Plan général, organisation

Partie 1: Généralités
Partie 2: Méthodes directes
Partie 3: Méthodes itératives
Partie 4: Problèmes aux valeurs et vecteurs propres
- Séance 5a: Généralités, Puissances itérées, puissances inverses
- Séance 6a: **Itérations orthogonales et algorithme QR**

Partie 5: Systèmes linéaires mal conditionnés

Computation of matrix spectra

General approach: use invariance of eigenvalues under similarity transformations:

- Let $A \in \mathbb{K}^{n\times n}$, let $X \in \mathbb{K}^{n\times n}$ invertible, set $T := X^{-1}AX$.
  Then $P_A(\lambda) = P_T(\lambda)$ (same eigenvalues and multiplicities).

Matrix spectra computations: find similarity decomposition $A = XTX^{-1}$ where eigenvalues of T "easy" to compute.

- Factorizations for direct methods (e.g. LU, Cholesky) not appropriate
  For instance: $A = LU$ reveals eigenvalues of L, U, but no connection to eigenvalues of A.
- In fact, we know LU etc cannot work (since direct eigenvalue algorithms inpossible)

Better starting point: the Schur decomposition of A:

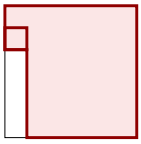
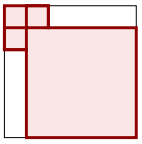
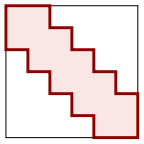
- Any $A \in \mathbb{K}^{n\times n}$ has a *Schur decomposition* $A = QTQ^{\mathrm{H}}$ ($Q \in \mathbb{K}^{n\times n}$ unitary, $T \in \mathbb{K}^{n\times n}$ upper triangular)
- Schur decomposition is a similarity transformation of A (so $P_A(\lambda) = P_T(\lambda)$)
- Since T triangular, $\mathrm{diag}(T)$ holds the eigenvalues of A.
- Even if $A \in \mathbb{R}^{n\times n}$, $Q, T \in \mathbb{C}^{n\times n}$ in general (as real matrices may have complex eigenvalues).
- If A is Hermitian, T is real and diagonal.

Computation of matrix spectra: a possible outline

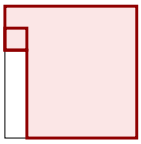
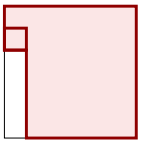
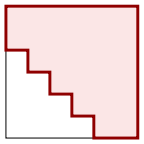
Ideal general approach: compute eigenvalues by finding Schur decomposition of A.

Towards finding $A = QTQ^{\mathrm{H}}$: introduce zeros in lower triangle of A (again!), but note carefully:

- Let F unitary; assume FA puts zeros in whole 1st column under diagonal.
  Then, similarity needs forming FAF^{H}, but right multiplication undoes zeroing-out (try with Householder reflector)
- Remedy: use instead (e.g. Householder) transformations such that (for Hermitian A)

$F_1 A =$  $F_1 A F_1^{\mathrm{H}} =$  $\ldots FAF^{\mathrm{H}} =$ 

  A Hermitian $\Longrightarrow$ FAF^{H} tridiagonal
- For non-Hermitian A, can reach FAF^{H} upper Hessenberg:

$F_1 A =$  $F_1 A F_1^{\mathrm{H}} =$  $\ldots FAF^{\mathrm{H}} =$ 

- This is not yet the Schur decomposition (but we get closer)
- Reduction to tridiagonal/Hessenberg takes fixed computational work (direct step)
  Then, the rest (e.g. tridiagonal/Hessenberg to Schur) is iterative
- Finding $A = QTQ^{\mathrm{H}}$ may need complex arithmetic even if A real (but then we prefer real arithmetic).

Computation of matrix spectra: orthogonal iterations

- Assume $|\lambda_1| > |\lambda_2| > \dots |\lambda_n|$, with corresponding eigenvectors $Q_1, \dots, q_n$.
- Starting idea: apply power iterations to a set of p vectors $X_p = [x_1, \dots, x_p] \in \mathbb{K}^{n\times p}$:

  $$X_p^{(0)} = X,\ X_p^{(1)} = AX_p^{(0)}, \dots\ X_p^{(k)} = AX_p^{(k-1)} \dots$$

  Then $E_p^{(k)} := \text{span}\big(x_1^{(k)}, \dots, x_p^{(k)}\big) \ \to\ E_p := \text{span}\big(q_1, \dots, q_p\big)$
- Conceivably: (a) run k iterations (until convergence of $\text{span}\big(x_1^{(k)}, \dots, x_p^{(k)}\big)$),
  (b) diagonalize smaller matrix $A_p^{(k)} := (X^{(k)})^{\mathsf{H}} A X^{(k)} \in \mathbb{K}^{p\times p}$.
- However, vectors of $X_p^{(k)}$ increasingly collinear
- Remedy: orthogonalization (again!), i.e. find next iterate $X_p^{(k)}$ via

  $$X_p^{(k)} R^{(k)} = AX_p^{(k-1)}$$ use QR decomposition on $AX_p^{(k-1)}$

Algorithm 13 Orthogonal iterations

1: $A \in \mathbb{K}^{n\times n}$ Hermitian, $X_p^{(0)} = [x_1^{(0)}, \dots, x_p^{(0)}] \in \mathbb{K}^{n\times p}$ with orthonormal columns (initialization)
2: **for** $k = 1, 2, \dots$ **do**
3: $\quad Z^{(k)} = AX_p^{(k-1)}$ (apply A)
4: $\quad X_p^{(k)} R_p^{(k)} = Z^{(k)}$ (compute reduced QR decomposition of $Z^{(k)} \in \mathbb{K}^{n\times p}$)
5: $\quad$ **Stop** if convergence, set $\lambda_i = (x_i^{(k)})^{\mathsf{H}} A q_i^{(k)}$, $q_i = x_i^{(k)}$
6: **end for**

How and why orthogonal iterations work

Focus on case $p = 2$ (recall $p = 1$ is standard power iteration):

$$AX_p^{(k-1)} = X_p^{(k)} R^{(k)} \qquad \text{with} \quad R^{(k)} = \begin{bmatrix} r_{11}^{(k)} & r_{12}^{(k)} \\ 0 & r_{22}^{(k)} \end{bmatrix} \qquad \text{i.e.} \quad \begin{cases} \text{(a)} & r_{11}^{(k)} x_1^{(k)} = A x_1^{(k-1)}, \\ \text{(b)} & r_{12}^{(k)} x_1^{(k)} + r_{22}^{(k)} x_2^{(k)} = A x_2^{(k-1)} \end{cases}$$

(a) $x_1^{(0)}, x_1^{(1)}, x_1^{(2)} \dots x_1^{(k)} \dots$ generated by power iterations, hence $x_1^{(k)} \to q_1$.

