Potential Theory and Practical Aspects of the Solution of Lyapunov Equations

Chris Beattie, Virginia Tech
Mark Embree, Rice University
John Sabino, Boeing

Computational Methods with Applications

Harrachov

21 August 2007

Overview

We wish to solve large Lyapunov equations via low-rank Smith/ADI methods. The adaptation of such methods from 'medium scale' to 'large scale' problems benefits from an understanding of the solutions (via potential theory) and numerous algorithmic improvements that enhance performance.

- Introduction
  - Applications of Lyapunov and Sylvester equations
  - Balanced truncation model reduction
  - Nonnormality and the state matrix
- Potential Theory and Decay of Singular Values
  - ADI approximations
  - Convergence bounds
  - Condenser capacity and Bagby points
- Practical Aspects of Lyapunov Solvers
  - Alternatives to asymptotically optimal ADI points
  - Other algorithmic improvements
  - Numerical results

Applications of Lyapunov and Sylvester equations

- Eigenvalue perturbation theory
  Sylvester operator arises naturally in perturbation theory for invariant subspaces, block diagonalizations, etc.; [Stewart 1973], [Higham 2002]
- Total Energy computation for dynamical systems
  Given $x'(t) = Ax(t)$, $x(0) = x_0$, with solution $x(t) = \mathrm{e}^{tA}x_0$.
  For some h.p.d. matrix E, the *total energy* of the system is

$$\mathcal{E}(x_0) = \int_0^\infty x_0^* e^{tA^*} E e^{tA} x_0 \mathrm{d}t = x_0^* \left(\int_0^\infty e^{tA^*} E e^{tA} \mathrm{d}t \right) x_0.$$

  Notice that $X := \int_0^\infty e^{tA^*} E e^{tA} \, \mathrm{d}t$ satisfies

$$A^*X + XA = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t} \left(\mathrm{e}^{tA^*} E \mathrm{e}^{tA} \right) \mathrm{d}t = -E.$$

- Control theory
  - Many applications . . .

Model reduction via balanced truncation

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t), \quad x(0) = x_0 \end{aligned}$$

The infinite reachability and observability gramians

$$P := \int_0^\infty \mathrm{e}^{tA} BB^* \mathrm{e}^{tA^*} \,\mathrm{d}t, \qquad Q := \int_0^\infty \mathrm{e}^{tA^*} C^* C \mathrm{e}^{tA} \,\mathrm{d}t$$

are Hermitian positive semi-definite (integrals of HPsD matrices) and can be characterized as the solutions to the Lyapunov equations

$$AP + PA^* = -BB^*, \qquad A^*Q + QA = -C^*C.$$

If $x_0 = 0$, the minimum energy of u required to drive x to state $\widehat{x}$ is

$$\widehat{x}^* P^{-1} \widehat{x}.$$

Starting from $x_0 = \widehat{x}$ with $u(t) \equiv 0$, the energy of output y is

$$\widehat{x}^* Q \widehat{x}.$$

Internal coordinates

$\widehat{x}^* P^{-1} \widehat{x}$: $\widehat{x}$ is hard to reach if it is rich in the lowest modes of P.
$\widehat{x}^* Q \widehat{x}$: $\widehat{x}$ is hard to observe if it is rich in the lowest modes of Q.

Balanced truncation transforms the internal coordinates to align
states that require much energy to reach
with
states that produce little output energy.

$$\begin{aligned}(Sx)'(t) &= (SAS^{-1})(Sx(t)) + (SB)u(t)\\ y(t) &= (CS^{-1})(Sx(t)) + Du(t), \qquad (Sx)(0) = Sx_0.\end{aligned}$$

With this transformation, the reachability and observability gramians are

$$\widehat{P} = SPS^*, \qquad \widehat{Q} = S^{-*} Q S^{-1}.$$

Balanced truncation effectively constructs an S so that $\widehat{P} = \widehat{Q}$.

Balancing transformations

$$\{s_j\} = \{\sqrt{\text{eigs of } PQ}\}: \qquad \Sigma = \operatorname{diag}(s_1, \ldots, s_n)$$

Cholesky factorization: $P = UU^*$

Cholesky factorization: $Q = LL^*$

Hermitian eigenvalue decomposition: $U^*QU = K\Sigma^2K^*$

singular value decomposition: $L^*U = V\Sigma W^*$

Suitable choices for balancing S:

$$S_1 = UK\Sigma^{-1/2} \qquad S_2 = UW\Sigma^{-1/2}.$$

Reduced system: upper left $k \times k$ part of the system in new coordinates.

Error bound: $\|H(z) - \widehat{H}(z)\|_{\mathcal{H}_\infty} \leq 2(s_{k+1} + s_{k+2} + \cdots + s_n)$.

The s_j are the Hankel singular values.

See [Antoulas 2005] for details and comments on numerics.

Invariance of internal coordinates

Hankel singular values:	square roots of eigenvalues of PQ
Transfer function:	$D + C(z - A)^{-1}B$
Markov parameters:	$D,\ CB,\ CAB,\ CA^2B,\ \ldots$

All of these quantities are independent of the internal coordinates.

For example, $\widehat{P}\,\widehat{Q} = (SPS^*)(S^{-*}QS^{-1}) = SPQS^{-1}$.

Hence the Hankel singular values—and the error bound—are independent of the state space coordinate system.

If A is stable, the reduced model is *stable* regardless of the internal coordinate.

Do the internal coordinates affect balancing?

Balancing transformation requires computation of

$$P = UU^*, \qquad Q = LL^*.$$

Coordinate transformations do not change the eigenvalues of PQ, but they do change the eigenvalues of P and Q individually:

$$\widehat{P} = SPS^*, \qquad \widehat{Q} = S^{-*}QS^{-1}.$$

These are *congruence transformations* (preserve symmetry, inertia), as opposed to *similarity transformations* (preserve eigenvalues).

What can be said of the spectral properties of P and Q?

State variables and moment matching model reduction

- The challenge of solving large-scale Lyapunov equations is an impediment to balanced truncation model reduction.
- A popular alternative are Krylov-based *moment-matching* methods. For example, use the Arnoldi method to compute the factorization

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^*, \qquad V_k^* V_k = I.$$

The reduced system takes the form

$$A_k = V_k^* A V_k, \qquad B_k = V_k^* B, \qquad C_k = C V_k.$$

The moments (Markov parameters) of the original system are

$$CB, \quad CAB, \quad CA^2B, \quad \ldots.$$

k-step Arnoldi reduction matches k moments:

$$C_k B_k = CB, \quad C_k A_k B_k = CAB, \quad \ldots, \quad C_k A_k^{k-1} B_k = CA^{k-1} B.$$

k-step bi-orthogonal Lanczos reduction matches $2k$ moments.

Moment matching: role of internal representation

The moments are invariant to the internal representation, but the reduced system is not.

State-space coordinates can significantly affect the Arnoldi algorithm.

If the numerical range of A contains points in the right half plane, it is possible that the reduced model will not even be stable.

$$\begin{aligned}(Sx)'(t) &= (SAS^{-1})(Sx(t)) + (SB)u(t)\\ y(t) &= (CS^{-1})(Sx(t)) + Du(t), \qquad (Sx)(0) = Sx_0.\end{aligned}$$

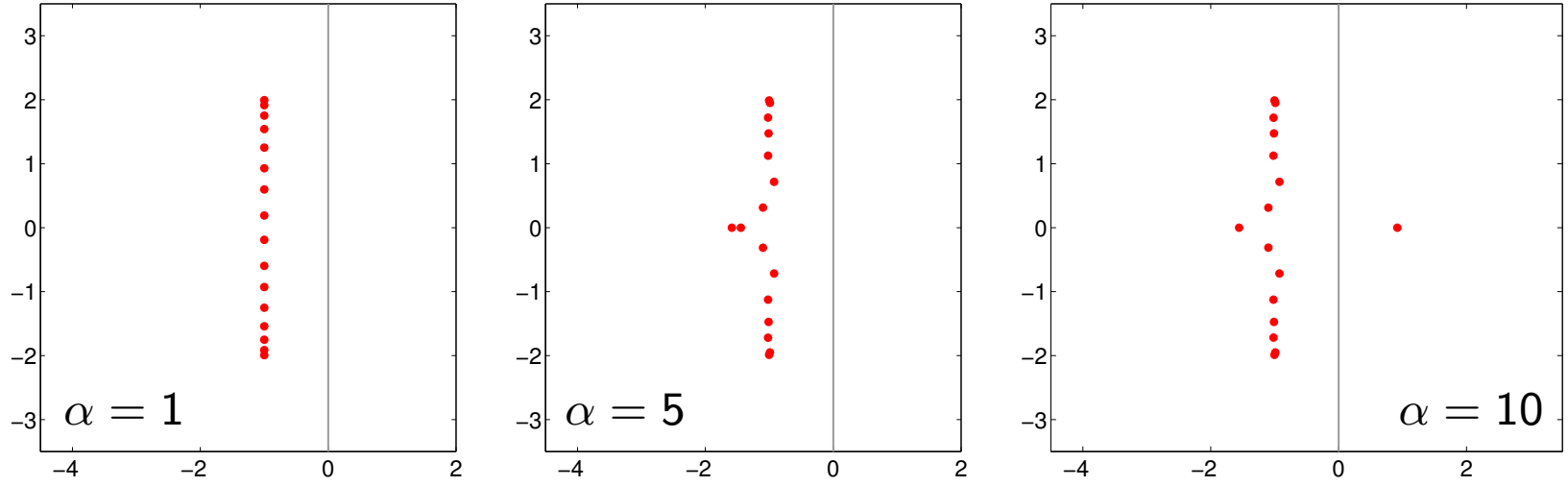
Elementary example: tridiagonal Toeplitz matrix

Take $S_\alpha = \mathrm{diag}(\alpha, \alpha^2, \ldots, \alpha^n)$, and

$$A_1 = \begin{bmatrix} -1 & -1 & & \\ 1 & -1 & \ddots & \\ & \ddots & \ddots & -1 \\ & & 1 & -1 \end{bmatrix}, \quad A_\alpha = S_\alpha A_1 S_\alpha^{-1} = \begin{bmatrix} -1 & -\alpha^{-1} & & \\ \alpha & -1 & \ddots & \\ & \ddots & \ddots & -\alpha^{-1} \\ & & \alpha & -1 \end{bmatrix}.$$

Then $\sigma(A_\alpha) \in [-1-2\mathrm{i}, -1+2\mathrm{i}]$, but $\kappa(S_\alpha) = \max\{|\alpha|^n, |\alpha|^{-n}\}$.

