Low-Rank Approximation

Lecture 4 – Matrix equations

Leonardo Robol, University of Pisa, Italy

Cagliari, 23–27 Sep 2019

Matrix equations

We consider a problem apparently unrelated to the low-rank approximation framework: assume we have A, B, C, D, E such that

$$AXE + DXB = C,$$

where all the matrices have compatible sizes.

- We would like to solve the equation for X.
- Several equations fall in this category with particular choices of A, B, C, D, E.
- The equation is linear. Indeed, up to reassembling the unknowns this is a linear system.
- Solving it as a linear system hides most of the structure.

Classification

$$AXE + DXB = C$$

Depending on the matrices we choose, we give this equation a different name:

- $E = D = I$, then this is a Sylvester equation:

$$AX + XB = C, \qquad A \in \mathbb{C}^{m\times m},\ B \in \mathbb{C}^{n\times n},\ X, C \in \mathbb{C}^{m\times n}$$

- If in addition $B = A^H$, then this is a Lyapunov equation $AX + XA^H = C$.

The equations above are called continuous time Sylvester and Lyapunov, and they have a discrete time counter part:

$$AXB + X = C, \qquad AXA^H + X = C.$$

They arise from control theory — from continuous and discrete time ODEs, respectively.

Tools for the job

We need some standard tools in linear algebra, so let's make a brief recap:

- We denote with $A \otimes B$ the Kronecker product of A and B, that is:

$$\begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}.$$

- We denote by $\mathrm{vec}(X)$ the vectorized of X, which is the vector obtained by stacking the column of X one on top of the other:

$$\mathrm{vec}(X) := \begin{bmatrix} Xe_1 \\ \vdots \\ Xe_n \end{bmatrix}.$$

- MATLAB notation for these two operations: `kron(A, B)` and `X(:)`.

Important properties

- Kronecker products and vec operations play very well together.
- $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$.
- $\|\mathrm{vec}(X)\|_2 = \|X\|_F$.
- $(A \otimes B)(C \otimes D) = AC \otimes BD$ (this is valid any time the dimensions are compatible, so is for vectors as well).
- $(A \otimes B)^T = A^T \otimes B^T$.
- $\mathrm{rank}(A \otimes B) = \mathrm{rank}(A)\mathrm{rank}(B)$.
- We have important properties on the spectrum and the singular values:

$$\Lambda(A \otimes B) = \{\lambda_i \mu_j,\ \lambda_i \in \Lambda(A),\ \mu_j \in \Lambda(B)\}$$
$$\sigma_{ij}(A \otimes B) = \sigma_i(A)\sigma_j(B),$$

where the last equality is valid up to proper reordering.

Matrix equations as linear systems

Knowing properties of Kronecker products immediately allows to recast matrix equations as linear systems:

$$AXE + DXB = C \iff (E^T \otimes A + B^T \otimes D)\mathrm{vec}(X) = \mathrm{vec}(C).$$

- Particular cases have stronger structure, for instance the Sylvester equation $AX + XB = C$ gives rise to the linear system

$$(B^T \otimes I + I \otimes A)\mathrm{vec}(X) = \mathrm{vec}(C).$$

- If the matrix equation involves an $m \times n$ matrix X, this is an $mn \times mn$ linear system; using a non-structured linear system solver, we need $\mathcal{O}(m^3 n^3)$ flops to solve it. Often too expensive!

Existence of solution

Using the characterization of eigenvalues, it is easy to check when the solution of a Sylvester equation exists is unique:

- $AX + XB = C \iff (B^T \otimes I + I \otimes A)\mathrm{vec}(X) = \mathrm{vec}(C)$.
- The above linear system is uniquely solvable iff the matrix has no zero eigenvalues. That is,

$$\lambda_i(A) + \lambda_j(B) \neq 0 \iff \lambda_i(A) \neq -\lambda_j(B),$$

  which is to say that A, B have disjoint spectrum.
- Similar conditions for $AXB + I = C$:

$$\lambda_i\lambda_j(B) + 1 \neq 0 \iff \lambda_i(A) \neq -\lambda_j^{-1}(B).$$

- Can we answer for the more general case?

Existence of solution

If we consider the general equation

$$AXE + DXB = C,$$

the associated linear system $B^T \otimes D + E^T \otimes A$ is nonsingular if and only if the pencils

$$(A - \lambda D), \qquad (B + \lambda E)$$

have disjoint spectrum.

Proof.
Let us assume that E, D are nonsingular. Then,

$$\det(B^T \otimes D + E^T \otimes A) \neq 0 \iff \det(E^{-T} B^T \otimes I + I \otimes D^{-1} A) \neq 0,$$

and the latter inequality holds if and only if

$$\lambda_i(E^{-T} B^T) \neq -\lambda_j(D^{-1} A),$$

which are exactly the eigenvalues of the pencils above. What if E, D singular? □

Shifting the equation

The previous hypothesis about E, D non-singular is not really restrictive. Given $AXE + DXB = C$, we can always introduce shifts

$$(A - \lambda D)X(B + \eta E) + (\eta D - A)X(\lambda E + B) = (\eta - \lambda)C$$

- If we cannot find λ, η such that $B + \eta D$ and $\eta E - A$ or $A - \lambda E$ and $\lambda + B$ are nonsingular, then we can easily construct a nonzero solution in the kernel.
- If you are familiar with matrix pencils, it means both have nontrivial singular blocks in the Kronecker Canonical Form.
- Otherwise, we can always shift the pencils to be in the valid situation for the proof.
- This shifting property will come useful later!

Operator theory

The equation makes also sense when A, B, C, D, X are operators on some Banach algebra. In that case, the condition for existence of solution becomes on the spectra of the pencils:

$$\text{unique solution} \iff \{\lambda \mid \|(A - \lambda D)^{-1}\| < \infty\} \cap \{\lambda \mid \|(B + \lambda E)^{-1}\| < \infty\} = \emptyset.$$

- We get again the condition on eigenvalues in the finite dimensional case.
- Spectra are more complicated in an infinite dimensional case, not all points in the spectrum are eigenalues. The condition is anyway "the spectra are disjoint".
- The proof we had before does not work straight-away.

Closed formula for the solution

It turns out that one can write closed formula for the solution of Sylvester equations. These apply also to the infinite-dimensional case and can be indeed used for the proof.

Theorem
Let Γ_A and Γ_B close the spectrum of A and B, respectively, and assume that the enclosed set is disjoint. Then, the solution of $AXE + DXB = C$ is

$$X = -\frac{1}{4\pi^2}\int_{\Gamma_A}\int_{\Gamma_B}\frac{(\lambda D - A)^{-1}C(\mu E - B)^{-1}}{\lambda + \mu}d\lambda d\mu$$

In addition, if Γ encloses the spectrum of A but not the one of B, then

$$X = -\frac{1}{2\pi i}\int_{\Gamma}(A - \lambda D)^{-1}C(B + \lambda E)^{-1}\, d\lambda.$$

The condition for having disjoint spectra is clear in this last result.