(b) Write $x_1^{(k)} = q_1 + \varepsilon_1^{(k)}$ with $\|\varepsilon_1^{(k)}\| \to 0$, then set

$$\begin{aligned} \widehat{A} &= (I - q_1 q_1^{\mathsf{H}})^{\mathsf{H}} A (I - q_1 q_1^{\mathsf{H}}) \\ &= A - \lambda_1 q_1 q_1^{\mathsf{H}} \qquad \Longrightarrow \ \widehat{A} q_1 = 0, \ \widehat{A} q_i = A q_i \ (i \geq 2) \end{aligned}$$

Consequently:

$$\begin{aligned} \widehat{A} x_2^{(k-1)} &= A x_2^{(k-1)} - \lambda_1 (q_1^{\mathsf{H}} x_2^{(k-1)}) q_1 \\ &= r_{12}^{(k)} x_1^{(k)} + r_{22}^{(k)} x_2^{(k)} - \lambda_1 (q_1^{\mathsf{H}} x_2^{(k-1)}) q_1 \end{aligned} \qquad \begin{aligned} r_{12}^{(k)} &= (x_1^{(k)})^{\mathsf{H}} A x_2^{(k-1)} \\ &= (q_1 + \varepsilon_1^{(k)})^{\mathsf{H}} A x_2^{(k-1)} \\ &= \lambda_1 (q_1^{\mathsf{H}} x_2^{(k-1)}) + (\varepsilon_1^{(k)})^{\mathsf{H}} A x_2^{(k-1)} \end{aligned}$$

$$\begin{aligned} \widehat{A} x_2^{(k-1)} &= r_{22}^{(k)} x_2^{(k)} + (q_1^{\mathsf{H}} x_2^{(k-1)}) \varepsilon_1^{(k)} + ((\varepsilon_1^{(k)})^{\mathsf{H}} A x_2^{(k-1)}) x_1^{(k)} \\ &= r_{22}^{(k)} x_2^{(k)} + (\varepsilon_1^{(k)}) \end{aligned}$$

$x_2^{(0)}, x_2^{(1)}, x_2^{(2)} \dots x_2^{(k)} \dots$ generated by power iterations for $\widehat{A}$.

Computation of matrix spectra: orthogonal iterations

Convergence of orthogonal iterations

Let $A \in \mathbb{K}^{n\times n}$ Hermitian with $|\lambda_1| > |\lambda_2| > \ldots |\lambda_p|$.
Assume all leading submatrices $(Q_p^{\mathsf{H}} X_p^{(0)})_{1:q,1:q}$ $(1 \leq q \leq p)$ of $Q_p^{\mathsf{H}} X_p^{(0)} \in \mathbb{K}^{p\times p}$ are nonsingular.
Let $X_p^{(k)} = [x_1^{(0)}, \ldots, x_p^{(0)}]$: set of orthonormal vectors produced by k orthogonal iterations. Then:

$$\|x_i^{(k)} \pm q_i\| = O(C^k), \qquad \text{with} \quad C := \max_{1\leq j\leq p} |\lambda_{j+1}|/|\lambda_j| < 1.$$

Computation of matrix spectra: QR iterations

- Adapt orthogonal iterations to complete spectrum of $A \in \mathbb{K}^{n\times n}$ (Hermitian);
- Remove restriction $|\lambda_1| > |\lambda_2| > \ldots |\lambda_p| > \ldots$

Focus on $T^{(k)} := X^{(k)\mathrm{H}} A X^{(k)}$ (note $T^{(k)} \to \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$):

$$T^{(k-1)} = X^{(k-1)\mathrm{H}} A X^{(k-1)} = X^{(k-1)\mathrm{H}} X^{(k)} R^{(k)} \qquad \text{(QR decomposition of } AX^{(k-1)}\text{)}$$
$$T^{(k)} = X^{(k)\mathrm{H}} A X^{(k)} = X^{(k)\mathrm{H}} A X^{(k-1)} X^{(k-1)\mathrm{H}} X^{(k)} = X^{(k)\mathrm{H}} X^{(k)} R^{(k)} X^{(k-1)\mathrm{H}} X^{(k)}$$
$$= R^{(k)} X^{(k-1)\mathrm{H}} X^{(k)},$$

Reformulate:

$$\text{(a) } T^{(k-1)} = Q^{(k)} R^{(k)}, \qquad \text{(b) } T^{(k)} = R^{(k)} Q^{(k)}, \qquad Q^{(k)} := X^{(k-1)\mathrm{H}} X^{(k)} \text{ unitary.}$$

Algorithm 14 Basic QR iterations

1: $A \in \mathbb{K}^{n\times n}$ Hermitian, $T^{(0)} = A$ (initialization)
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad Q^{(k)} R^{(k)} = T^{(k-1)}$ (compute QR decomposition of $T^{(k-1)} \in \mathbb{K}^{n\times p}$)
4: $\quad T^{(k)} = R^{(k)} Q^{(k)}$ (update $T^{(k)}$)
5: $\quad$ **Stop** if convergence, $\mathrm{diag}(T^{(k)})$ contains the eigenvalues of A
6: **end for**

Computation of matrix spectra: shortcomings of basic QR iterations

Basic QR iterations are workable (in particular backward stable) but lack efficiency:

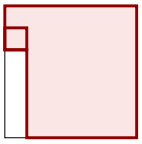
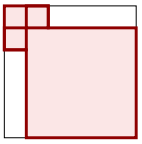
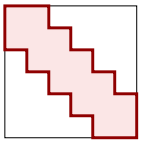
- Each QR factorization costs $O(n^3)$ operations (see lecture 2)
- Expect $O(n)$ QR iterations needed, so $O(n^4)$ computing work overall.
- Rate of convergence of eigenvalues depends on their distribution
- Convergence may fail if $|\lambda_j| = |\lambda_{j+1}|$ for some j.

