at 0; repeat for $f(x) = x\,|x|$.

References

- See course references in `syllabus.md` and the instructor notes (`PRE2-CLASS6-notes.pdf`).
- Companion notes and code snippets: `lesson-06.md` and `slides/lesson-06/py/`.