Closed formula for the solution (continues)

These formula are mostly of theretical interest, because numerically it can be difficult to approximate them. Another which draws a connection with matrix functions is, for $AX + XB = C$, for A, B with spectrum in the left half plane,

$$X = -\int_0^\infty e^{tA} C e^{tB}\, dt.$$

- Note that this only makes sense for operators which are stable, i.e., for which $\|e^{tA}\| \to 0$ as $t \to \infty$.
- The same formula could be given when A, B have eigenvalues with positive real parts — up to switching the sign.
- The formula is valid under the weaker condition that A and $-B$ have spectra separated by a vertical line. Why?

Solution of a Sylvester equation – small scale

We now consider the case where A, B, are reasonably small, say below 1000×1000.

- We ruled out the possibility to use a Kroneckerized operator. That would only be feasible for n up to ≈ 60 or so.
- Luckily, a Sylvester equation with $n \times n$ matrices can be solved in $\mathcal{O}(n^3)$ flops, making use of Schur decompositions.
- Brief reminder: given any matrix A, we have a unitary matrix Q such that

$$Q^H A Q = T,$$

  with T upper triangular.
- The decomposition can be computed in $\mathcal{O}(n^3)$ time using the QR method.

Solution of a Sylvester equation – small scale

Here is the algorithm to solve the equation $AX + XB = C$.

- Compute the Schur forms for A and B:

$$Q_A^H A Q_A = T_A, \qquad Q_B^H B Q_B = T_B.$$

- Transform the equation using Q_A and Q_B:

$$Q_A^H C Q_B = Q_A^H (AX + XB) Q_B = T_A Q_A^H X Q_B + Q_A^H X Q_B T_B = T_A Y + Y T_B,$$

  where T_A, T_B are upper triangular.
- Solve the equation $T_A Y + Y T_B = Q_A^H C Q_B$ by back-substitution, computing the entries starting from the entry in position (n, n) up to $(1, 1)$.
- Recover the original solution as $X = Q_A Y Q_B^H$.

Cost: $\mathcal{O}(n^3)$. Known as Bartels-Stewart algorithm. More efficient variants are nowadays available (for instance, the Hessenberg-Schur method).

Stability

The Bartels-Stewart algorithm is stable, as a linear system solver. That is, the computed solution $X + \delta X$ of $AX + XB = C$ will satisfy

$$(\mathcal{A} + \delta A)\mathrm{vec}(X + \delta X) = \mathrm{vec}(C + \delta C),$$

where $\mathcal{A} = B^T \otimes I + I \otimes A$. Does δA have the same structure of $\mathcal{A}$? That is, does it hold that:

$$\left[(B + \delta B)^T \otimes I + I \otimes (A + \delta A)\right] \mathrm{vec}(X + \delta X) = \mathrm{vec}(C + \delta C).$$

Unfortunately, not. However, error bounds in the Frobenius norm are easily derivable from the linear system.

Condition number

Associated with stability one is also interested in understanding the condition number of a matrix equation $AX + XB = C$ (or a more general one).

If we use a stable algorithm, then the computed solution will satisfy

$$\|\delta X\| \leq \kappa \cdot \epsilon + \mathcal{O}(\epsilon^2),$$

where ϵ is the size of the backward error, and κ the condition number.

- We cannot go into details, but essentially:

$$\kappa \approx \frac{1}{\mathrm{sep}[(\mathrm{A}, -\mathrm{B})]},$$

  where sep measures the distance between the spectra of A and $-B$.
- Related to the (structured) condition number of $B^T \otimes I + I \otimes A$.

Solution for large A and small B

Suppose $A \in \mathbb{C}^{m\times m}$ is large, and $B \in \mathbb{C}^{n\times n}$ is small. Then,

$$AX + XB = C$$

can be solved efficiently by the following steps:

- We transform B to upper triangular form, which yields the transformed system:

$$AY + YT_B = CQ_B, \qquad Y = XQ_B, \qquad Q_B^H B Q_B = T_B.$$

- Then, we have the following equation for the first column of Y:

$$AYe_1 + YT_B e_1 = CQ_B e_1 \iff (A + (T_B)_{11} I)Ye_1 = CQ_B e_1.$$

- Once we solve it, we can continue with the second column:

$$AYe_2 = YT_B e_2 = CQ_B e_2 \iff (A + (T_B)_{22} I)Ye_2 = CQ_B e_2 - (T_B)_{12} Ye_1.$$

- We continue until we get all the columns if Y (which are a few).
- Requires to solve a bunch of shifted linear systems with A, so it makes it easy to exploit sparsity and other features of the larger scale matrix.

Large scale case

We now consider the case where both A, B are large scale, so that the strategies described previously are not applicable to the problem.

- In this setting, we need to make additional hypotheses on the data we have at hand.
- This will draw a connection with low-rank approximation.

Standard setup from now on:

$$AX + XB = C = UV^T, \qquad U \in \mathbb{C}^{m\times k}, \quad V \in \mathbb{C}^{n\times k}.$$

In practice, we will consider $k = 1$, as it simplifies the discussion and everything generalizes easily. We will add a few comments on the general case only when the differences are relevant.

The plan

We will cover a few possibilities for the approximation of the solution X. In all cases, we will assume that X has numerically low-rank; we will later on see how this is justified.

- We will use projection methods, and in particular Krylov subspaces.
- Then, we will see that the solution can be computed much faster with a particular iteration related to rational functions: the ADI method. This will be highly relevant from a theoretical perspective as well.
- Finally, we will see that rational Krylov subspaces can put together all the advantages of the previous tools.

Projection methods

Projection methods, including Krylov subspaces, can all be formulated in a similar manner. We will derive them by looking again at the Kronecker system.

- This is a slightly unusual derivation, but I feel it gives some insight on why these methods are so powerful.
- It also helps to bridge the gap with iterative methods for linear system.
- Once this approach is clear, the more "usual" derivation follows very easily.

Before starting, we need a very short 1-minute recap on iterative methods for linear system.

Solving linear system iteratively (by projection methods)

Let's switch focus for a second, and assume we are solving a linear system $Ax = b$.

- Suppose A is large scale, so we only have a a fast matrix-vector operation available.
- We cannot look at the entries of A explicitly: no sparse factorization methods. In particular, no `A\b` in MATLAB.

General idea: construct a sequence of low-dimensional subspaces $\mathcal{U}_1 \subseteq \mathcal{U}_2 \subseteq \dots$ such that

- The dimension of $\mathcal{U}_\ell$ is typically ℓ; we denote U_ℓ an orthogonal basis for it.
- For each ℓ, we find a solution $x_\ell = U_\ell \hat{x}_\ell$ that is the "best" under some appropriate metric. For instance, it might minimize $\|Ax_\ell - b\|_2$ among all possible x_ℓ of the above form.
- If we choose our space wisely, then $x_\ell \to x$ and $\|Ax_\ell - b\|_2 \to 0$ fast enough.

Krylov subspaces

The usual choice of the projection subspace is a Krylov subspace generated by A and b:

$$U_\ell = \mathcal{K}_\ell(A, b) := \operatorname{span}\{b, Ab, \ldots, A^{\ell-1}b\}.$$

This has many advantages:

- The projected matrix $A_\ell := U_\ell^T A U_\ell$ is automatically available from the Arnoldi iteration, and the basis can be computed with $\ell - 1$ matrix-vector products + reorthogonalizations.
- One can interpret the solution of the linear system as evaluating the matrix function $f(z) = \frac{1}{z}$, this gives the FOM method.
- Other strategies possible — using the rectangular Hessenberg matrix gives GMRES, and so on.