Two major improvements adresss these issues:

- First put A in *tridiagonal form* (needs $O(n^3)$ work).
  QR decompositions of a tridiagonal matrix then take $O(n^2)$ work $\implies$ $O(n^3)$ overall work.
- Accelerate convergence using a shifted form of QR algorithm

Computation of matrix spectra: reduction to tridiagonal form

Method 1: symmetric application of Householder reflectors

$F_1 A =$  $F_1 A F_1^{\mathrm{H}} =$  $\ldots F A F^{\mathrm{H}} =$ 

- Requires storage of A
- Proved stability

Method 2: Lanczos orthogonalization iterations

- Recall Arnoldi iterations (used for GMRES):

  $\boxed{A = QHQ^{\mathrm{H}}}$ Q unitary, H upper Hessenberg

  Here, A Hermitian $\implies$ H tridiagonal.
- Specialize Arnoldi iterations to A Hermitian (so H tridiagonal) $\rightarrow$ Lanczos iterations
- Only requires matrix-vector products $q \mapsto Aq$, i.e. suitable for large sparse matrices

Lanczos iterations

Column-by-column enforcement of equality (starting with q_1 arbitrary unit vector)

$$A[q_1 \ \dots \ q_k] = [q_1 \ \dots \ q_k, q_{k+1}] \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \beta_{n-1} \\ 0 & \dots & 0 & \beta_{n-1} & \alpha_n \end{bmatrix}$$

column 1: $Aq_1 = \alpha_1 q_1 + \beta_1 q_2 \quad \Longrightarrow \alpha_1, \beta_1, q_2$
$(q_1^{\mathsf{H}} q_2 = 0, \ \|q_2\| = 1)$

column 2: $Aq_2 = \beta_1 q_1 + \alpha_2 q_2 + \beta_2 q_3 \quad \Longrightarrow \alpha_2, \beta_2, q_3$
$(q_1^{\mathsf{H}} q_3 = q_2^{\mathsf{H}} q_3 = 0, \ \|q_3\| = 1)$

column k: $Aq_k = \beta_{k-1} q_{k-1} + \alpha_k q_k + \beta_k q_{k+1} \quad \Longrightarrow \alpha_k, \beta_k, q_{k+1}$
$(q_{k-1}^{\mathsf{H}} q_{k+1} = q_k^{\mathsf{H}} q_{k+1} = 0, \ \|q_{k+1}\| = 1) \dots$

column n: $Aq_n = \beta_{n-k1} q_{n-1} + \alpha_n q_n \quad \Longrightarrow \alpha_n$

By induction: $q_k \in \mathrm{span}(q_1, Aq_1, \dots, A^{k-1}q_1)$ for $k = 1, 2, 3 \dots$
$\mathrm{span}(q_1, q_2, \dots, q_k) = \mathrm{span}(q_1, Aq_1, \dots, A^{k-1}q_1) = \mathcal{K}_k(A, b)$

Inverse-iteration interpretation of the QR algorithm

- Generic orthogonal iteration at root of basic QR algorithm:
  $X^{(k)}R^{(k)} = AX^{(k-1)\mathrm{H}} \implies A = X^{(k)}R^{(k)}X^{(k-1)\mathrm{H}}$.
- Evaluate $A^{-1} = (A^{-1})^{\mathrm{H}}$ (since A Hermitian):
  $A^{-1} = X^{(k-1)}(R^{(k)})^{-1}X^{(k)\mathrm{H}} = X^{(k)}(R^{(k)})^{-\mathrm{H}}X^{(k-1)\mathrm{H}}$
  Rewrite using "flipped identity" P (properties: $P^2 = I$, $P[x_1, \ldots, x_n] = [x_n, \ldots, x_1]$):

$$\boxed{A^{-1} = \left(X^{(k)}P\right)\left(P(R^{(k)})^{-\mathrm{H}}P\right)\left(X^{(k-1)}P\right)^{\mathrm{H}},} \qquad P := \begin{bmatrix} 0 & \cdots & & 1 \\ \vdots & & \iddots & \\ & \iddots & & \vdots \\ 1 & & \cdots & 0 \end{bmatrix}$$

- Observe (i) $X^{(k-1)}P$, $X^{(k)}P$ unitary; (ii) $P(R^{(k)})^{-\mathrm{H}}P$ upper triangular.

Orthogonal iteration for A on $X^{(k)}$ equivalent to orthogonal iteration for A^{-1} on $X^{(k)}P$

In particular, 1st column of $X^{(k)}P$, i.e. $x_n^{(k)}$, undergoes inverse iteration (without shift).

Computation of matrix spectra: shifted QR algorithm

Inverse-iteration interpretation suggests using a shift $\mu^{(k)}$; main steps become

$$\text{(a) } T^{(k-1)} - \mu^{(k)} I = Q^{(k)} R^{(k)} \quad \text{and} \quad \text{(b) } T^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I.$$

How to (adaptively) choose shifts?

→ $\mu^{(k)} := t_{nn}^{(k-1)}$ natural choice (Rayleigh quotient for $q_n^{(k-1)}$), but known to fail on some "nice" matrices.

→ Wilkinson shift (eigenvalue of bottom rightmost 2×2 block of $T^{(k-1)}$ closest to $t_{nn}^{(k-1)}$)

Algorithm 15 Shifted QR iterations

1: $A \in \mathbb{K}^{n \times n}$ Hermitian (data)
2: $(Q^{(0)})^{\mathsf{H}} T^{(0)} Q^{(0)} = A$ (Tridiagonalization of A)
3: **for** $k = 1, 2, \ldots$ **do**
4: Choose $\mu^{(k)}$ (shift value, e.g. use the *Wilkinson shift*)
5: $Q^{(k)} R^{(k)} = T^{(k-1)} - \mu^{(k)} I$ (compute QR factorization of $T^{(k-1)} - \mu^{(k)} I \in \mathbb{K}^{n \times p}$)
6: $T^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I$ (update $T^{(k)}$)
7: If any off-diagonal entry $t_{j,j+1}^{(k)}$ is sufficiently small,

set $t_{j,j+1}^{(k)} = t_{j+1,j}^{(k)} = 0$ to obtain $T^{(k)} = \begin{bmatrix} T_1^{(k)} & 0 \\ 0 & T_2^{(k)} \end{bmatrix}$.