Eigenvalues of reduced model A_{32} for $n = 256$:



Algorithms for Lyapunov and Sylvester equations

$$AX - XB = C, \quad A \in \mathbf{C}^{n\times n}, B \in \mathbf{C}^{m\times m}$$

- Can be formulated as an *nm*-by-*nm* linear system:

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

  which requires $O(n^3m^3)$ flops to solve. The spectrum of this matrix,

$$\sigma\Big((I \otimes A) - (B^T \otimes I)\Big) = \Big\{\lambda_j - \mu_j : \lambda_j \in \sigma(A),\ \mu_j \in \sigma(B)\Big\},$$

  shows that a unique solution X exists if an only if $\sigma(A) \cap \sigma(B) = \emptyset$.
- Dense methods need Schur factorization of A, B, $O(n^3 + m^3)$ flops. [Bartels, Stewart], [Hammarling], [Sorensen, Zhou]
- Numerous iterative approaches:
  Smith/ADI methods [Smith; Wachspress; Penzl; Li & White; etc.]
  Krylov, rational Krylov methods [Saad; Simoncini]
  "Approximate power iteration" [Hodel, Poolla, Tenison; Sorensen]

The ADI iteration for $AX - XB = C$

- The conventional ADI method [Peaceman & Rachford 1955] was designed to solve $(\mathcal{H} + \mathcal{V})x = b$, where $\mathcal{H}$ and $\mathcal{V}$ commute; see [Wachspress 1966].
- Set $\mathcal{H}X := AX$, $\mathcal{V}X := -XB$, with $\mathcal{H}(\mathcal{V}X) = -AXB = \mathcal{V}(\mathcal{H}X)$ [Ellner & Wachspress 1986], [Wachspress 1988].
- Written as two stages:

$$\begin{aligned}(A + p_k)X_{k+1/2} &= X_k(B + p_k) + C \\ X_{k+1}(B - q_k) &= (A - q_k)X_{k+1/2} - C.\end{aligned}$$

- Written as one stage:

$$\begin{aligned}X_{k+1} = {} & (A - q_k)(A + p_k)^{-1}X_k(B + p_k)(B - q_k)^{-1} \\ & + ((A - q_k)(A + p_k)^{-1} - I)C(B - q_k)^{-1}.\end{aligned}$$

- Exact solution is a fixed point, which gives an error formula, bounds...

Error bounds for ADI iteration

- Error formula:

$$X - X_k = \left(\prod_{j=0}^{k-1} (A - q_j)(A + p_j)^{-1} \right) (X - X_0) \left(\prod_{j=0}^{k-1} (B + p_j)(B - q_j)^{-1} \right).$$

- Define the rational function

$$\phi_k(z) := \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j},$$

  so that

$$X - X_k = \phi_k(A)(X - X_0)\phi_k(B)^{-1}.$$

- With $X_0 = 0$ we have

$$\frac{\|X - X_k\|}{\|X\|} \leq \|\phi_k(A)\| \|\phi_k(B)^{-1}\|,$$

  see, e.g., [Wachspress 1966].

ADI convergence theory

- Considerable work on ADI for Lyapunov/Sylvester equations in early 1990s: [Starke 1989, 1991, 1993a, 1993b], [Levenberg & Reichel 1993], ...
- $X \in \mathbf{C}^{n \times m}$ is typically dense, even when A, B are sparse.

ADI convergence theory

- Considerable work on ADI for Lyapunov/Sylvester equations in early 1990s: [Starke 1989, 1991, 1993a, 1993b], [Levenberg & Reichel 1993], ...

- $X \in \mathbf{C}^{n\times m}$ is typically dense, even when A, B are sparse.

- However, with $X_0 = 0$, the formula

$$\begin{aligned} X_{k+1} \;=\; & (A - q_k)(A + p_k)^{-1} X_k (B + p_k)(B - q_k)^{-1} \\ & + ((A - q_k)(A + p_k)^{-1} - I) C (B - q_k)^{-1}. \end{aligned}$$

  indicates that

$$\operatorname{rank}(X_k) \;\leq\; k \operatorname{rank}(C).$$

  [Penzl 2000]

- When C has low rank, as typical for balanced truncation model reduction (e.g., SISO systems), X may be well approximated by low-rank matrices.

Decay rate for singular values

- Let $r = \text{rank}(C)$, and recall $\phi_k(z) := \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j}$.

  Error bound, rank of X_k give

$$\frac{\sigma_{kr+1}(X)}{\sigma_1(X)} \leq \|\phi_k(A)\| \|\phi_k(B)^{-1}\|.$$

Decay rate for singular values

- Let $r = \text{rank}(C)$, and recall $\phi_k(z) := \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j}$.

  Error bound, rank of X_k give

  $$\frac{\sigma_{kr+1}(X)}{\sigma_1(X)} \leq \|\phi_k(A)\| \|\phi_k(B)^{-1}\|.$$

- Pick compact sets $\Omega_A, \Omega_B \subset \mathbf{C}$ with $\sigma(A) \subset \Omega_A$, $\sigma(B) \subset \Omega_B$, and let $\kappa(\Omega; Z)$ denote the smallest constant for which

  $$\|f(Z)\| \leq \kappa(\Omega; Z) \max_{z \in \Omega} |f(z)|$$

  uniformly over all f analytic on Ω [Beattie, E., Rossi 2004] (related to *k-spectral sets* in operator theory [Paulsen 1986]).

Decay rate for singular values

- Let $r = \text{rank}(C)$, and recall $\phi_k(z) := \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j}$.

  Error bound, rank of X_k give

  $$\frac{\sigma_{kr+1}(X)}{\sigma_1(X)} \leq \|\phi_k(A)\| \|\phi_k(B)^{-1}\|.$$

- Pick compact sets $\Omega_A, \Omega_B \subset \mathbf{C}$ with $\sigma(A) \subset \Omega_A$, $\sigma(B) \subset \Omega_B$, and let $\kappa(\Omega; Z)$ denote the smallest constant for which

  $$\|f(Z)\| \leq \kappa(\Omega; Z) \max_{z \in \Omega} |f(z)|$$

  uniformly over all f analytic on Ω [Beattie, E., Rossi 2004] (related to *k-spectral sets* in operator theory [Paulsen 1986]).

$$\frac{\sigma_{kr+1}(X)}{\sigma_1(X)} \leq \kappa(\Omega_A; A)\kappa(\Omega_B; B) \frac{\max\{|\phi_k(z)| : z \in \Omega_A\}}{\min\{|\phi_k(z)| : z \in \Omega_B\}}.$$

Decay rate for singular values, cont'd

- Choices for Ω_A, Ω_B depend on nonnormality of A, B.
- If $\Omega_A = \sigma(A)$, $\Omega_B = \sigma(B)$, then we can take

$$\kappa(\Omega_A; A) = \|V\|\|V^{-1}\|, \qquad \kappa(\Omega_B; B) = \|U\|\|U^{-1}\|$$

  with diagonalizations $A = V\Lambda V^{-1}$, $B = UMU^{-1}$.
  This bound was established by [Beckermann] in the context of displacement rank; cf. [Levenberg & Reichel 1993].
- For A, B far from normal, better to use the numerical range, polynomial numerical hulls, pseudospectra etc.
- Preferred choice for Ω_A and Ω_B may change with degree k.

Ω_A and Ω_B for nonnormal A, B

- Note that Ω_A and Ω_B must be disjoint for the rational approximation problem to converge.
- If A and $-B$ are both stable (e.g., $B = -A^*$ for Lyapunov eqns.), then there always exist disjoint $\Omega_A \supset \sigma(A)$ and $\Omega_B \supset \sigma(B)$ with finite $\kappa(\Omega_A; A)$, $\kappa(\Omega_B; B)$.
- For many problems, the numerical ranges $W(A)$ and $W(-B)$ have a nontrivial intersection.
- An anomaly not captured by these bounds:
  - Consider the Lyapunov equation $AX + XA^* = -bb^T$.
  - If there is *no decay*, wlog $X = I$, then $A + A^* = -bb^T$.
  - This implies the rightmost eigenvalue of $A + A^*$ is zero.
  - Thus the numerical range $W(A)$ is contained in the closed left half plane.
  - Further increasing nonnormality must *improve* decay of singular values.

Any decay is possible for any spectrum

Theorem. [cf. Penzl 2000]
Let $A \in \mathbf{C}^{n\times n}$, $B \in \mathbf{C}^{m\times m}$ with $\sigma(A) \cap \sigma(B) = \emptyset$,
and $C = cd^*$ with (A, C) reachable and (C, B) observable.
Then for any full-rank $Y \in \mathbf{C}^{n\times m}$, there exist invertible $S \in \mathbf{C}^{n\times n}$ and $T \in \mathbf{C}^{m\times m}$ such that Y solves the Sylvester equation

$$(SAS^{-1})Y - Y(T^{-1}BT) = -(Sc)(d^*T).$$

Any decay is possible for any spectrum

Theorem. [cf. Penzl 2000]
Let $A \in \mathbf{C}^{n\times n}$, $B \in \mathbf{C}^{m\times m}$ with $\sigma(A) \cap \sigma(B) = \emptyset$,
and $C = cd^*$ with (A, C) reachable and (C, B) observable.
Then for any full-rank $Y \in \mathbf{C}^{n\times m}$, there exist invertible $S \in \mathbf{C}^{n\times n}$ and $T \in \mathbf{C}^{m\times m}$ such that Y solves the Sylvester equation

$$(SAS^{-1})Y - Y(T^{-1}BT) = -(Sc)(d^*T).$$

- Given disjoint sets of eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_m\}$ and positive singular values $\{s_1, \ldots, s_{\min\{m,n\}}\}$, there exist $\widehat{A}$, $\widehat{B}$, and rank-1 $\widehat{C}$ such that

$$\begin{aligned} \sigma(\widehat{A}) &= \{\lambda_1, \ldots, \lambda_n\} \\ \sigma(\widehat{B}) &= \{\mu_1, \ldots, \mu_m\}, \\ \text{singular values of } Y &= \{s_1, \ldots, s_{\min\{m,n\}}\}, \end{aligned}$$

  and

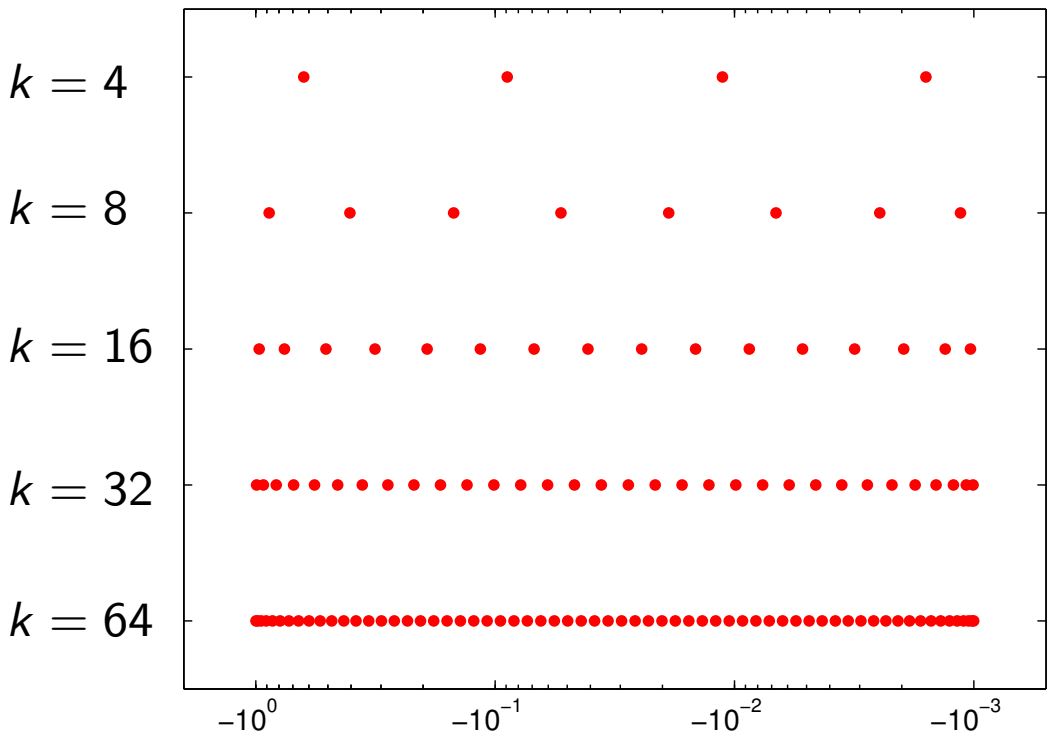
$$\widehat{A}Y - Y\widehat{B} = \widehat{C}.$$

- Specified eigenvalues can be defective but not derogatory.
- Proof is constructive.