Convegence analysis

The convergence can be linked to a polynomial approximation problem. The residual at step ℓ satisfies:

$$\frac{\rho_\ell}{\rho_0} \leq \min_{\substack{\deg p(x) \leq \ell - 1 \\ p(0)=1}} \|p(A)\|_2 \cdot \|b\|_2, \qquad \rho_\ell := \|Ax_\ell - b\|_2.$$

- We have already discussed the issues in evaluating $\|p(A)\|$: this is connected with the eigenvalues only in the normal case.
- In the latter case, the convergence is very fast is the spectrum is clustered around 1 (or any other value).
- Otherwise, polynomial approximation might be quite slow (and therefore, we need preconditioning).
- Similar convergence properties among all the methods in this class (FOM, GMRES, ...).
- $x_\ell \in \mathcal{U}_\ell \implies (Ax_\ell - b) \in \mathcal{U}_{\ell+1}$: checking the residual is very economical.

Choosing a smarter space

In our case, the matrix of the linear system has a particular structure:

$$\mathcal{A} := I \otimes A + B^T \otimes I.$$

- We can make a smarter choice for the space, trying to preserve this structure.
- If we choose $\mathcal{U}_\ell = \mathcal{V}_\ell \times \mathcal{W}_\ell$, then we have

$$U_\ell = V_\ell \otimes W_\ell \implies U_\ell^T \mathcal{A} U_\ell = I \otimes V_\ell^T A V_\ell + W_\ell^T B^T W_\ell \otimes I.$$

- The projected problem still retains the same structure, i.e., it is still a matrix equation.
- The dimension of U_ℓ is, in general, ℓ^2.

Choosing $\mathcal{U}_j$

Since we have seen that Krylov subspace work so well, we may set (recall that $C = UV^T$):

$$\mathcal{V}_\ell = \mathcal{K}_\ell(A, U), \qquad \mathcal{W}_\ell = \mathcal{K}_\ell(B^T, V).$$

Then, it follows that the projected linear system takes the form

$$U_\ell^T \mathcal{A} U_\ell = I \otimes A_\ell + B_\ell^T \otimes I,$$

where $A_\ell := V_\ell^T A V_\ell$ and $B_\ell = W_\ell^T B W_\ell$. This is equivalent to solving a $\ell \times \ell$ matrix equation

$$A_\ell Y + Y B_\ell = V_\ell^T C W_\ell,$$

- Since this is now small scale, we can solver with our favorite dense solver (Bartels-Stewart or Hessenberg-Schur).
- The approximation is given by $X_\ell = V_\ell Y W_\ell^T$.

Computing the residual

In order to stop the iteration, we would like to verify that the residual norm is smaller than some tolerance

$$\|R_\ell\|_F = \|AX_\ell + X_\ell B - C\|_F \leq \tau$$

- Very natural condition, since it ensures a small backward error on the underlying linear system.
- In principle, it looks super-expensive to compute!
- Luckily, we have that $R_\ell \in \mathcal{U}_{\ell+1}$.

A sketched algorithm to solve the problem

- Compute the tensorized Krylov subspace

$$\mathcal{U}_{\ell+1} := \mathcal{K}_{\ell+1}(A, U) \times \mathcal{K}_{\ell+1}(B^T, V).$$

- Solve the projected equation $A_\ell Y + Y B_\ell = V_\ell^T C_\ell$ in the smaller subspace $\mathcal{U}_\ell$, and call Y the solution of the projected problem. Cost: $\mathcal{O}(\ell^3)$!.
- Evaluate the residual as

$$\|R_\ell\|_2 = \max\left\{\|e_{\ell+1}^T A_{\ell+1} Y\|_2, \|Y B_{\ell+1} e_{\ell+1}\|_2\right\}$$

- . . . or

$$\|R_\ell\|_F = \sqrt{\|e_{\ell+1}^T A_{\ell+1} Y\|_2^2 + \|Y B_{\ell+1} e_{\ell+1}\|_2^2}$$

- Quite easy proof!

Take-home messages (part 1)

- Exploiting the structure of the problem at hand makes the projected problem of the same type.
- This enables a much more efficient solution of the small problem.
- It also enables to construct a space of dimension ℓ^2 at the cost of one of dimension ℓ.

However, the approach is not drawback-free:

- Convergence with rate $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, where $\kappa \approx \kappa(B^T \otimes I + I \otimes A)$:

  $$\|R_\ell\|_2 \leq C\rho^\ell$$

  The expression for κ can be given in terms of separation of the spectra for normal matrices.
- For positive definite problems where A or B are ill-conditioned, this is very slow.
- Very often the case in practice! This made the choice of Krylov methods not so attractive for a long time in this setting.

ADI iteration: back to rational functions

Recall that $AX + XB = UV^T$ holds iff

$$(A - \lambda I)X(B + \eta I) + (\eta I - A)X(\lambda I + B) = (\eta - \lambda)C = (\eta - \lambda)UV^T$$

Assuming $\lambda \notin \Lambda(A)$ and $\eta \notin \Lambda(-B)$, and $C = UV^T$,

$$X = r(A)^{-1} X r(-B) + (\eta - \lambda)(A - \lambda I)^{-1} UV^T (B + \eta I)^{-1},$$

where $r(z) = -(z - \lambda)/(z - \eta)$.

- This observation will be the basis of the ADI method.
- Can also be used as a theoretical tool to prove that X is numerically low-rank.

A closer look at the solution

$$X = (\eta - \lambda)(A - \lambda I)^{-1} U V^T (B + \eta I)^{-1} + r(A)^{-1} X r(-B)$$

- Here η, λ are completely arbitrary — so we may consider the relation it for different parameters

$$\eta_1, \lambda_1, \ldots, \eta_\ell, \lambda_\ell.$$

- We can compose the relation with itself over and over:

$$X = (\eta_1 - \lambda_1)(A - \lambda_1 I)^{-1} U V^T (B + \eta_1 I)^{-1} + r_1(A)^{-1} X r_1(-B)$$

$$X = (\eta_2 - \lambda_2)(A - \lambda_2 I)^{-1} U V^T (B + \eta_2 I)^{-1}$$
$$+ r_2(A)^{-1}(\eta_1 - \lambda_1)(A - \lambda_1 I)^{-1} U V^T (B + \eta_1 I)^{-1} r_1(-B) + r_{12}(A)^{-1} X r_{12}(-B),$$

where $r_{12}(z) = r_1(z) r_2(z)$.

A closer look at this relation

$$X = (\eta - \lambda)(A - \lambda I)^{-1} U V^T (B + \eta I)^{-1} + r(A)^{-1} X r(-B)$$

- Reiterating we obtain that, for any rational function $r(z)$ of degree (ℓ, ℓ),

$$X = X_\ell + r(A)^{-1} X r(-B), \qquad X_\ell \in \mathcal{U} \otimes \mathcal{V}$$

- $X_\ell \approx X$ if $\|r(A)^{-1} X r(-B)\|$ is small, indeed:

$$\|X - X_\ell\|_F \leq \|X\|_F \cdot \|r(A)^{-1}\|_2 \|r(-B)\|_2 = \|X\|_F \frac{\max\limits_{z \in \Lambda(B)} |r(-z)|}{\min\limits_{z \in \Lambda(A)} |r(z)|},$$

  where the last equality only holds for normal matrices.