From now, apply the QR algorithm separately to $T_1^{(k)}$ and $T_2^{(k)}$ ("deflation")

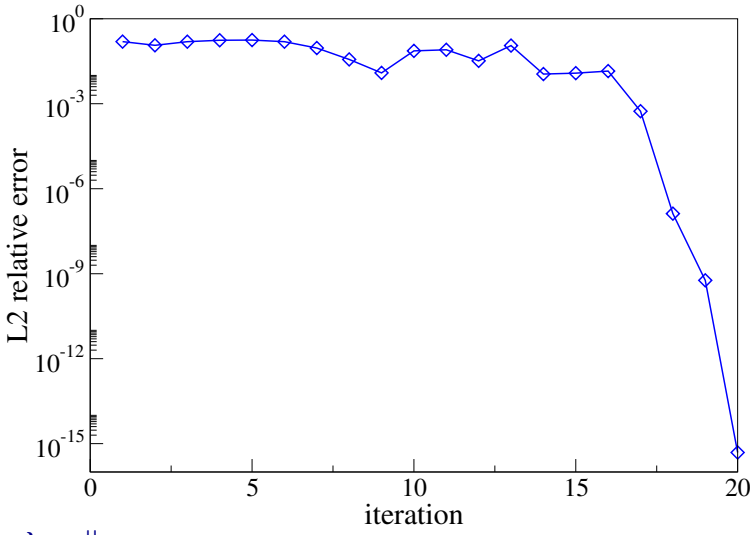
8: **Stop** if convergence, $\mathrm{diag}(T^{(k)})$ contains the eigenvalues of A
9: **end for**

Computation of matrix spectra: example

$A \in \mathbb{R}^{n \times n}$ a random real symmetric matrix.

Stopping criterion of QR iterations: $\dfrac{\| T^{(k+1)} - T^{(k)} \|}{\| T^{(k)} \|} \leq 10^{-12}$

n	iterations (deflation)	iterations (no deflation)	spectrum error
10	17	254	5.1954e-16
20	37	1275	1.0795e-15
50	96	22365	1.1e-15
100	185		8.6751e-16
200	370		1.6564e-15
500	880		1.5776e-15



Spectrum error: $e_\lambda = \dfrac{\| \lambda - \lambda_{QR} \|}{\| \lambda \|}$, $\lambda = \{ \lambda_1, \ldots, \lambda_n \}$

- Deflation (here not fully implemented) dramatically reduces iteration count
- Very high accuracy on whole spectrum achievable (note however: comparison spectrum also numerical $\leftarrow$ `eig(A)`)

Computation of matrix spectra: extension to unsymmetric problems

$\boxed{(A-\lambda I)x=0}$, A unsymmetric

- Set A in upper Hessenberg form (e.g. using Householder reflectors):

$$A = QHQ^{\mathrm{H}}, \qquad H = \begin{bmatrix} \times & \times & \times & \dots & \times \\ \times & \times & & & \times \\ 0 & \times & \times & & \times \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \times & \times \end{bmatrix}$$

  Spectra of A and H coincide.

- (shifted) QR iterations applicable to H:

  $T^{(0)} = H$, then (a) $T^{(k-1)} - \mu^{(k)} I = Q^{(k)} R^{(k)}$ and (b) $T^{(k)} = R^{(k)} Q^{(k)} + \mu^{(k)} I$.

- If $A \in \mathbb{C}^{n\times n}$, $T^{(k)} \to T$ upper triangular;
- If $A \in \mathbb{R}^{n\times n}$ and QR algorithm in real arithmetic, $T^{(k)} \to T$ "almost upper triangular" (2×2 diagonal blocks $\to$ pairs of conjugate complex eigenvalues);

Plan général

Partie 1: Généralités

Partie 2: Méthodes directes

Partie 3: Méthodes itératives

Partie 4: Problèmes aux valeurs et vecteurs propres

Partie 5: Systèmes linéaires mal conditionnés

Séance 6b: **Compression et approximation de systèmes mal conditionnés**

Séance 7a: Recherche de solutions parcimonieuses.

Ill-conditioned problems

- In some areas of applications, linear systems $Ax = b$ ($A \in \mathbb{K}^{m\times n}$, most often $m \geq n$) with "unpleasant" properties, e.g.
  - $\rightarrow$ A has, in theory, full column rank;
  - $\rightarrow$ However, A ill-conditioned with very fast decay of singular values, i.e. numerically rank-deficient:

    $$\boxed{\|A - A_r\| \ll \|A\|} \quad \text{for some rank-}r\text{ matrix } A_r,\ r \ll n$$

  - $\rightarrow$ imperfect data b (e.g. measurement errors)
- Such cases occur e.g. for
  - $\rightarrow$ Inverse and identification problems (infer "hidden" physical properties from indirect measurements)
  - $\rightarrow$ Image processing and image restoration
  - $\rightarrow$ Data analysis

- In what follows: least-squares solutions of $Ax = b$ ($A \in \mathbb{K}^{m\times n}$, $m \geq n$, $\mathrm{Rank}(A) = n$).

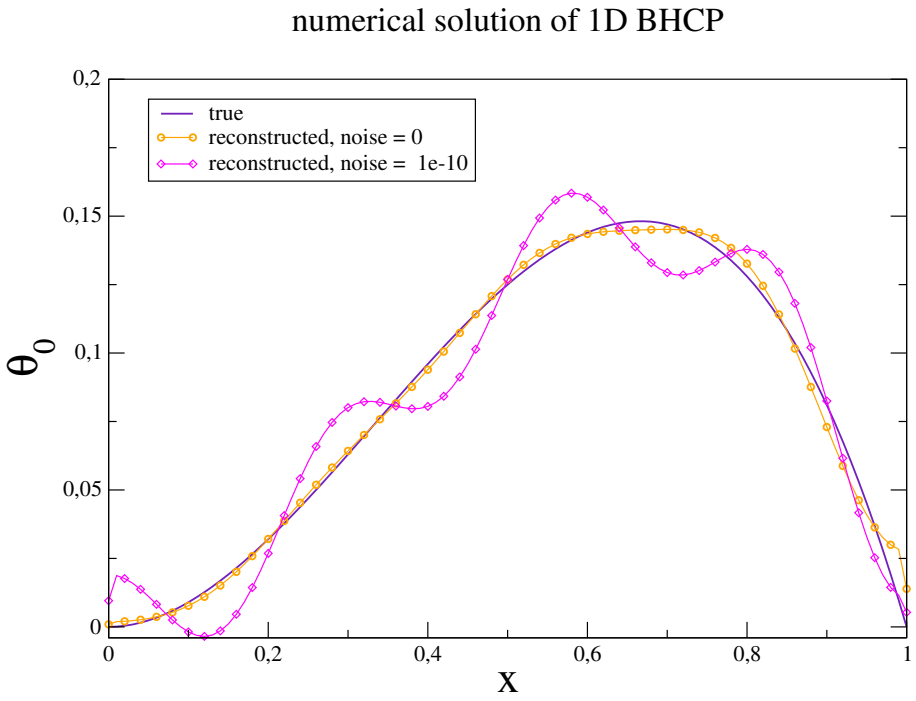
Recall matrix SVD (see lecture 2):

$$A = USV^{\mathrm{H}} = \sum_{i=1}^{n} \sigma_i u_i v_i^{\mathrm{H}} \qquad \begin{cases} U = [u_1, \ldots, u_n] \in \mathbb{K}^{m\times n} \\ V = [v_1, \ldots, v_n] \in \mathbb{K}^{n\times n} \\ S = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n} \end{cases}$$