Distribution of optimal ADI points, $A = -B = [-1, -10^{-3}]$

Optimal parameters are known in terms of elliptic functions
[Lebedev, *USSR Computational Math. Math. Phys.*, 1977]

Abstract: "THE MAIN results obtained by foreign mathematicians in the theory of optimal parameters in the method of alternating directions are shown to be contained in Zolotarev's results on best bilinear approximations (1877)...."



Rational approximation problem

Let $R_{j,k}$ = set of rational functions of degree (j,k).

- The quantity

$$e_k(\Omega_A,\Omega_B) := \min_{\phi\in R_{k,k}} \frac{\max\{|\phi(z)| : z\in\Omega_A\}}{\min\{|\phi(z)| : z\in\Omega_B\}}.$$

  was introduced and computed in a special case by Zolotarev [1877]; it is known as his "Third Problem".

- Gončar [1969] shows that for compact sets Ω_A, Ω_B with positive capacity, connected complements:

$$\lim_{k\to\infty} e_k(\Omega_A,\Omega_B)^{1/k} = \mathrm{e}^{-1/\mathrm{cap}(\Omega_A,\Omega_B)},$$

  where $\mathrm{cap}(\Omega_A,\Omega_B)$ is the *capacity* of the *condenser* (Ω_A,Ω_B) [Polya & Szego 1962; Bagby 1967; Levin & Saff 1994; Saff & Totik 1997].

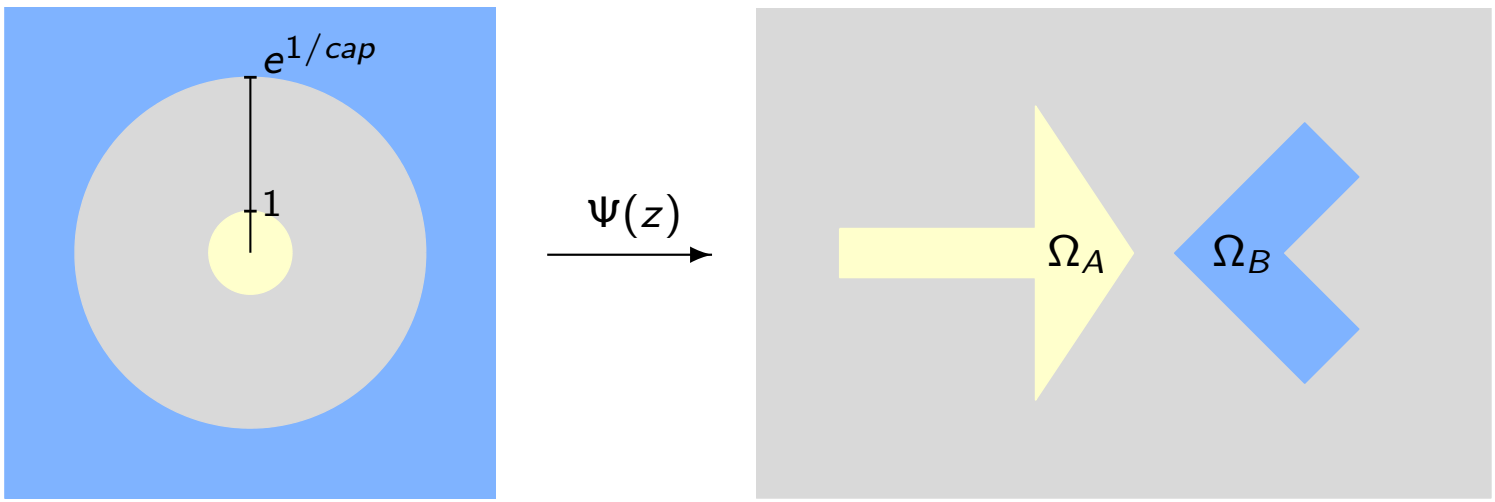
- Generalizations to $R_{j,k}$ with j/k fixed as $j,k\to\infty$ can be useful when $\Omega_B \neq -\Omega_A$ [Levenberg & Reichel 1993, Levin & Saff 1994].

Rational approximation problem

Condenser capacity can be derived from a doubly-connected conformal map $\Psi(z)$ that takes the annulus

$$\{z \in \mathbf{C} : 1 < |z| < \exp(1/\mathrm{cap}(\Omega_A, \Omega_B))\}$$

to the complement of $\Omega_A \cup \Omega_B$.



Could compute Ψ using Hu's DSCPACK [1995];
cf. [Starke 1993]; DeLillo, Driscoll, Elcrat, Pfaltzgraff [2006].

Optimal rational interpolation points

We wish to find parameters $\{p_j\}_{j=0}^{k-1}$ and $\{q_j\}_{j=0}^{k-1}$ for the rational function

$$\phi_k(z) = \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j} \quad \text{to minimize} \quad \frac{\max\{|\phi_k(z)| : z \in \Omega_A\}}{\min\{|\phi_k(z)| : z \in \Omega_B\}}.$$

Asymptotically optimal choices:

- Fejér–Walsh points. Given the conformal map Ψ, set

$$\begin{aligned} q_j &= \Psi(e^{2\pi i j/k}) \in \partial\Omega_A \\ -p_j &= \Psi(e^{2\pi i j/k+1/\mathrm{cap}(\Omega_A,\Omega_B)}) \in \partial\Omega_B \end{aligned}$$

  [Walsh, 1965], [Starke 1993]

Optimal rational interpolation points

We wish to find parameters $\{p_j\}_{j=0}^{k-1}$ and $\{q_j\}_{j=0}^{k-1}$ for the rational function

$$\phi_k(z) = \prod_{j=0}^{k-1} \frac{z - q_j}{z + p_j} \quad \text{to minimize} \quad \frac{\max\{|\phi_k(z)| : z \in \Omega_A\}}{\min\{|\phi_k(z)| : z \in \Omega_B\}}.$$

Asymptotically optimal choices:

- Fejér–Walsh points. Given the conformal map Ψ, set

$$\begin{aligned} q_j &= \Psi(e^{2\pi i j/k}) \in \partial\Omega_A \\ -p_j &= \Psi(e^{2\pi i j/k+1/\mathrm{cap}(\Omega_A,\Omega_B)}) \in \partial\Omega_B \end{aligned}$$

  [Walsh, 1965], [Starke 1993]

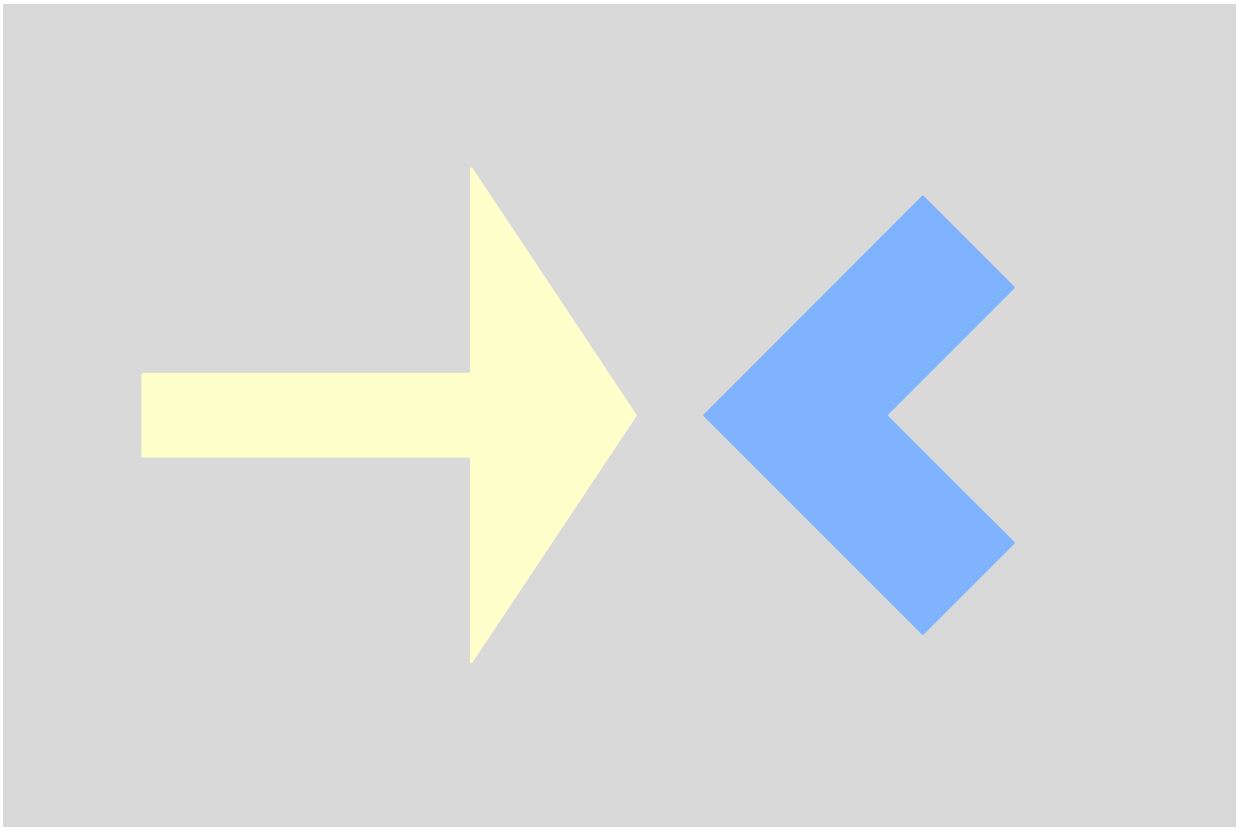
- Leja–Bagby points. Given $\{p_j\}_{j=0}^{\ell-1}$, $\{q_j\}_{j=0}^{\ell-1}$, pick $-p_\ell \in \Omega_B$, $q_\ell \in \Omega_A$ s.t.:

$$\prod_{j=0}^{\ell-1} \frac{|p_\ell + p_j|}{|p_\ell - q_j|} = \max_{-p\in\Omega_B} \prod_{j=0}^{\ell-1} \frac{|p + p_j|}{|p - q_j|}, \qquad \prod_{j=0}^{\ell-1} \frac{|q_\ell - q_j|}{|q_\ell + p_j|} = \max_{q\in\Omega_A} \prod_{j=0}^{\ell-1} \frac{|q - q_j|}{|q + p_j|}.$$