- Related to Zolotarev problem, explicit solution if

$$\Lambda(A) = \Lambda(B) \subseteq [a, b] \subseteq \mathbb{R}_+.$$

ADI iteration

- We observe that, after ℓ steps, the decomposition of X can be written as:

$$X = X_\ell + r(A)^{-1} X r(-B), \qquad r(z) = \prod_{j=1}^{\ell} \frac{\lambda - z}{z - \eta}.$$

- X_ℓ does not depend on X:

$$\begin{aligned} X_1 &= (\eta_1 - \lambda_1)(A - \lambda_1 I)^{-1} U V^T (B + \eta_1 I)^{-1} \\ X_2 &= (\eta_2 - \lambda_2)(A - \lambda_2 I)^{-1} U V^T (B + \eta_2 I)^{-1} \\ &\quad + r_2(A)^{-1}(\eta_1 - \lambda_1)(A - \lambda_1 I)^{-1} U V^T (B + \eta_1 I)^{-1} r_1(-B) \\ &= X_1 + r_2(A)^{-1} X_1 r_1(-B) \\ &\vdots \end{aligned}$$

- Is X_ℓ a good approximation to X? This only depends on the remaining term:

$$\|X - X_\ell\|_2 = \|r(A)^{-1} X r(-B)\|_2 \leq \|r(A)^{-1}\| \|X\|_2 \|r(-B)\|_2.$$

ADI iteration

This iteration is known as ADI iteration, and the equation

$$X - X_\ell = r(A)^{-1} X r(-B)$$

is often referred as the ADI residual representation.

- The iteration can be formulated explicitly as a matrix iteration involving shifted inverses of A and B.
- The rank of X_ℓ is at most ℓ (or ℓk if $UV^T = C$ has rank k): it is the sum of ℓ rank 1 (or rank k) matrices.
- The explicit error representation makes it very powerful: we exactly know what we need to minimize when choosing η_j, λ_j.

Factored ADI

- Since X_ℓ has rank $k\ell$ — at most — we can make the iteration very efficient.
- We have a low-rank parametrization for X_1 from the start:

$$X_1 = \underbrace{(\eta_1 - \lambda_1)(A - \lambda_1 I)^{-1} U}_{W_1} \underbrace{V^T (B + \eta_1 I)^{-1}}_{Z_1^T} = W_1 Z_1^T.$$

- Then, we set $X_2 = r_2(A)^{-1} X_1 r_2(-B) + X_1$, which yields:

$$X_2 = [W_1 \ r_2(A)^{-1} W_1] \, [Z_1 \ r_2(-B^T) Z_1]^T.$$

- It may seem that the above has rank 2^ℓ at step ℓ, but we could write it in a smarter way so that it becomes obvious that it has rank at most ℓ. Key observation[1]:

$$W_\ell \in \operatorname{span} \left\{ U, (A - \lambda_1 I)^{-1} U, (A - \lambda_2 I)^{-1} (A - \lambda_1 I)^{-1} U, \ldots, \prod_{j=1}^{\ell} (A - \lambda_j I)^{-1} U \right\}.$$

[1] Note that the order in which we take matrix products does not matter, because they all commute.

Convergence theory

The key question is: how fast does this converge, if it does?

- Clearly the iterations are more expensive that the Krylov case: they require matrix inversions.
- Unclear how we can make $\|r(A)^{-1}\|_2$ and $\|r(-B)\|$ small.

Lemma
Let A, B be normal matrices. Then,

$$\|r(A)^{-1}\|_2 \|r(-B)\| \leq \frac{\max_{z\in\Lambda(B)} |r(-z)|}{\min_{z\in\Lambda(A)} |r(z)|}.$$

Intuition: at least in the normal case, we need to find a rational function that is large on $\Lambda(A)$ and small on $\Lambda(-B)$. This is known as the fourth Zolotarev problem.

Zolotarev problems

In his paper in 1877, Zolotarev posed (and solved!) the following problems:

1. What is the solution of the following minimax problem?
$$\min_{\deg p(z)\le \ell} \frac{\max_{z\in[a,b]} |p(z)|}{\min_{z\in[-b,-a]|p(z)|}}.$$
2. What is the best polynomial approximant of degree at most (ℓ,ℓ) for the sign function over $[a,b]\cup[-b,-a]$?
3. What is the solution of the following minimax problem?
$$\min_{\deg r(z)\le (\ell,\ell)} \frac{\max_{z\in[a,b]} |r(z)|}{\min_{z\in[-b,-a]|r(z)|}}.$$
4. What is the best rational approximant to the sign function over $[-b,-a]\cup[a,b]$ of degree at most (ℓ,ℓ)?

We are mainly interested in problem 3.

Equivalence of the problems and the square root

It can be seen that problem 3. and 4. are indeed equivalent: the same method used for the minimax problem yields an approximant for the sign function as well.

We also note that this yields a rational approximant for the square root. Indeed, if $r(x)$ is the best rational approximant for $\mathrm{sign}(x)$ we need to have

$$r(x) = x\frac{p(x^2)}{q(x^2)},$$

because the optimal approximant needs to be odd for symmetry reasons. Then, note that, for $x \geq 0$,

$$\frac{\sqrt{x}}{\mathrm{sign}(\sqrt{x})} \approx \frac{q(x)}{p(x)},$$

and therefore $q(x)/p(x)$ is the best rational approximant to $\sqrt{x}$ with respect to relative accuracy: it has the minimum possible $\|\epsilon(x)\|_\infty$ such that:

$$\sqrt{x} = \frac{q(x)}{p(x)}(1 + \epsilon(x)),$$

Convergence speed

Zolotarev found the best approximant, and also described the convergence speed. We denote by

$$Z_\ell([a,b]) := \min_{\deg r(z)\le(\ell,\ell)} \frac{\max_{z\in[a,b]}|r(z)|}{\min_{z\in[-b,-a]|r(z)|}}.$$

Then,

$$Z_\ell([a,b]) \le 4e^{-\frac{\ell\pi^2}{\log(4\frac{b}{a})}}.$$

The original bound would take the form

$$Z_\ell([a,b]) \le 4e^{-\frac{\ell\pi^2}{\mu(\frac{a}{b})}},$$

where μ is the Grötzsch ring function. But the logarithm is almost as sharp, and much more easily computable.

ADI error formula

Plugging this bound into the result for the ADI iteration, we have the following convergence bound.

Theorem
Let X_ℓ the result of the ADI iteration applied to a Sylvester equation $AX + XB = UV^T$ such that A, B are symmetric positive definite with spectrum contained in $[a, b]$. Then,

$$\|X - X_\ell\|_2 \leq \|X\|_2 \cdot Z_\ell([a, b]) \leq 4\|X\|_2 \cdot e^{-\frac{\ell\pi^2}{\log(4\frac{b}{a})}}.$$

Natural question: how does this compare with the convergence of the polynomial Krylov method that we saw before?