Example: backward heat equation

Physical problem: find the temperature distribution in a system before thermal measurements are made (example: space shuttle re-entry).

$$\underbrace{\Theta(\cdot, T)}_{\text{measurement}} = \underbrace{\mathcal{A}(T)}_{\text{heat eq.}} \underbrace{\Theta(\cdot, 0)}_{\text{unknown}}$$



Example: backward heat equation

Physical problem: find the temperature distribution in a system before thermal measurements are made (example: space shuttle re-entry).

$$\underbrace{\Theta(\cdot, T)}_{\text{measurement}} = \underbrace{\mathcal{A}(T)}_{\text{heat eq.}} \underbrace{\Theta(\cdot, 0)}_{\text{unknown}}$$

$$\begin{aligned} \kappa \partial_{xx}\Theta - \partial_t \Theta &= 0 && (0 \le t \le T,\ 0 \le x \le \ell) \qquad \kappa := k/(\rho c) \\ \Theta(0,t) = \Theta(\ell,t) &= 0 && (0 \le t \le T) \\ \Theta(x,0) &= \Theta_0(x) && (0 \le x \le \ell) \end{aligned}$$

- General solution (Fourier series): $\Theta(x,t) = \sum_{n \ge 0} a_n \sin \frac{n\pi x}{\ell}\, e^{-(n\pi)^2 \kappa t/\ell^2}$
- Initial temperature: $\Theta(x,0) = \sum_{n \ge 0} a_n \sin \frac{n\pi x}{\ell}, \quad a_n = \frac{2}{\ell}\int_0^\ell \Theta_0(x) \sin \frac{n\pi x}{\ell}\, \mathrm{d}x$
- Final temperature: $\Theta(x,T) = \sum_{n \ge 0} b_n \sin \frac{n\pi x}{\ell}, \quad b_n = a_n \underbrace{e^{-(n\pi)^2 \kappa T/\ell^2}}_{\lambda_n}$

$$\{ b_0, b_1, b_2 \dots \}^\mathsf{T} = \operatorname{diag}[\lambda_0, \lambda_1, \lambda_2 \dots] \{ a_0, a_1, a_2 \dots \}^\mathsf{T}$$

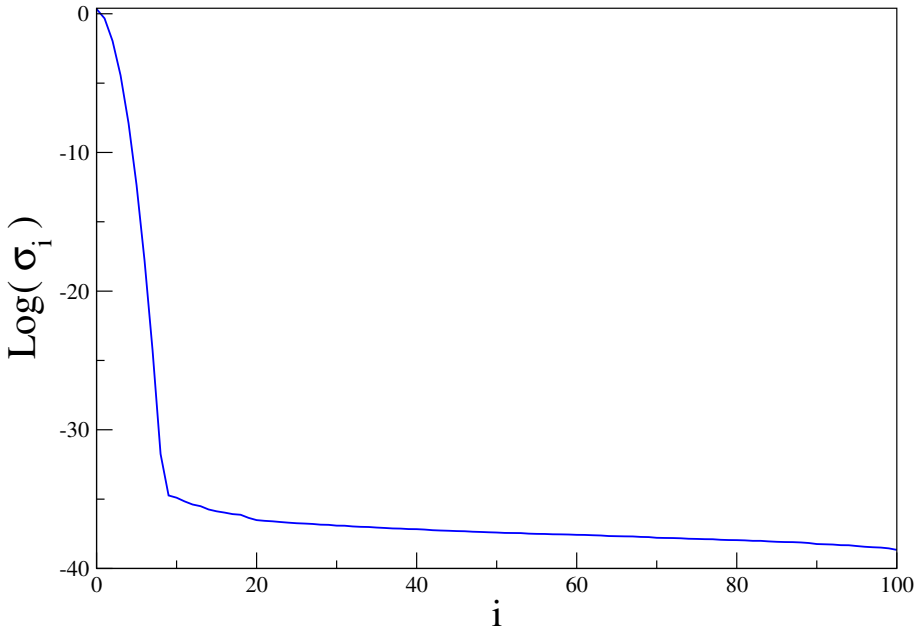
- Reconstruction de $\Theta(\cdot, 0)$ given $\Theta(\cdot, T)$ (explicit inversion):

$$\Theta(x,0) = \sum_{n \ge 0} \lambda_n^{-1} b_n \sin \frac{n\pi x}{\ell} \qquad \lambda_n^{-1} = O(e^{Cn^2})\,!$$

Example: backward heat equation

Physical problem: find the temperature distribution in a system before thermal measurements are made (example: space shuttle re-entry).

$$\underbrace{\Theta(\cdot, T)}_{\text{measurement}} = \underbrace{\mathcal{A}(T)}_{\text{heat eq.}} \underbrace{\Theta(\cdot, 0)}_{\text{unknown}} \quad \Longrightarrow \quad \boxed{A(T)\Theta_0 = \Theta_T} \quad \text{after space discretization of } \Theta$$



Singular values of $A(T)$ ($x \in [0,1]$, $\Delta x = 1/100$)

- Matrix A: exact rank 100, numerical rank < 10.

Sensitivity of least squares solutions to data errors

Goal: solve $\boxed{\min_{x\in\mathbb{K}^n} \|Ax-b\|^2}$ with A "bad" (ill-conditioned, numerically rank-deficient).

Recall (again!) SVD of A: $A = USV^{\mathrm{H}} = \sum_{i=1}^{n} \sigma_i u_i v_i^{\mathrm{H}}$ $\quad \begin{cases} U = [u_1,\dots,u_n] \in \mathbb{K}^{m\times n} \\ V = [v_1,\dots,v_n] \in \mathbb{K}^{n\times n} \\ S = \mathrm{diag}(\sigma_1,\dots,\sigma_n) \in \mathbb{R}^{n\times n} \end{cases}$

- Unique solution ($\mathrm{rank}(A) = n$ assumed), given by
$$x = \sum_{i=1}^{n} \frac{u_i^{\mathrm{H}} b}{\sigma_i} v_i.$$
- Noisy data $b_\delta = b + w$ with $\|b - b_\delta\|_2 = \|w\|_2 = \delta$ (δ: size of data error).
Solution error:
$$x_\delta = \sum_{i=1}^{n} \frac{b_\delta^{\mathrm{H}} u_i}{\sigma_i} v_i, \qquad x_\delta - x = \sum_{i=1}^{n} \frac{w^{\mathrm{H}} u_i}{\sigma_i} v_i, \qquad \frac{\|x_\delta - x\|_2}{\delta} = \frac{1}{\delta}\Big(\sum_{i=1}^{n} \frac{|w^{\mathrm{H}} u_i|^2}{\sigma_i^2}\Big)^{1/2}.$$
- If A numerically rank deficient, may have $\boxed{\frac{w^{\mathrm{H}} u_1}{\sigma_1} \text{ small, but } \frac{w^{\mathrm{H}} u_i}{\sigma_i} \text{ large for some } i.}$
In some cases, exponential decay of σ_i: $|w^{\mathrm{H}} u_i|/\sigma_i$ very large even if δ small.