  [Bagby, 1967], [Starke 1991], [Levenberg & Reichel 1993]

Approximation of Leja–Bagby points

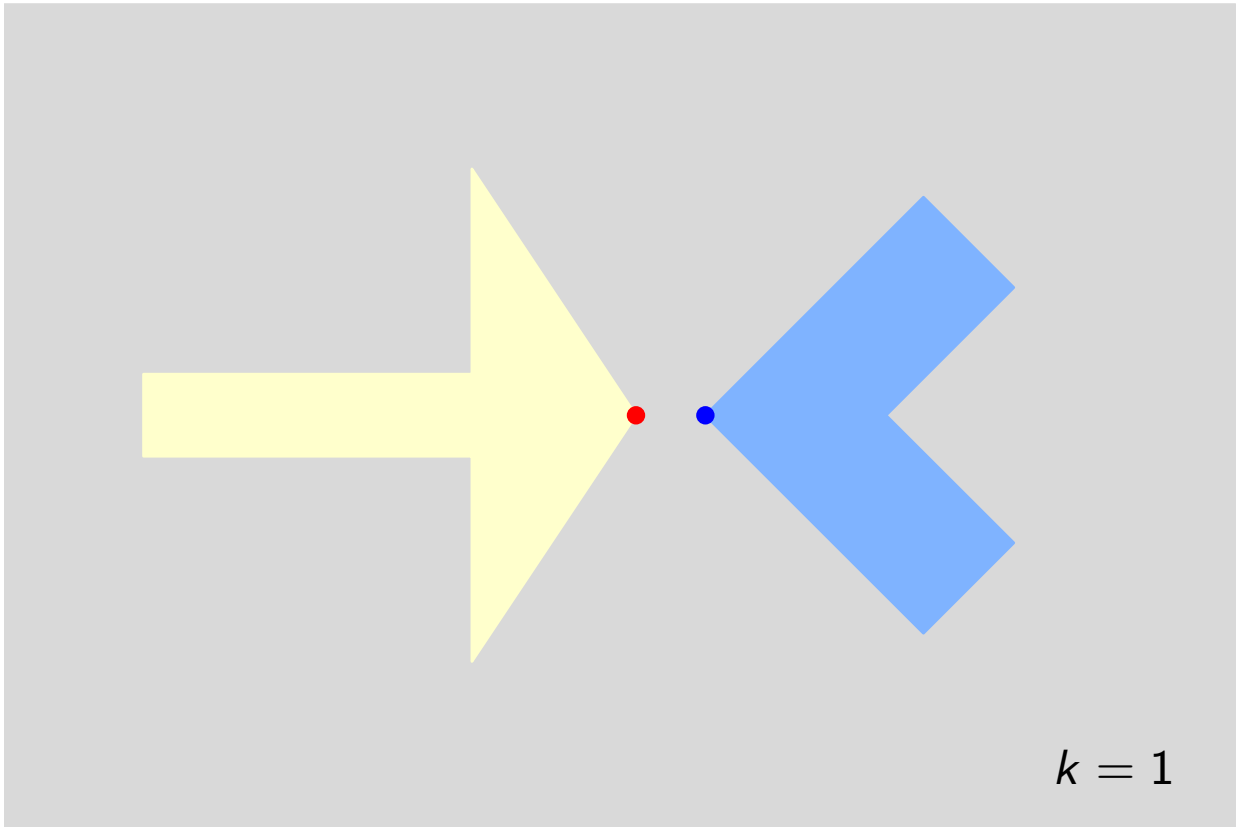
- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



$\bullet = q_j$ $\bullet = -p_j$

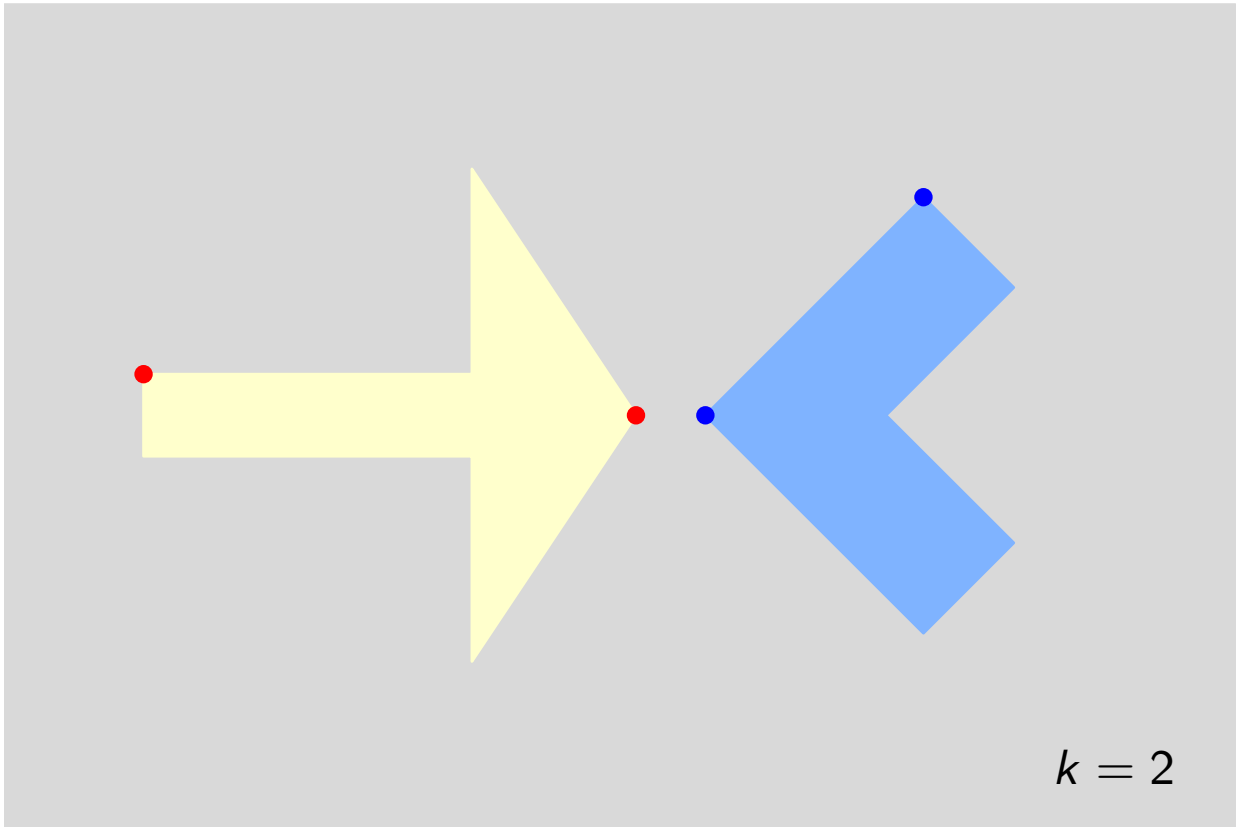
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



Approximation of Leja–Bagby points

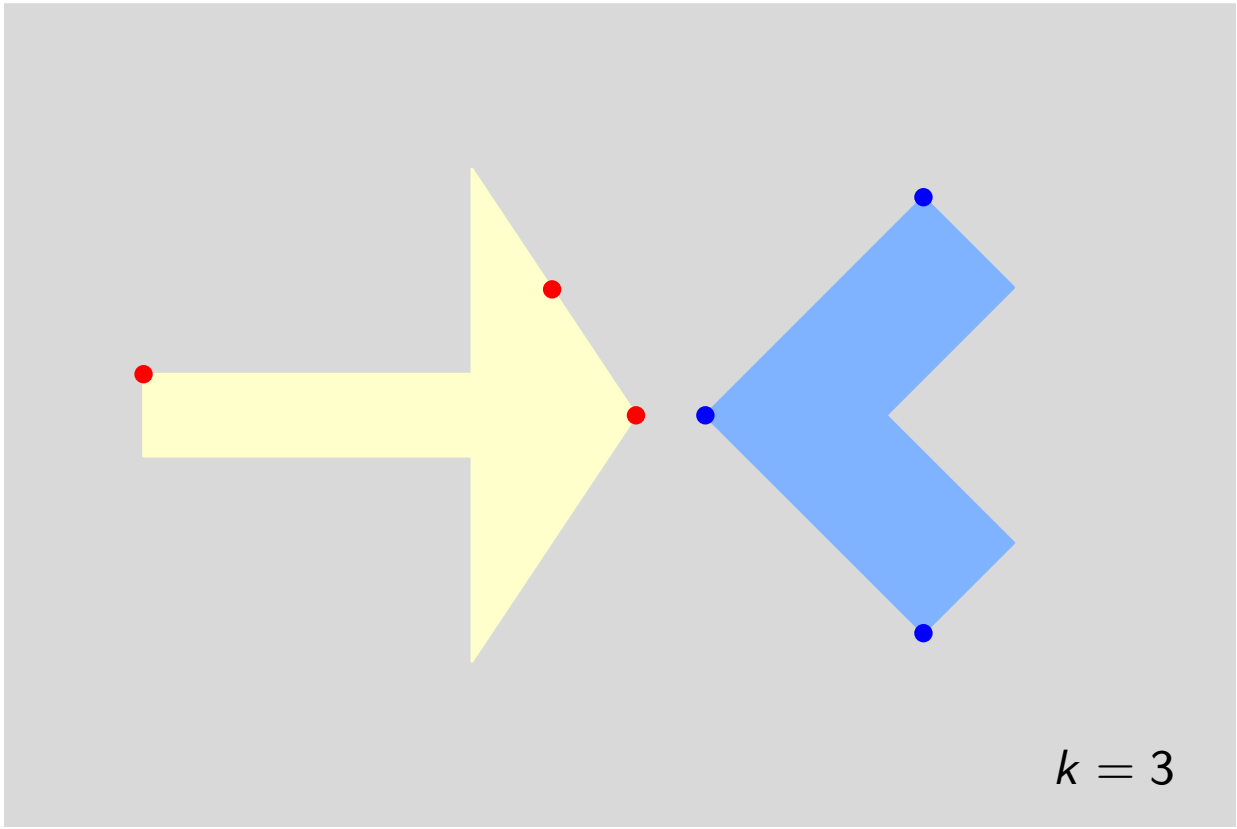
- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