Convergence speeds (in practice)

Polynomial Krylov

b/a:	10	100	10^3	10^6
5 steps	3.8e-02	3.7e-01	7.3e-01	9.9e-01
10 steps	1.4e-03	1.3e-01	5.3e-01	9.8e-01
20 steps	2.0e-06	1.8e-02	2.8e-01	9.6e-01

ADI method

b/a:	10	100	10^3	10^6
5 steps	1.5e-06	2.6e-04	2.6e-03	3.9e-02
10 steps	2.4e-12	7.0e-08	6.8e-06	1.5e-03
20 steps	5.8e-24	4.9e-15	4.6e-11	2.3e-06

Singular value decay and low-rank approximation

The ADI iteration has another important theoretical conseqeuence: it allows to predict the numerical rank of the solution X.

We may define:

$$\mathrm{rank}_\epsilon(X) = \min_{\|\delta X\|_2 \le \epsilon \|X\|_2} \mathrm{rank}(X + \delta X).$$

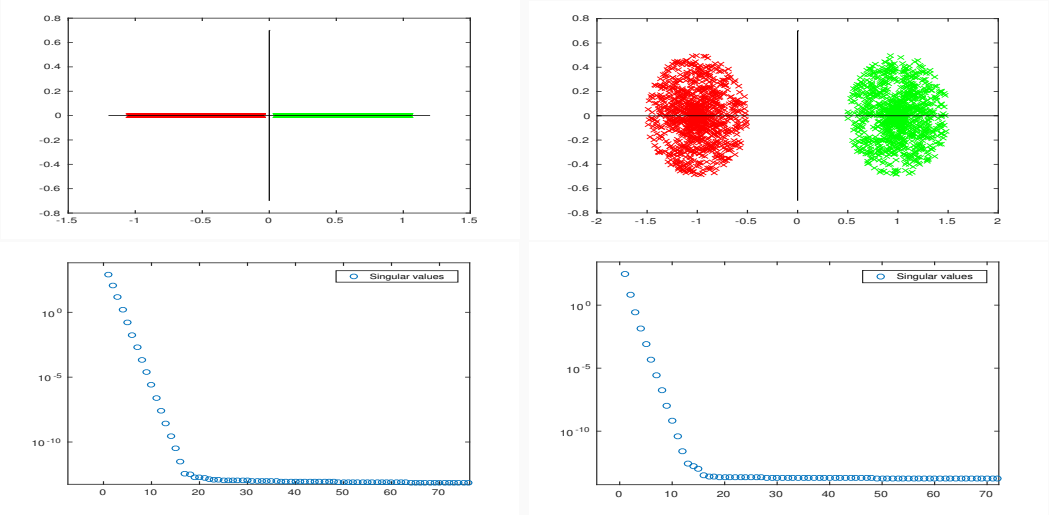
Clearly, $\mathrm{rank}_\epsilon(X) = \#\{\sigma_j(X) > \sigma_1(X)\epsilon\}$. Therefore, we can estimate the numerical rank if we know how fast the singular values decay.

Theorem
Let X be the solution of $AX + XB = UV^T$, with A, B posdef with spectrum inside $[a, b]$. Then,

$$\sigma_{\ell+1}(X) \le \sigma_1(X) Z_\ell([a, b]) \le 4\sigma_1(X) e^{-\frac{\ell\pi^2}{\log(4\frac{b}{a})}}.$$

Proof.
Use ADI error equation. □

Numerical rank of the solution



Singular values decay in the solution of $AX + XB = C$ with $\mathrm{rank}(C) = 1$, for two different configurations of the spectra of A and $-B$.

Take home messages (ADI) - part 2

- ADI can be a very powerful method for approximating the solution if A, B are not so well-conditioned: much faster convergence than in the polynomial case.
- Error equation of ADI is also a nice theoretical tool.
- The factored version of ADI can be even more efficient.
- For the Hermitian case, there is explicit knowledge of the Zolotarev numbers $Z_\ell([a, b])$.
- The solution requires several shifted linear systems: this might be quite expensive, since it reduces the possibility of reusing LU factorizations and/or preconditioners.

Rational Krylov subspaces

- X_ℓ, the solution at step ℓ of the ADI iteration, belongs to the subspace $\mathcal{V}_\ell \otimes \mathcal{W}_\ell$, where

$$\mathcal{V}_\ell = \text{span}\left\{ (A - \lambda_1 I)^{-1} U, \ldots, \prod_{j=1}^{\ell} (A - \lambda_j I)^{-1} U \right\}$$

$$\mathcal{W}_\ell = \text{span}\left\{ (B^T - \eta_1 I)^{-1} V, \ldots, \prod_{j=1}^{\ell} (B^T - \eta_j I)^{-1} V \right\},$$

  assuming λ_i, η_i are the zeros and poles of $r(z)$.

- $\mathcal{V}_j, \mathcal{W}_j$ known as rational Krylov subspaces.

This suggests an idea: can we use these subspaces instead of the usual Krylov ones for approximating the solution?

Rational Krylov subspaces

A rational Krylov subspace is completely determined by the matrix A, a vector b, and a sequence of poles $\xi_1, \ldots, \xi_\ell$. With our previous definition:

$$\mathcal{RK}_\ell(A, b, \{\xi_1, \ldots, \xi_\ell\}) := \operatorname{span}\left\{ (A - \xi_1 I)^{-1} b, \ldots, \prod_{j=1}^{\ell} (A - \xi_j I)^{-1} b \right\}.$$

- The space does not depend on the order of the poles ξ_j.
- Analogously to the Krylov case, it contains all the vectors of the form $r(A)b$, where $r(z)$ is a rational function $p(z)/q(z)$, with poles included in $\{\xi_1, \ldots, \xi_1 \ell\}$ and $\deg p \leq \ell - 1$.
- It can be written as:

$$\mathcal{RK}_\ell(A, b, \{\xi_1, \ldots, \xi_\ell\}) = \prod_{j=1}^{\ell} (A - \xi_j I)^{-1} \mathcal{K}_\ell(A, b).$$

Poles at infinity

Using the last definition, we allow for some of the ξ_j to be ∞. This just means:

$$\mathcal{RK}_\ell(A, b, \{\xi_1, \ldots, \xi_\ell\}) = \left[\prod_{\substack{j=1 \\ \xi_j \neq \infty}}^{\ell} (A - \xi_j I)^{-1} \right] \mathcal{K}_\ell(A, b).$$

- Note that if $\xi_1 = \ldots = \xi_\ell = \infty$, then $\mathcal{RK}_\ell(A, b, \{\xi_1, \ldots, \xi_\ell\}) = \mathcal{K}_\ell(A, b)$.
- If we include at least one infinity pole, then $b \in \mathcal{RK}_\ell(A, b, \{\xi_1, \ldots, \xi_\ell\})$.
- Natural interpretation in terms of homogeneous polynomials and poles on the Riemann sphere (a polynomial is a rational function with $\deg p$ poles at infinity).