Low-rank approximations

- Often useful to replace A with low-rank approximation A_r with $\|A - A_r\|$ "small enough"
- Natural choice of A_r: truncated SVD (TSVD) of A

$$\boxed{\widehat{A}_r := U_r S_r V_r = \sum_{i=1}^{r} \sigma_i u_i v_i^{\mathsf{H}}} \qquad \begin{cases} U_r = [u_1, \ldots, u_r] \in \mathbb{K}^{m \times r} \\ V_r = [v_1, \ldots, v_r] \in \mathbb{K}^{n \times r} \\ S_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r \times r} \end{cases}$$

Eckart-Young-Mirsky theorem

Let $A \in \mathbb{K}^{m \times n}$, $r \leq n$. $\widehat{A}_r$ is best rank-r approximation of A (spectral and Frobenius norms):

$$\widehat{A}_r = \underset{\substack{B \in \mathbb{K}^{m \times n} \\ \operatorname{rank}(B) = r}}{\arg\min} \left\{ \|A - B\|_2 \text{ or } \|A - B\|_{\mathsf{F}}. \right\}; \qquad \|A - \widehat{A}_r\|_2 \leq \sigma_{r+1}, \quad \|A - \widehat{A}_r\|_{\mathsf{F}}^2 \leq \sum_{i=r+1}^{n} \sigma_i^2.$$

- Corresponding estimates for relative truncation errors:

$$\frac{\|A - \widehat{A}_r\|_2}{\|A\|_2} \leq \frac{\sigma_{r+1}}{\sigma_1}, \qquad \frac{\|A - \widehat{A}_r\|_{\mathsf{F}}}{\|A\|_{\mathsf{F}}} \leq \frac{\sqrt{\sum_{i=r+1}^{n} \sigma_i^2}}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}.$$

  Smallest rank r such that $\|A - \widehat{A}_r\| \leq \varepsilon$ can be found knowing $\sigma_1, \ldots, \sigma_n$.

Practical computation of $\widehat{A}_r$ given A potentially expensive (needs SVD of A, $O(m^2 n)$ cost).

Regularized least squares

- Alternative to low-rank truncation by SVD: modified minimization problem (Tikhonov(-Phillips) regularization)

$$\min_{x\in\mathbb{K}^n} J_\alpha(x;b), \quad J_\alpha(x;b) := \|Ax-b\|_2^2 + \alpha\|x\|_2^2 \qquad (\alpha \geq 0 \text{ "small"})$$

- Heuristic idea: $\alpha\|x\|_2^2$ "penalizes" solutions x with $\|x\|$ large.
  Supplementary *prior information*: prefer solutions with smaller $\|x\|$.
- Analysis: use SVD of A (note that $\|x\|_2^2 = \|V^{\mathrm{H}}x\|_2^2$):

$$\boxed{J_\alpha(x;b) = \sum_{i=1}^{n}\left\{|\sigma_i y_i - z_i|^2 + \alpha|y_i|^2\right\} + \sum_{i=n+1}^{m}|z_i|^2} \qquad (y_i := v_i^{\mathrm{H}}x,\ z_i := u_i^{\mathrm{H}}b)$$

  Minimization uncouples into n univariate quadratic minimizations, hence

$$y_i = \frac{\sigma_i z_i}{\sigma_i^2+\alpha}, \qquad x_\alpha = \sum_{i=1}^{n}\frac{\sigma_i z_i}{\sigma_i^2+\alpha}v_i$$

- Properties of regularized least squares solution x_α:
  - $\rightarrow$ x_α is unique for any $\alpha > 0$ (even if $\mathrm{Rank}(A) < n$);
  - $\rightarrow$ For $\alpha > 0$, x_α does not minimize $\|Ax-b\|_2^2$;
  - $\rightarrow$ Limit of x_α as $\alpha \to 0$ is minimum-norm least-squares solution of $Ax = b$.

Regularized least squares, noisy data

- Solve $Ax = b_\delta$ with noisy data $b_\delta = b + w$ (with $\|w\|_2 = \delta$).
- Regularized solution for data b_δ:

$$x_{\alpha,\delta} := \arg\min_x J_\alpha(x; b_\delta) = \sum_{i=1}^{n} \frac{\sigma_i z_i^\delta}{\sigma_i^2 + \alpha} v_i = x_\alpha + \sum_{i=1}^{n} \frac{\sigma_i (w^H u_i)}{\sigma_i^2 + \alpha} v_i$$

- Regularized solution error $e_{\alpha,\delta} := x_{\alpha,\delta} - x$:

$$e_{\alpha,\delta} = e_{\alpha,\delta}^{\text{reg}} + e_{\alpha,\delta}^{\text{noise}}, \qquad e_{\alpha,\delta}^{\text{reg}} = -\alpha \sum_{i=1}^{n} \frac{z_i}{\sigma_i(\sigma_i^2 + \alpha)} v_i, \quad e_{\alpha,\delta}^{\text{noise}} = \sum_{i=1}^{n} \frac{\sigma_i (w^H u_i)}{\sigma_i^2 + \alpha} v_i.$$

Regularized least squares: choice of α using L-curve

Optimal choice method for α?