● $= q_j$ ● $= -p_j$

Approximation of Leja–Bagby points

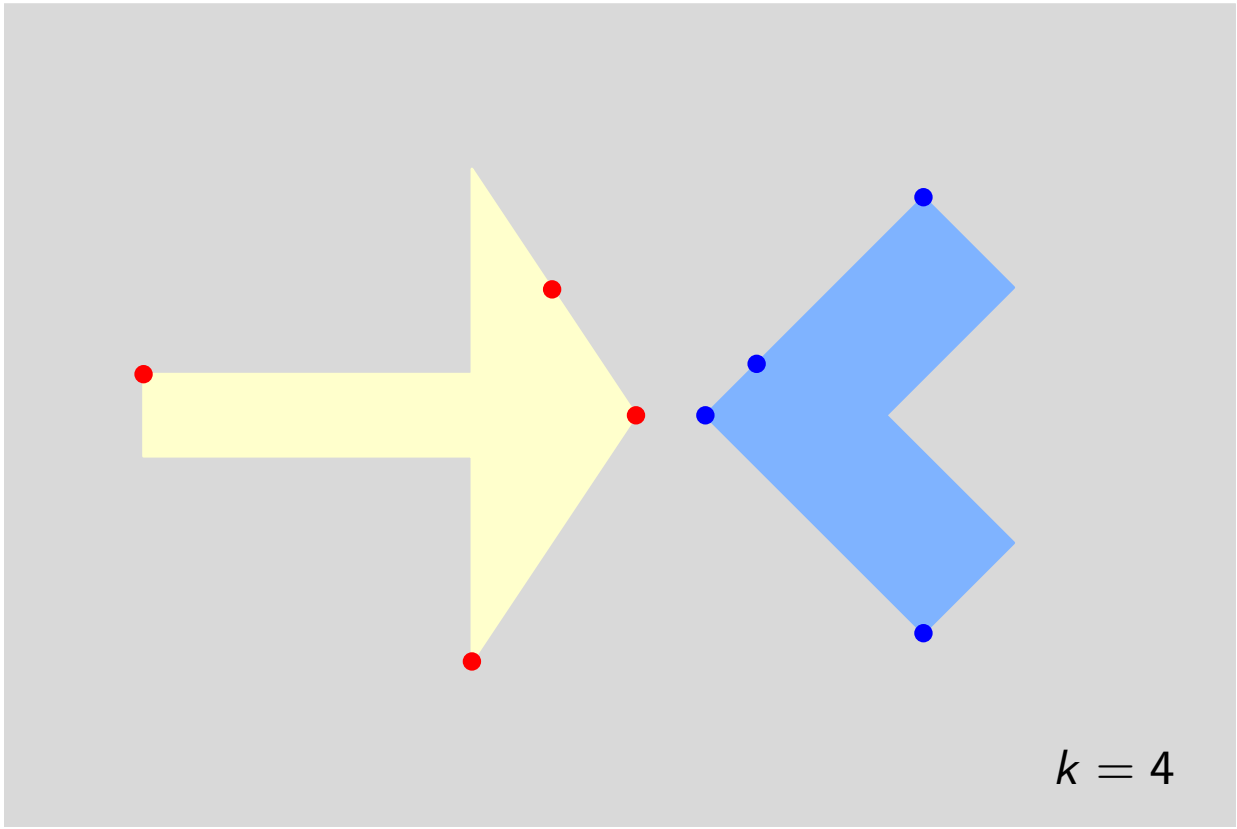
- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



● $= q_j$ ● $= -p_j$

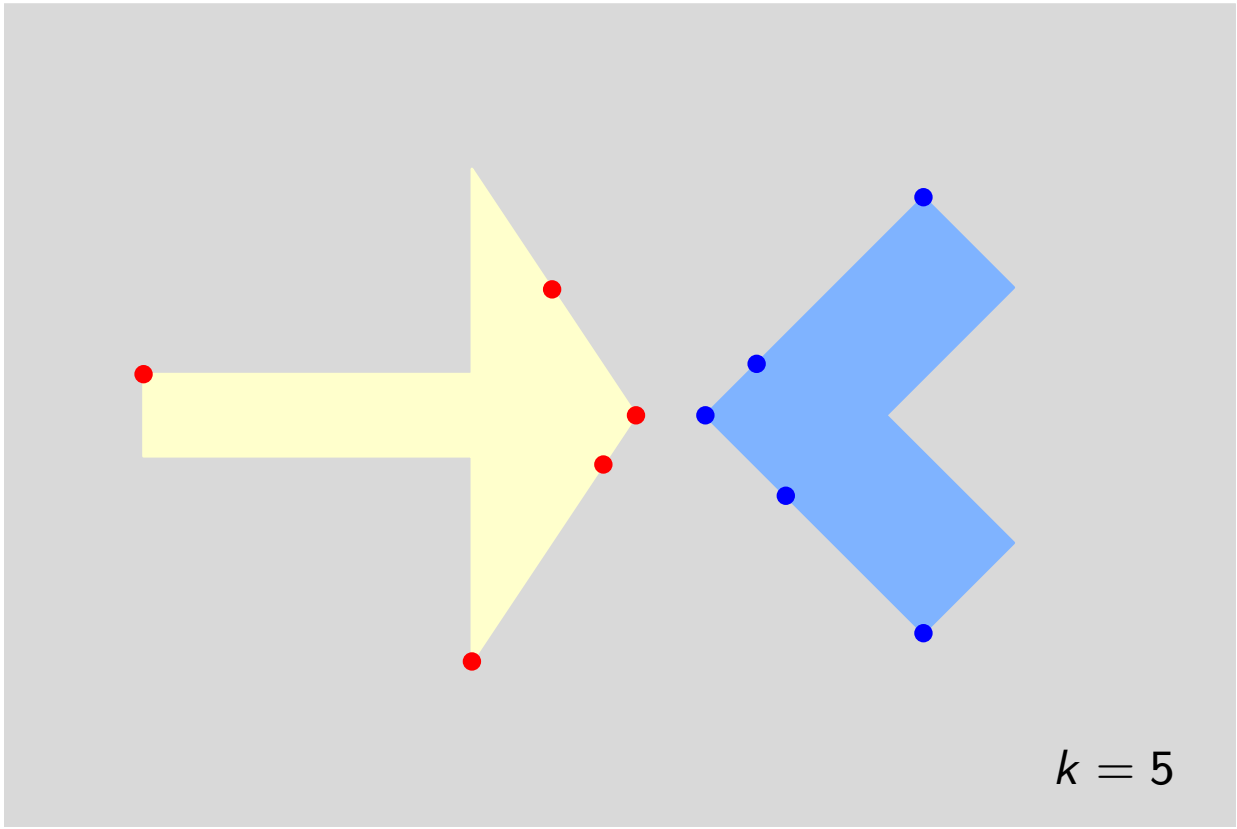
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



Approximation of Leja–Bagby points

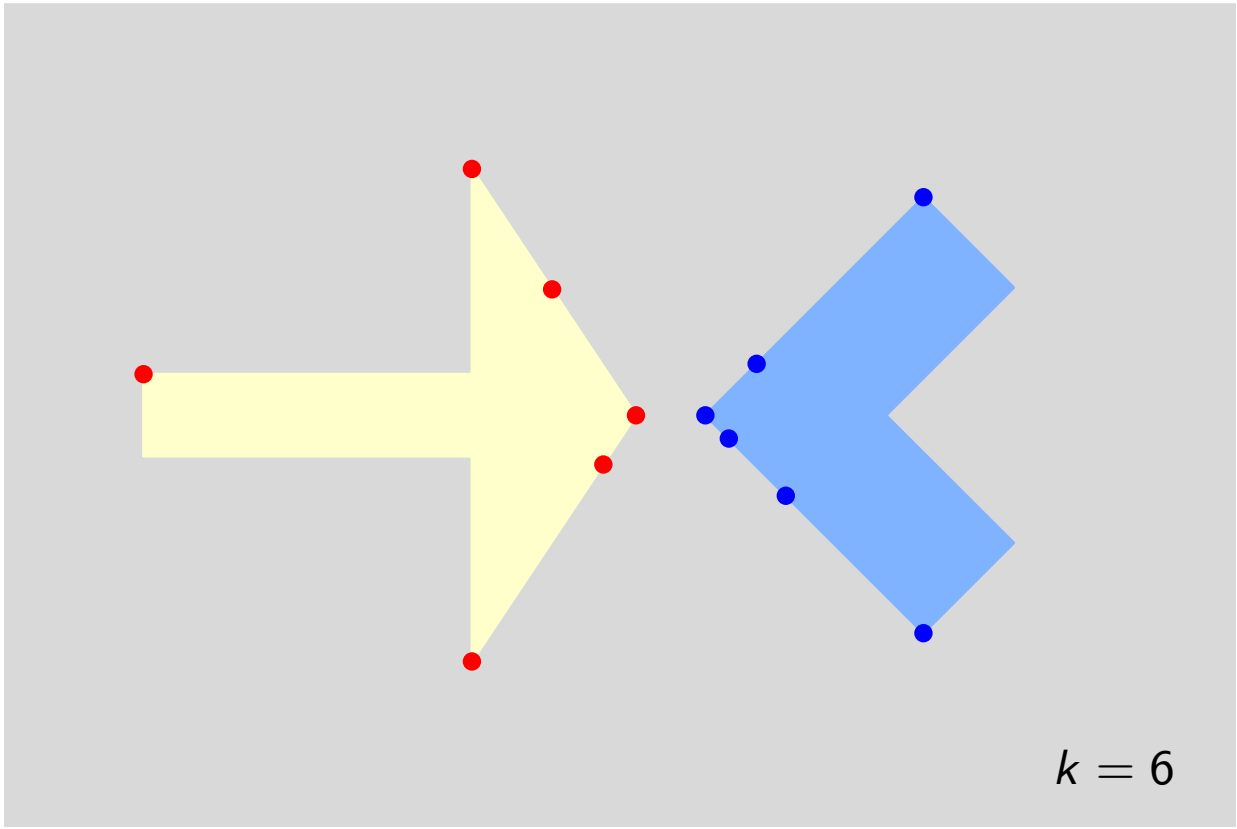
- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