Rational Arnoldi method

For polynomial Krylov subspaces, the Arnoldi method allows to build the basis of $\mathcal{K}_\ell(A, b)$ by iteratively extending an Hessenberg matrix H that satifies

$$AU_\ell = U_\ell A_\ell + e_{\ell+1} v_j^T.$$

If we allow A_ℓ to be $(\ell + 1) \times \ell$, instead of square, we may rephrase this more compactly:

$$AU_\ell = U_{\ell+1} \begin{bmatrix} A_\ell \\ \beta_{\ell+1} e_\ell^T \end{bmatrix}$$

Let us call the above rectangular matrix H_ℓ. If we call K_ℓ the rectangular matrix with the $\ell \times \ell$ identity on top, and zero elsewhere:

$$AU_{\ell+1} K_\ell = U_{\ell+1} H_\ell$$

The above is an instance of a rational Arnoldi decomposition.

Rational Arnoldi decompositions

We can associate a rational Krylov space $\mathcal{RK}_\ell(A, b)$ with a rational Arnoldi decomposition (RAD):

$$AU_{\ell+1}K_\ell = U_{\ell+1}H_\ell,$$

where:

- U_ℓ spans $\mathcal{RK}_\ell(A, b)$, and $U_{\ell+1}$ spans $\mathcal{RK}_{\ell+1}(A, b)$.
- K_ℓ, H_ℓ are upper Hessenberg.
- The matrices can be extended iteratively by solving shifted linear system + reorthogonalization.

Rational Arnoldi procedure

- Initially, we compute the first vector in the basis: $u_1 := \alpha_1^{-1}(A - \xi_1)^{-1}b$, where α_1 is used to renormalize it.
- This yields the starting relation:

$$AU_1K_1 = U_1H_1, \qquad U_1, K_1 \in \mathbb{C}^{1\times 0}.$$

- To go from $j-1$ to j, we select a continuation vector c_j, and compute:

$$u_j := \alpha_j^{-1}(A - \xi_j I)^{-1}(U_{j-1}c_j) - U_{j-1}\begin{bmatrix}\theta_1\\ \vdots \\ \theta_{j-1}\end{bmatrix}.$$

  Here θ_i and α_j are the reorthogonalization and renormalization coefficients.
- Rearrange the above relation as:

$$\alpha_j A u_j - \alpha_j \xi_j u_j = U_{j-1}c_j - \alpha_j A U_{j-1}\theta + \alpha_j \xi_j U_{j-1}\theta$$

Rational Arnoldi procedure (continued)

- Rearrange the above relation as:

$$\alpha_j A u_j - \alpha_j \xi_j u_j = U_{j-1} c_j - \alpha_j A U_{j-1} \theta + \alpha_j \xi_j U_{j-1} \theta$$

- Extend $U_j := [U_{j-1}\ u_j]$ and then:

$$A \begin{bmatrix} U_{j-1} & u_j \end{bmatrix} \begin{bmatrix} K_{j-1} & \alpha_j \theta \\ & \alpha_j \end{bmatrix} = \begin{bmatrix} U_{j-1} & u_j \end{bmatrix} \begin{bmatrix} H_{j-1} & c_j + \alpha_j \xi_j \theta \\ & \alpha_j \xi_j \end{bmatrix}$$

Some notes:

- Choice of the continuation vector can be tricky, needs to be designed in a way to avoid breakdown (cannot just take the last vector as in Krylov subspaces).
- The above procedure needs to be adjusted slightly for infinity poles.

Rational Arnoldi solver for Sylvester equations

We may replicate the solver we had for the polynomial case, just changing the projection space:

- Compute the tensorized rational Krylov subspace

$$\mathcal{U}_{\ell+1} := \mathcal{RK}_{\ell+1}(A, U, \{\xi, \dots, \xi_\ell, \infty\}) \times \mathcal{RK}_{\ell+1}(B^T, V, \{\sigma_1, \dots, \sigma_\ell, \infty\}).$$

- Solve the projected equation $A_\ell Y + Y B_\ell = V_\ell^T C W_\ell$ in the smaller subspace without the infinity pole. Cost: $\mathcal{O}(\ell^3)$!
- Evaluate the residual as

$$\|R_\ell\|_2 = \max\left\{\|e_{\ell+1}^T A_{\ell+1} Y\|_2, \|Y B_{\ell+1} e_{\ell+1}\|_2\right\}$$

- ... or

$$\|R_\ell\|_F = \sqrt{\|e_{\ell+1}^T A_{\ell+1} Y\|_2^2 + \|Y B_{\ell+1} e_{\ell+1}\|_2^2}$$

- The above formulas still work because we added the infinity pole at the end!

Convergence theory

- We know that the projection space we chose contains a very good solution: the one by ADI with the poles and zeros chosen as ξ_j and σ_j.
- However, there is no guarantee that exactly this solution will be extracted from the space.

Theorem (Beckermann)
The rational Galerkin method for Sylvester equations obtains a solution $\tilde{X}_\ell$ that satisfies:

$$\|A\tilde{X}_\ell + \tilde{X}_\ell B - C\|_2 \leq C\|AX_\ell + X_\ell B - C\|_2,$$

where X_ℓ is any solution given by ADI with the chosen poles, and
$C \approx \max\{|\lambda_i(A) - \lambda_j(-B)|\} / \min\{|\lambda_i(A) - \lambda_j(-B)|\}$ *(for normal matrices).*

Convergence theory (continued)

The message appears to be that this method is almost as good as ADI, but indeed it can be made more specific.

Theorem
In the rational Galerkin method as before, with normal matrices A and B, it holds

$$\|A\tilde{X}_\ell + \tilde{X}_\ell B - C\|_2 \le C \min_{r_A(z) \text{ with poles } \xi_j} \frac{\max_{z\in\Lambda(A)} |r_A(z)|}{\min_{z\in\Lambda(B)} |r_A(-z)|} + C \min_{r_B(z) \text{ with poles } \sigma_j} \frac{\max_{z\in\Lambda(B)} |r_B(z)|}{\min_{z\in\Lambda(A)} |r_B(-z)|}$$

where C is a moderate multiple of $\max\{|\lambda_i(A) - \lambda_j(-B)|\} / \min\{|\lambda_i(A) - \lambda_j(-B)|\}$.

This solves one problem of ADI: if the parameters are chosen inaccurately, the method might converge slowly (or not converge). Rational Galerkin has automatic optimization of the numerator built-in, which largely solves the issue.

Rational Krylov: retrieving the poles

Given a RAD decomposition, we can read the poles by taking the ratio of the subdiagonal elements in the pencil $H - \lambda K$:

$$AU_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times \\ & \alpha_2 & \times & \times \\ & & \alpha_3 & \times \\ & & & \alpha_4 \end{bmatrix} = U_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \beta_1 & \times & \times & \times \\ & \beta_2 & \times & \times \\ & & \beta_3 & \times \\ & & & \beta_4 \end{bmatrix}$$

The poles in this examples are: $\frac{\beta_j}{\alpha_j}$, for $j = 1, \ldots, 4$.

Since we need the last pole to be infinity (for checking the residual), we will have to choose $\alpha_\ell = 0$.