- Define

  $$J_\alpha(x_\alpha; b) = D(\alpha) + \alpha R(\alpha), \qquad D(\alpha) := \|Ax_\alpha - b\|_2^2, \quad R(\alpha) := \|x_\alpha\|_2^2$$

  and study behavior of $\alpha \mapsto \{ D(\alpha),\, R(\alpha) \}$

- We find

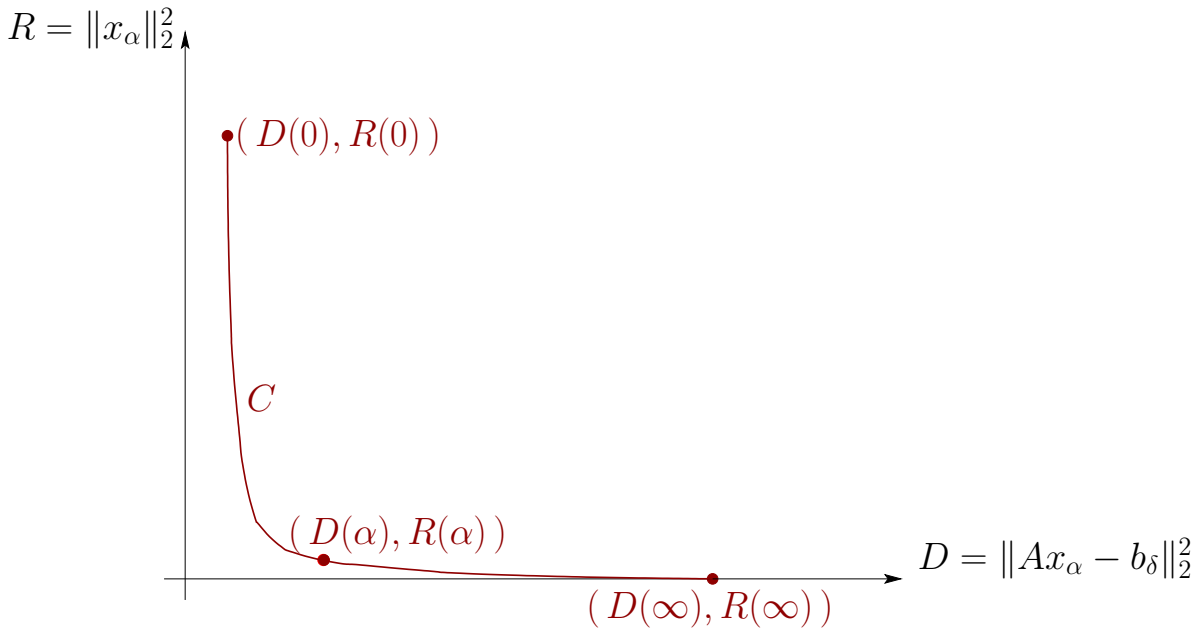
  $$D(\alpha) = \sum_{i=1}^{n} \frac{\alpha^2}{(\sigma_i^2+\alpha)^2} |z_i|^2 + \sum_{i=n+1}^{m} |z_i|^2, \qquad R(\alpha) = \sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2+\alpha)^2} |z_i|^2$$

  $$D'(\alpha) = 2\alpha \sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2+\alpha)^2} |z_i|^2 > 0, \qquad R'(\alpha) = -2 \sum_{i=1}^{n} \frac{\sigma_i^2}{(\sigma_i^2+\alpha)^3} |z_i|^2 < 0$$

- Outcome: $\alpha \mapsto D(\alpha)$ increasing and $\alpha \mapsto R(\alpha)$ decreasing, i.e.:

The L-curve $\alpha \in [0, \infty[\mapsto \big(D(\alpha), R(\alpha) \big)$ is convex

Regularized least squares: choice of α using L-curve



L-curve properties

- $C = (D, R)$ is monotonically decreasing (R a decreasing function of D) and convex;
- Extremities $A = \big(D(0), R(0)\big)$ and $A = \big(D(\infty), R(\infty)\big)$ of C given by

  $$D(0) = \|Ax - b\|_2^2, \quad R(0) = \|x\|_2^2, \qquad D(\infty) = \|b\|_2^2, \quad R(\infty) = 0$$

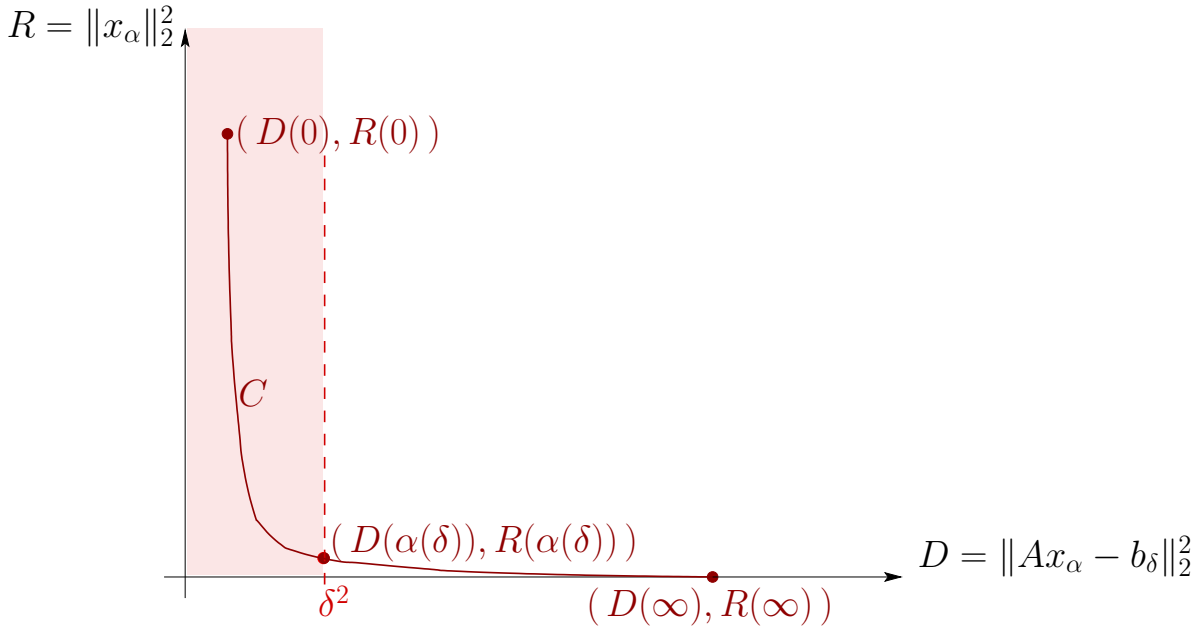
  ($x = x_0$: basic least-squares solution).