$\bullet = q_j$ $\bullet = -p_j$

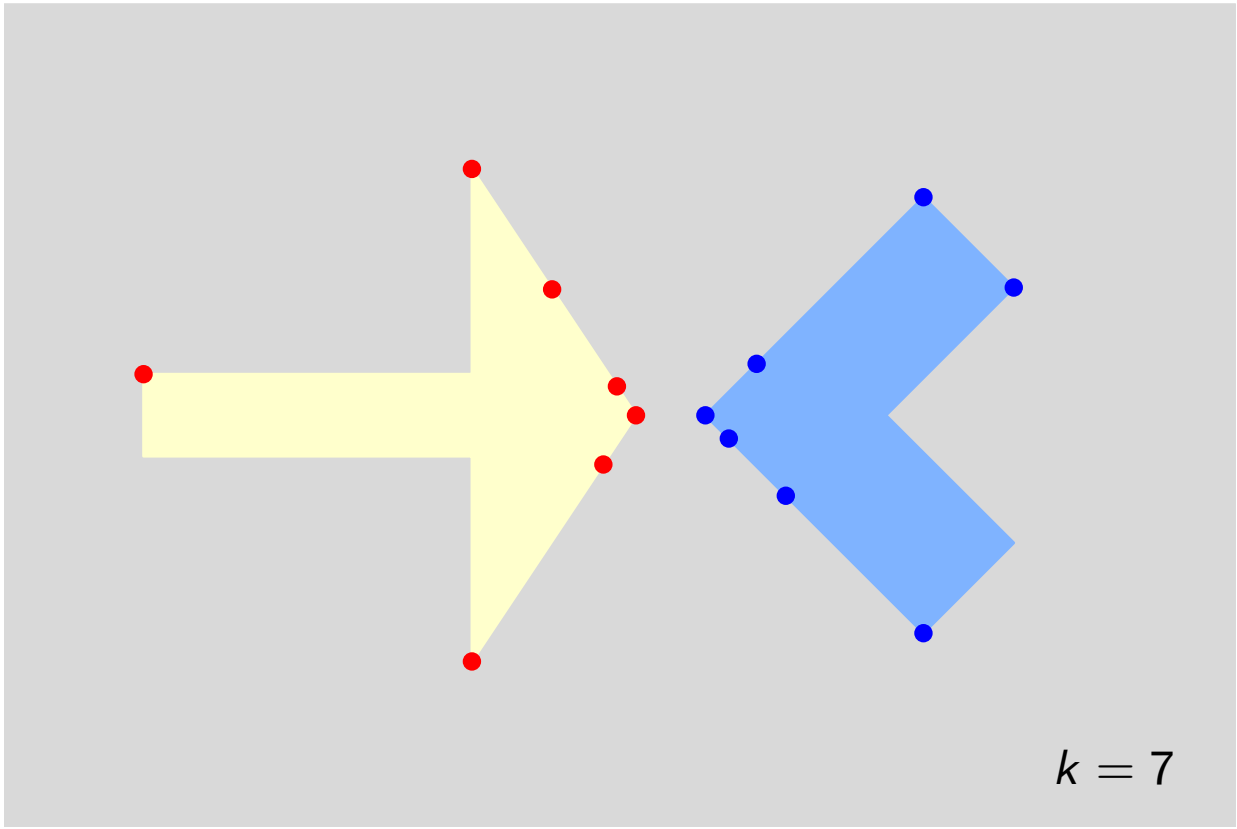
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



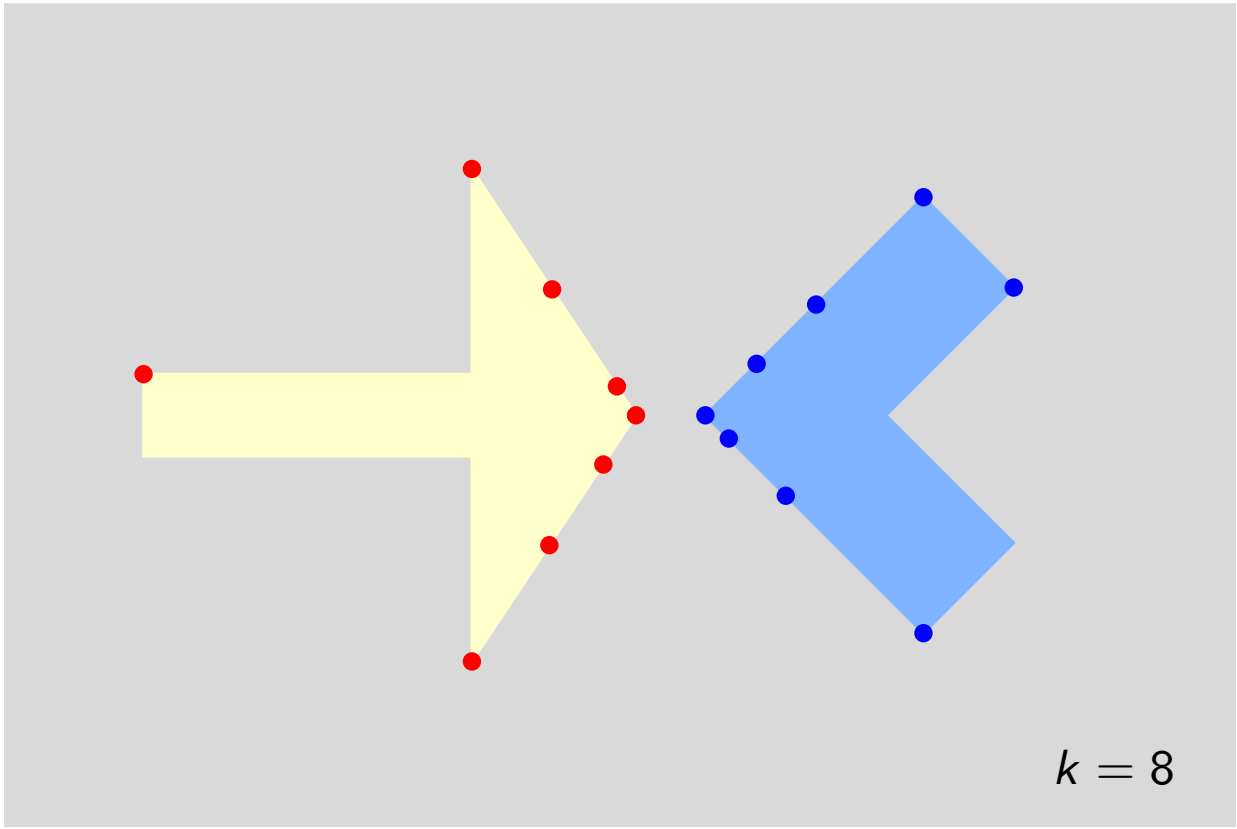
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



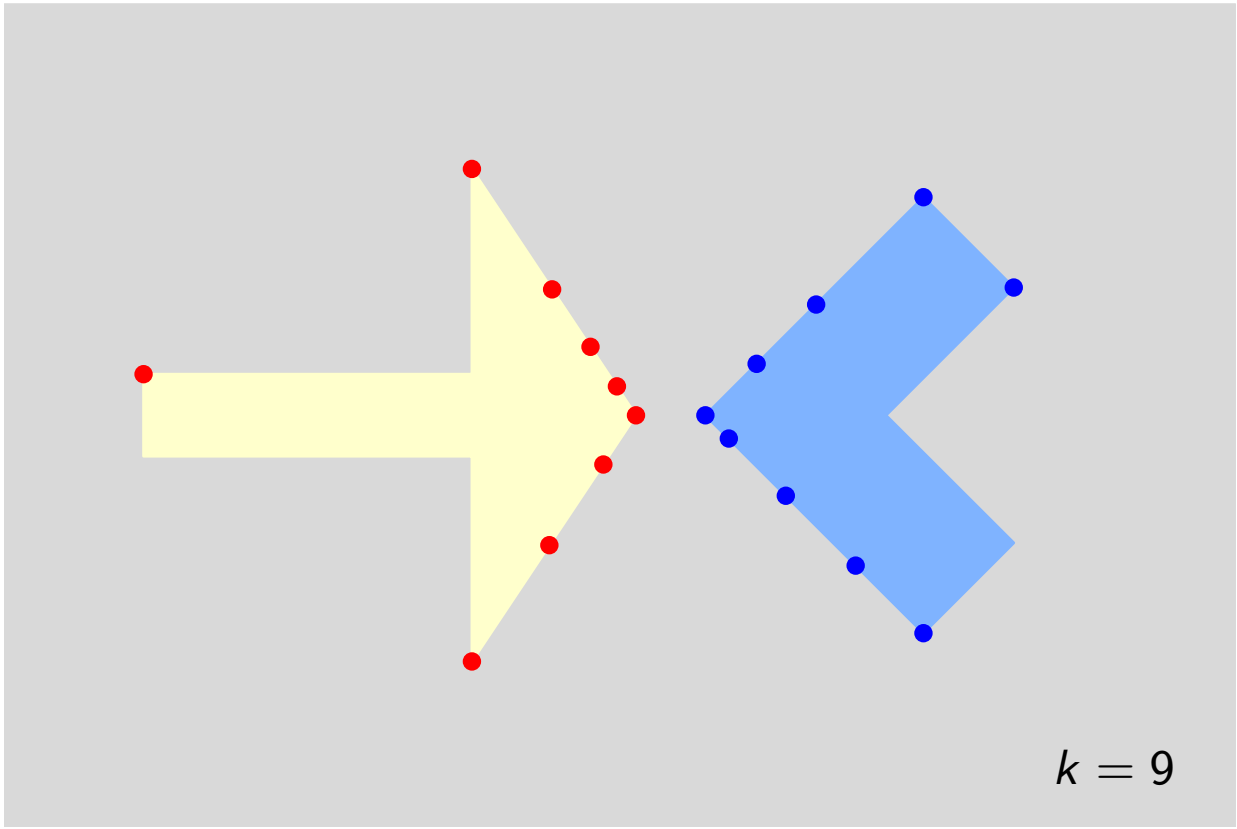
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