Reordering the poles

Classical trick with rotation. Consider the following upper triangular matrices:

$$T = \begin{bmatrix} \alpha_1 & \times \\ & \alpha_2 \end{bmatrix}, \qquad S = \begin{bmatrix} \beta_1 & \times \\ & \beta_2 \end{bmatrix}$$

Then, considering the pencil $T - \lambda S$, we can construct unitary matrices Q, Z such that QTZ^H and QSZ^H have the diagonal entries swapped (up to rescaling):

- Construct $M = \beta_1 T - \alpha_1 S$. Notice that this matrix has rank 1, and its row span is orthogonal to e_1^T, the left eigenvector of α_1/β_1.
- Compute a rotation G such that

$$GM = \begin{bmatrix} \times & \times \\ 0 & 0 \end{bmatrix}$$

- Construct a rotation H such that GTH has a zero in position $(2, 1)$. Then also S does, and the diagonal entries are swapped (Why?).

Reordering the poles

We can use the trick to reorder the poles in our RAD; note that this does not change the space, but reorders the basis; for checking the residual, we need our infinity poles at the end!

For instance, consider rotations G, H acting on rows 2 and 3, computed with the matrices

$$\begin{bmatrix} \alpha_2 & \times \\ & \alpha_3 \end{bmatrix}, \qquad \begin{bmatrix} \beta_2 & \times \\ & \beta_3 \end{bmatrix}.$$

Then,

$$AU_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times \\ & \alpha_2 & \times & \times \\ & & \alpha_3 & \times \\ & & & \alpha_4 \end{bmatrix} = U_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \beta_1 & \times & \times & \times \\ & \beta_2 & \times & \times \\ & & \beta_3 & \times \\ & & & \beta_4 \end{bmatrix}$$

Reordering the poles

We can use the trick to reorder the poles in our RAD; note that this does not change the space, but reorders the basis; for checking the residual, we need our infinity poles at the end!

For instance, consider rotations G, H acting on rows 2 and 3, computed with the matrices

$$\begin{bmatrix} \alpha_2 & \times \\ & \alpha_3 \end{bmatrix}, \qquad \begin{bmatrix} \beta_2 & \times \\ & \beta_3 \end{bmatrix}.$$

Then,

$$AU_{j+1}G^HG \begin{bmatrix} \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times \\ & \alpha_2 & \times & \times \\ & & \alpha_3 & \times \\ & & & \alpha_4 \end{bmatrix} H = U_{j+1}G^HG \begin{bmatrix} \times & \times & \times & \times \\ \beta_1 & \times & \times & \times \\ & \beta_2 & \times & \times \\ & & \beta_3 & \times \\ & & & \beta_4 \end{bmatrix} H$$

Reordering the poles

We can use the trick to reorder the poles in our RAD; note that this does not change the space, but reorders the basis; for checking the residual, we need our infinity poles at the end!

For instance, consider rotations G, H acting on rows 2 and 3, computed with the matrices

$$\begin{bmatrix} \alpha_2 & \times \\ & \alpha_3 \end{bmatrix}, \qquad \begin{bmatrix} \beta_2 & \times \\ & \beta_3 \end{bmatrix}.$$

Then,

$$AU_{j+1}G^H \begin{bmatrix} \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times \\ & \alpha_3 & \times & \times \\ & & \alpha_2 & \times \\ & & & \alpha_4 \end{bmatrix} = U_{j+1}G^H \begin{bmatrix} \times & \times & \times & \times \\ \beta_1 & \times & \times & \times \\ & \beta_3 & \times & \times \\ & & \beta_2 & \times \\ & & & \beta_4 \end{bmatrix}$$

Reordering the poles

We can use the trick to reorder the poles in our RAD; note that this does not change the space, but reorders the basis; for checking the residual, we need our infinity poles at the end!

For instance, consider rotations G, H acting on rows 2 and 3, computed with the matrices

$$\begin{bmatrix} \alpha_2 & \times \\ & \alpha_3 \end{bmatrix}, \qquad \begin{bmatrix} \beta_2 & \times \\ & \beta_3 \end{bmatrix}.$$

Then,

$$A\tilde{U}_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times \\ & \alpha_3 & \times & \times \\ & & \alpha_2 & \times \\ & & & \alpha_4 \end{bmatrix} = \tilde{U}_{j+1} \begin{bmatrix} \times & \times & \times & \times \\ \beta_1 & \times & \times & \times \\ & \beta_3 & \times & \times \\ & & \beta_2 & \times \\ & & & \beta_4 \end{bmatrix}$$

Reordering the poles

- Reordering the poles changes the matrix H, K.
- It also updates the matrix $\tilde{U}_j$ to have the infinity poles at the end, for instance.
- In this way, we can run the algorithm and check convergence at a minimal cost.

A few more details: the projected matrix in this case is given by

$$AU_{j+1}K_j = U_{j+1}H_j \implies AU_j\tilde{K}_j = U_j\tilde{H}_j \implies U_j^T AU_j = \tilde{H}_j\tilde{K}_j^{-1},$$

where $\tilde{K}_j, \tilde{H}_j$ are the Hessenberg matrices with the last row removed.

If we wish, we could avoid inverting $\tilde{K}_j$ by projecting down to a generalized Sylvester equation (little changes, but $\tilde{K}_j$ is usually well-conditioned).

Take home message (part 3)

We have seen three methods for solving matrix equations:

- Polynomial Krylov is easy to formulate, works OK for well-conditioned matrix equations.
- ADI very powerful, especially a a theoretical tool to predict singular values decay in the solution.
- rational Galerkin takes the best of both worlds: very fast convergence, and less fragile than ADI.

Sylvester equations and low-rank structure

- We have observed that solutions of low-rank Sylvester equations have usually numerical low rank.
- The solvers for Sylvester equations are good candidates for being low-rank approximation methods: fast and guaranteed convergence under suitable hypotheses.
- We now see that several matrices of interest in applications satisfy particular Sylvester equations.

Recall that we have a bound for the decay of Zolotarev numbers in case of symmetric real intervals:

$$Z_\ell([-b,-a],[a,b]) \leq 4\rho^\ell, \qquad \rho := e^{-\frac{\pi^2}{\log(4\frac{b}{a})}}.$$

Extending the bound to other intervals

We may consider the Zolotarev minimax problem on more general domains:

$$Z_\ell(E, F) \leq ???$$

If we can get a bound, the following extension of the previous theorem holds:

Theorem
Let $A, -B$ be normal matrices with spectra contained in E, F, respectively, and X a solution of

$$AX + XB = UV^T, \qquad \operatorname{rank}(UV^T) = k$$

Then,

$$\sigma_{k\ell+1}(X) \leq \sigma_1(X) \cdot Z_\ell(E, F).$$

Möbius transforms

Consider the following rational map:

$$M_A(z) = \frac{\alpha z + \beta}{\gamma z + \delta}, \qquad A := \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix},$$

such that A is invertible. Then, $M_A(z)$ is an automorphism of the Riemann sphere. In particular, note that for any rational function $r(z)$ of degree at most (ℓ, ℓ) $r(M_A(z))$ is still a rational function of degree at most (ℓ, ℓ).

Lemma
Let $A, -B$ be normal matrices with spectra contained in E, F, respectively, with $E \subseteq M_A([a, b])$ and $F \subseteq M_A([-b, -a])$. If X is a solution of

$$AX + XB = UV^T, \qquad \operatorname{rank}(UV^T) = k$$

Then,

$$\sigma_{k\ell+1}(X) \leq \sigma_1(X) \cdot Z_\ell(E, F) = Z_\ell([-b, -a], [a, b]) \leq 4\rho^\ell,$$

where $\rho = e^{-\frac{\pi^2}{\log(4\frac{b}{a})}}$.