Regularized least squares: choice of α using L-curve

- Assume data noise level δ is known (realistic in some cases, e.g. mechanical testing using digital image correlation).
- Use that L-curve is convex, reformulate regularized least-squares:

$$\min_{x \in \mathbb{K}^n} \|x\|_2^2, \qquad \text{subject to} \quad \|Ax - b\|_2^2 \leq \delta^2$$

- Select α such that $D(\alpha) = \delta^2$ (i.e. set LS residual equal to data noise)
- Unique solution provided $\delta < \|b_\delta\|_2$



Regularized solution using truncated SVD

- Matrices with fast decay of σ_i: truncated SVD as alternative to Tikhonov regularization:

$$x_r := \arg\min_{x \in \mathbb{K}^n} \|\widehat{A}_r x - b\|^2 = \sum_{i=1}^{r} \frac{u_i^{\mathrm{H}} b}{\sigma_i} v_i$$

- By analogy to regularized least squares, define

$$D_r := \|\widehat{A}_r x_r - b\|_2^2 = \sum_{i=r+1}^{m} |u_i^{\mathrm{H}} b|^2, \qquad R_r := \|x_r\|_2^2 = \sum_{i=1}^{r} \frac{|u_i^{\mathrm{H}} b|^2}{\sigma_i^2}$$

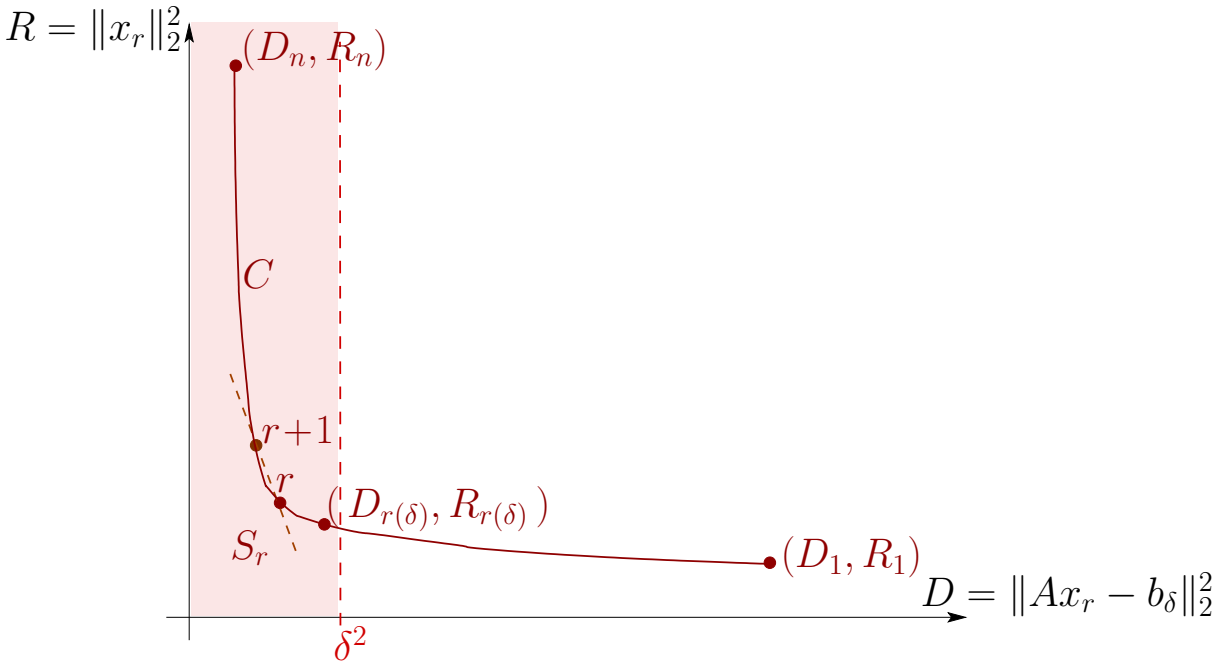
- Clearly $r \mapsto D_r$ decreasing and $r \mapsto R_r$ increasing.
- L-curve C_n: interpolates points (D_r, R_r) $(1 \le r \le n)$. C_n is convex:

$$S_r := \frac{R_r - R_{r+1}}{D_r - D_{r+1}} = -\frac{|z_{r+1}|^2}{\sigma_{r+1}^2} \frac{1}{|z_{r+1}|^2} = -\frac{1}{\sigma_{r+1}^2}, \qquad r \mapsto S_r \text{ increasing}$$

- Discrete parameter $1/r$ plays role of regularization parameter α. Optimal value:

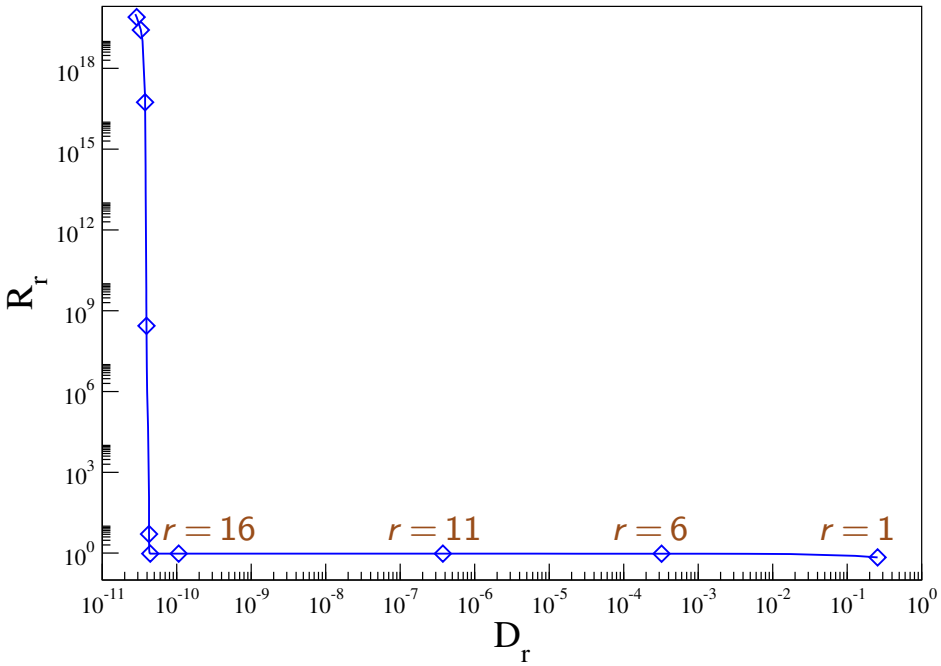
$$r(\delta) = \min_{1 \le r \le n} R_r \quad \text{subject to} \quad D_r \le \delta^2$$

Discrete L-curve



Example: backward heat equation

Discrete L-curve, simulated data with $\delta = 10^{-5}$,



- Optimal choice of r (L-curve for noise level $\delta = 10^{-5}$);
- Lowest actual temperature reconstruction error: $\approx 10^{-2}$ (in relative L^2 norm) for $r = 19$.