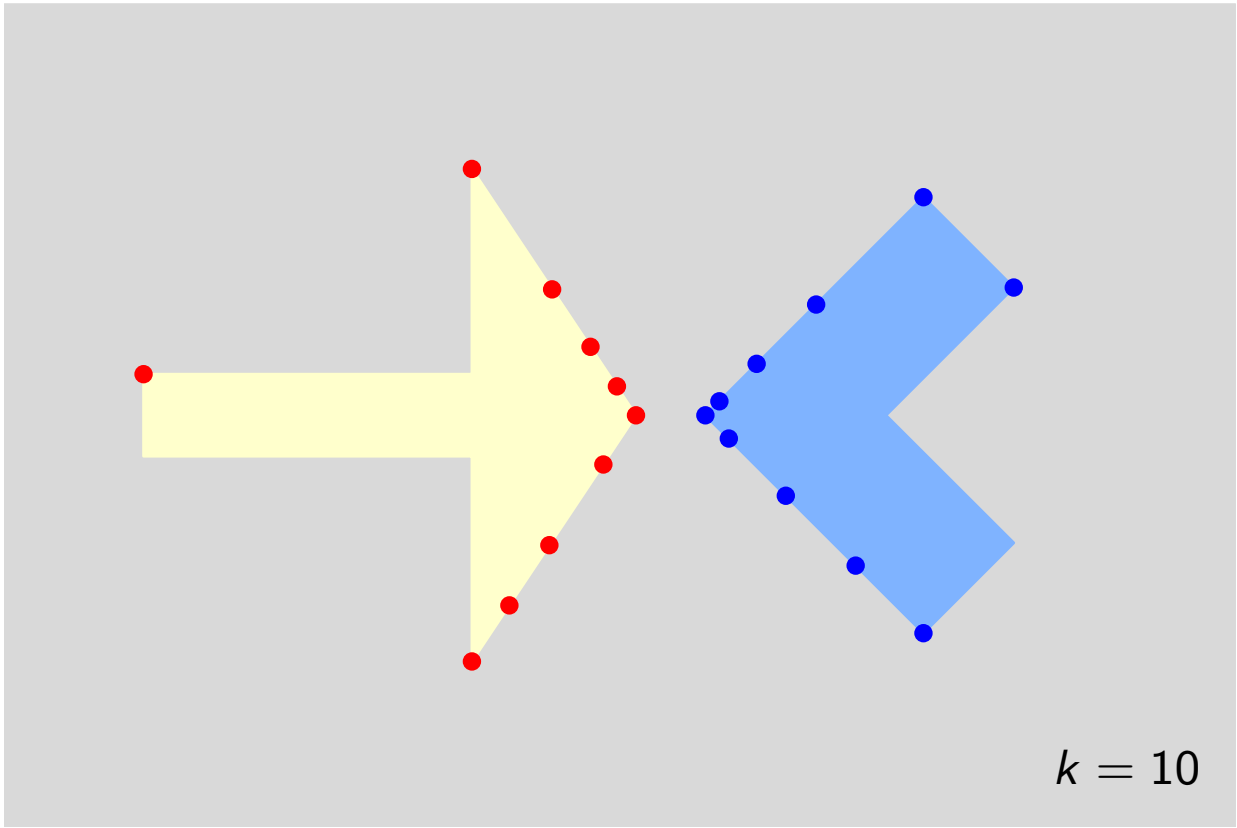
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



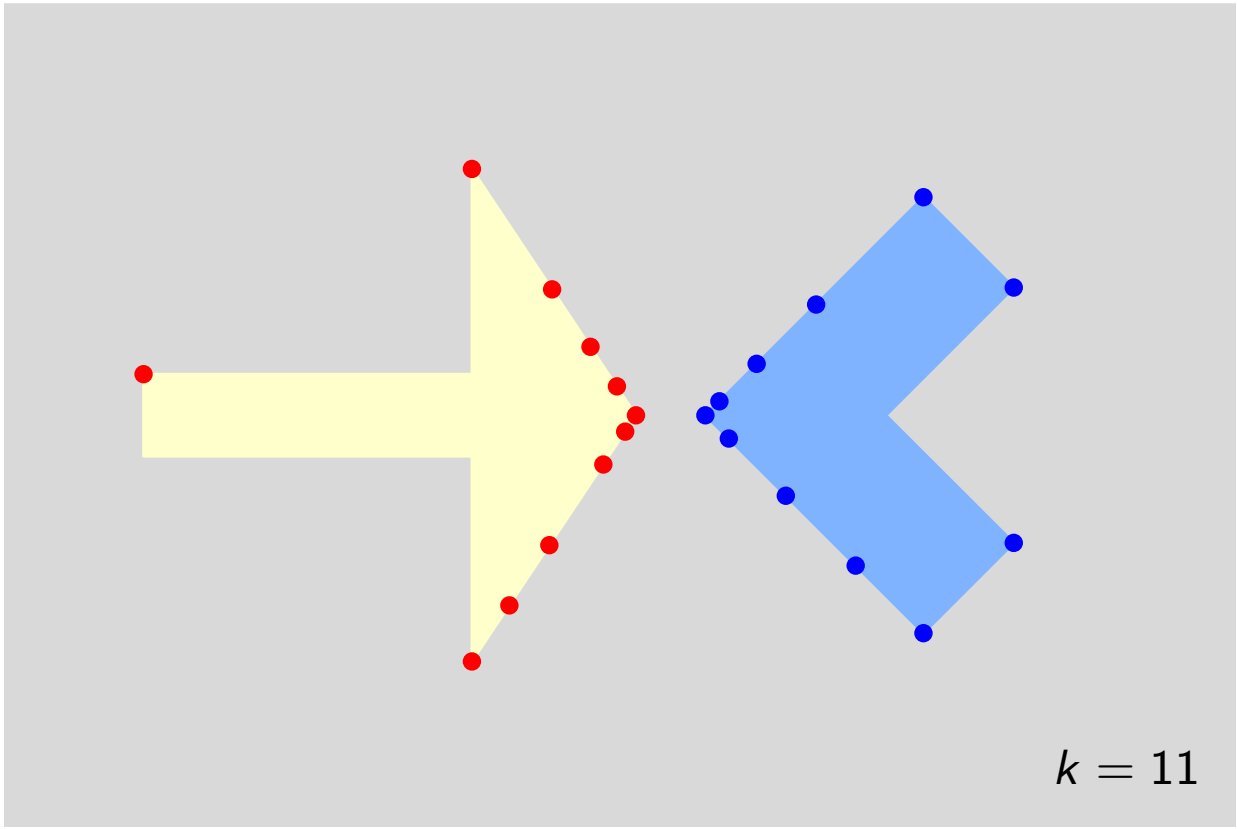
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



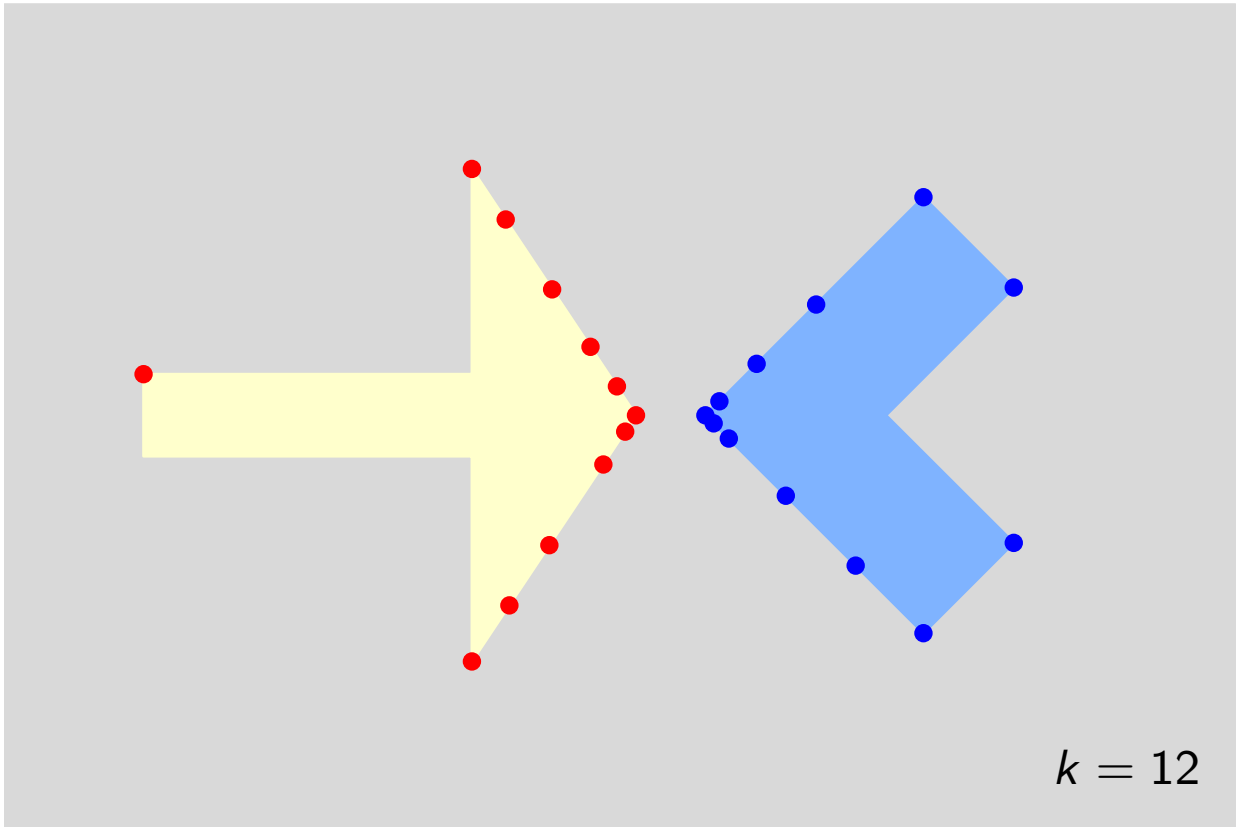
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



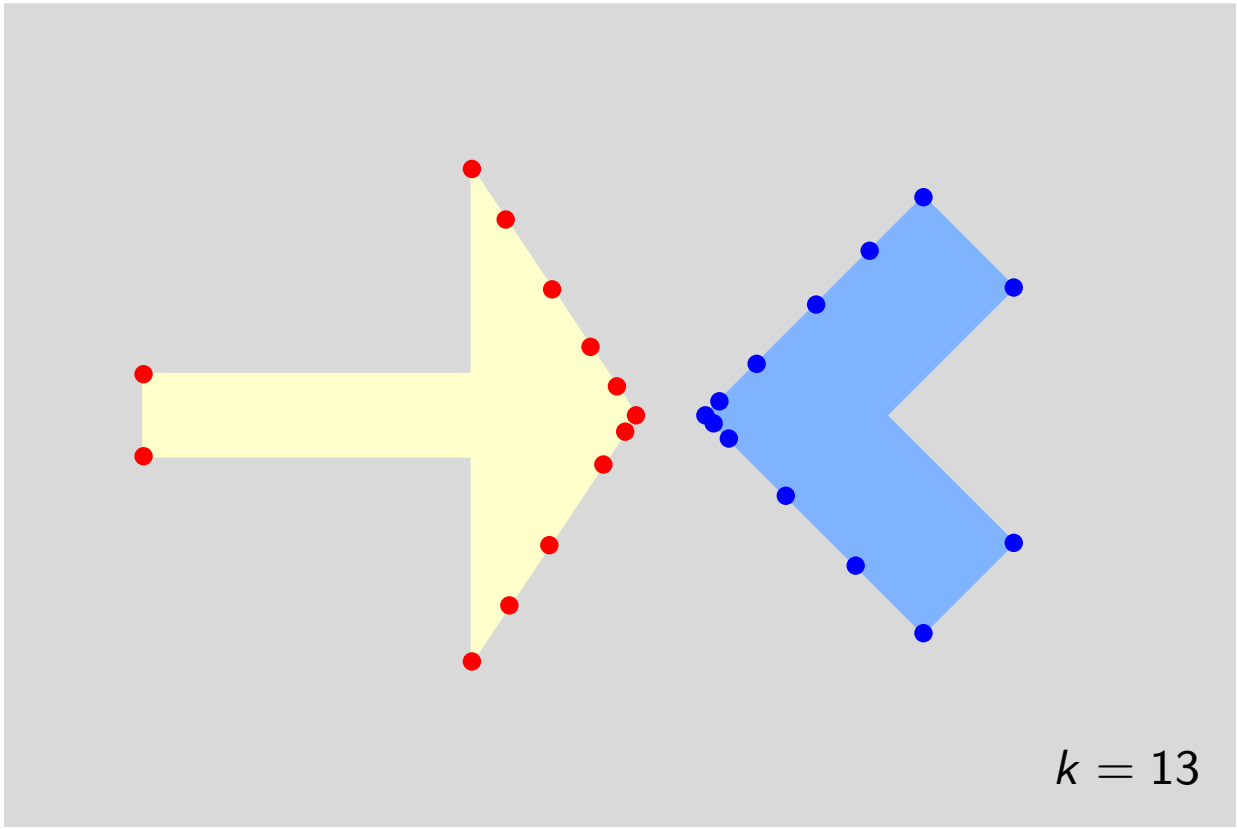
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