Matrices with displacement structure

In the context of structured matrices, Sylvester equations are called displacement equations. We say that X has displacement rank k if

$$AX + XB = UV^T, \qquad \operatorname{rank}(UV^T) = k,$$

for some A, B.

- The point of view is rather different: we are not trying to compute X.
- Indeed, the equation says that X is determined by far less parameters than mn, as a generic $m \times n$ matrix.
- If:
  1. A, B normal
  2. The spectra of A and $-B$ are well separated.

  Then, the matrix X will have numerically low-rank.

Example: Toeplitz matrices

A matrices is said to be Toeplitz if its diagonals are constant:

$$T = \begin{bmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_{-m+1} & \dots & a_{-1} & a_0 \end{bmatrix}$$

Let Z be the downshift matrix, that sends e_j into e_{j+1}. Then, any Toeplitz matrix has displacement rank 2:

$$ZT - TZ = \begin{bmatrix} -1 & 0 \\ 0 & a_{-1} \\ \vdots & \vdots \\ 0 & a_{m+2} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ a_1 & 0 \\ \vdots & \vdots \\ a_{n-2} & 0 \end{bmatrix}^T$$

Example: Toeplitz matrices

Does the previous result mean that Toeplitz matrices have low numerical rank?

- The identity matrix is Toeplitz, so this subtly suggests that no, Toeplitz matrices do not have numerically low-rank (in general).
- Let us look at the matrices defining the displacement relation:
  - Z is singular, and indeed has all eigenvalues equal to 0: the spectrum of Z and $-(-Z)$ are not well separated.
  - Z is also horribly non-normal.

This examples essentially puts together all the things that can go wrong in our argument. However, there are many structured matrices for which this approach succeed.

Another example: Cauchy matrices

Let x, y be two vectors with length m, n. The Cauchy matrices associated with these vectors is defined as:

$$[C(x,y)]_{ij} = \frac{1}{x_i + y_j}, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n.$$

Lemma
A Cauchy matrix $C(x, y)$ solves the displacement equation

$$D_x C(x,y) + D_y C(x,y) = ee^T,$$

where $e = [1 \ldots 1]^T$ is the vector of all ones, and D_x and D_y are the diagonal matrices with x and y on the diagonal.

Another example: Cauchy matrices

Are the hypothese verified in this case?

- Are the matrices D_x, D_y normal? Yes!
- Are the spectra of D_x and $-D_y$ well separated? This is only true if the vectors x and $-y$ are included in disjoint intervals.

Theorem
Let $C(x, y)$ be a Cauchy matrix with real x, y, and such that $x_j \in [a, b]$ and $-y_j \in [c, d]$, with $[a, b] \cap [c, d] = \emptyset$. Then,

$$\sigma_{j+1}(C(x,y)) \leq \|C(x,y)\|_2 \cdot \rho^{\ell}, \qquad \rho := e^{-\frac{\pi^2}{\log 4\kappa}}, \qquad \kappa = \sqrt{\left|\frac{(c-a)(d-b)}{(c-b)(d-a)}\right|}.$$

See: `example_cauchy.m`

Krylov matrices

One may think of considering as basis for the Krylov subspace $\mathcal{K}_\ell(A, b)$ the matrix:

$$K := \begin{bmatrix} b & Ab & \dots & A^{\ell-1}b \end{bmatrix}.$$

- Is it a good basis? If it's not, how bad it is?
- Everybody think it's not, and indeed we use the Arnoldi process to build a much better one (which is orthogonal).

We will see that it can be "exponentially bad".

Note: By a "good basis", we mean a well-conditioned basis, with singular values close to 1; such a basis would imply that if we write

$$y = Kx \implies \|y\|_2 \approx \|x\|_2,$$

and therefore perturbation analysis on x can be transferred to y.

Lemma
Krylov matrices generated by A satisfy the displacement relation

$$AK - KQ = (A^\ell + I)be_\ell^T,$$

where

$$Q = \begin{bmatrix} & & & -1 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}, \qquad K := \begin{bmatrix} b & Ab & \dots & A^{\ell-1}b \end{bmatrix}.$$

Proof.
Direct check. □

What does this say on the decay of the singular values? Clearly, it depends on A.

Vandermonde matrices are Krylov matrices

For interpolating a polynomial, we often need to solve a linear system

$$Vx = f, \qquad V := \begin{bmatrix} 1 & x_1 & \dots & x^{\ell-1} \\ \vdots & & \vdots & \\ 1 & x_n & \dots & x_n^{\ell-1} \end{bmatrix}$$

We note that V is nothing else the the Krylov matrix with:

$$A = D_x, \qquad b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

It is rather well-known that Vandermonde matrices can be horribly ill-conditioned, but that's not always true. If, for instance, we consider the case $x_j = \omega_n^j$, where the nodes are the roots of the unity, then V is unitary, and therefore all the singular values are one.

What does our result say in that case?

Fourier matrix and normal matrices

For the Fourier matrix, observe that the eigenvalues of D_x are the roots of the unity, exactly the same of the eigenvalues of Q[2]! Therefore, our theorem gives us the trivial bound:

$$\sigma_\ell(V) \leq \sigma_1(V) \cdot 1$$

It turns out, that in this case is sharp: all the singular values are one.

We shall consider another case: what happens if the matrix A generating the Krylov matrix is Hermitian? Then,

$$AK - KQ$$

is a Sylvester equation with normal matrices. If we assume that n is even (not really restrictive!) then A and Q have disjoint spectra.

[2] Assuming we choose the top-right entry to be 1. Otherwise, they are shifted roots of the unity, they do not coincide, but are completely interlaced.

Theorem
If A is Hermitian, and n is even, then the K Krylov matrix with the vectors $A^{j-1}b$ as columns satisfies

$$\sigma_{1+2j}(K) \leq 4\rho^j \cdot \|K\|, \qquad \rho := e^{\frac{-\pi^2}{4\log(8\frac{n}{2\pi})}}$$

Proof.
Complicated — at the blackboard if we have time. □

Positive definite Hankel matrices

Theorem
Any real $n \times n$ positive semi-definite Hankel matrix H satisfies

$$\sigma_{1+2\ell}(H) \leq 16\rho^{\ell+1}, \qquad \rho := e^{\frac{-\pi^2}{4\log(8\frac{n}{2\pi})}}.$$

Proof.
H is Hankel and positive semidefinite if and only if: $H_{ij} = \int_{-\infty}^{\infty} x^{i+j-2} d\mu(x)$, for some nonnegative measure $d\mu(x)$. Let w, x be Gauss quadrature weights and nodes of order n associated with $d\mu(x)$. Since they integrate exactly polynomials of degree up to $2n-1$, we have:

$$H_{ij} = \sum_{s=1}^{n} w_s x_s^{i+j-2} = \sum_{s=1}^{n} (\sqrt{w_s} x^{i-1})(\sqrt{w_s} x^{j-1}).$$

Therefore, $H = K^H K$, where K is the Krylov matrix with $\mathrm{diag}(x)$ and initial vector w. In particular $\sigma_{1+\ell}(H) = \sigma_{1+\ell}(K)^2$. □