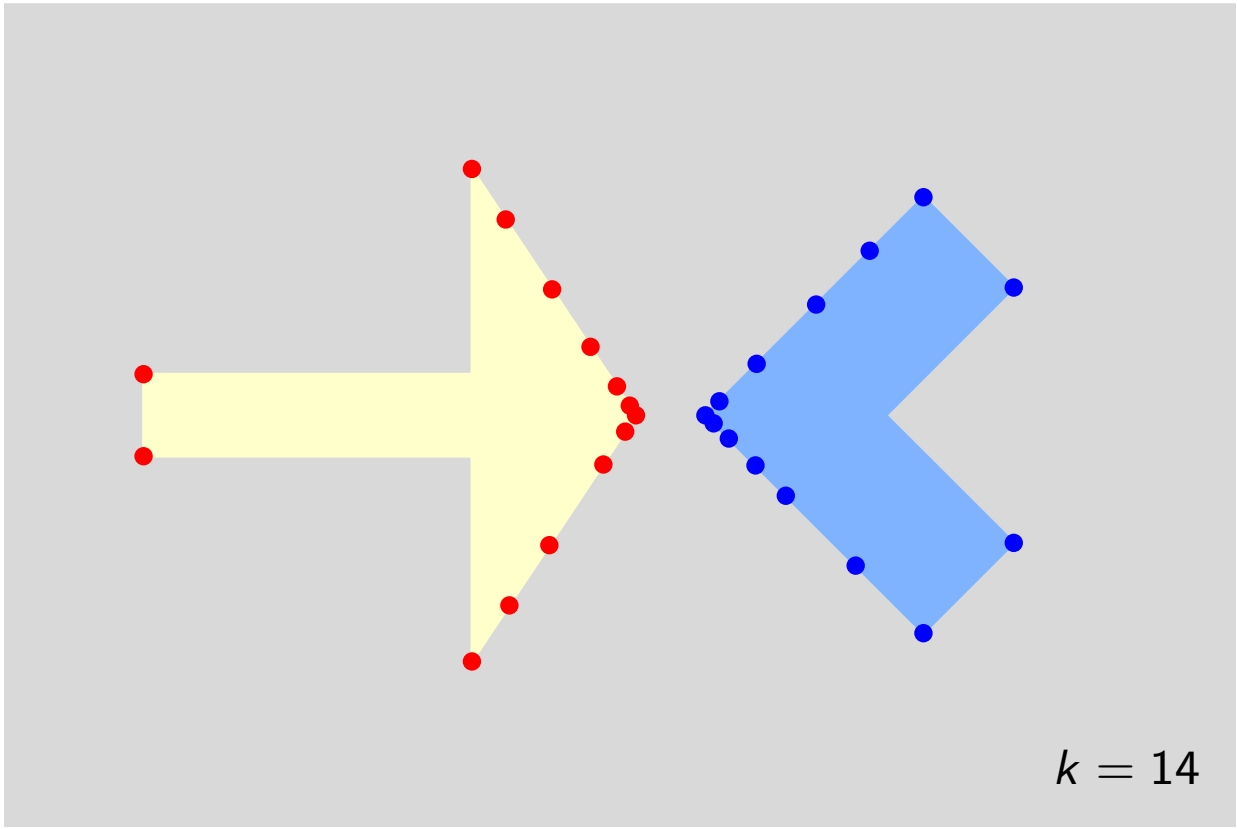
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



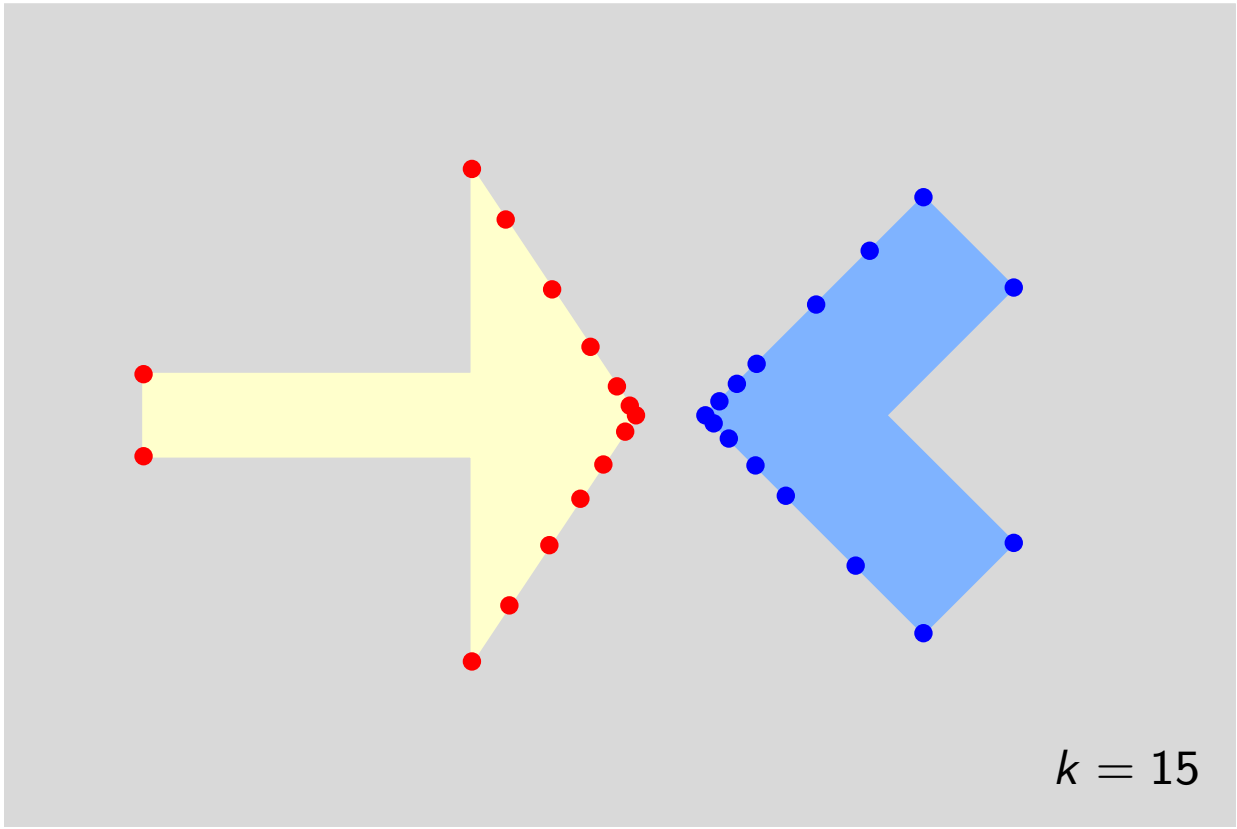
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



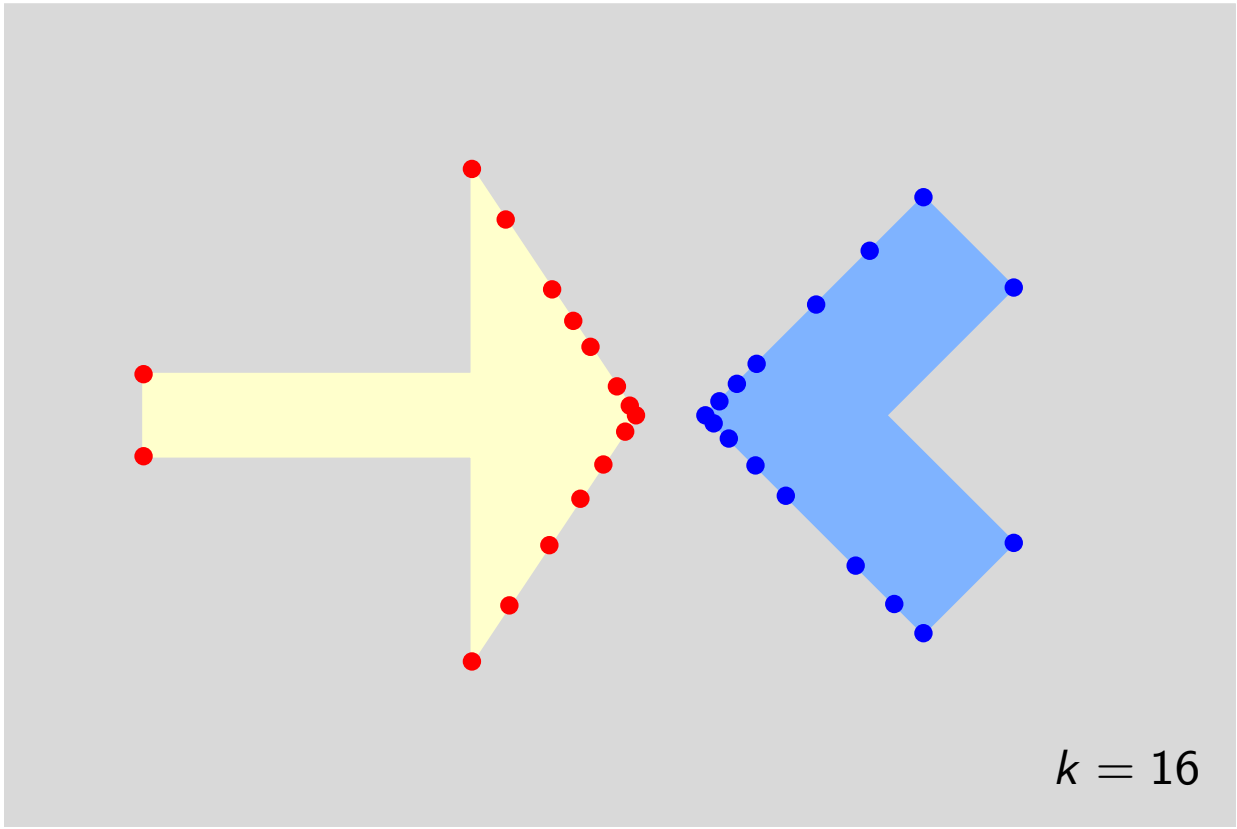
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



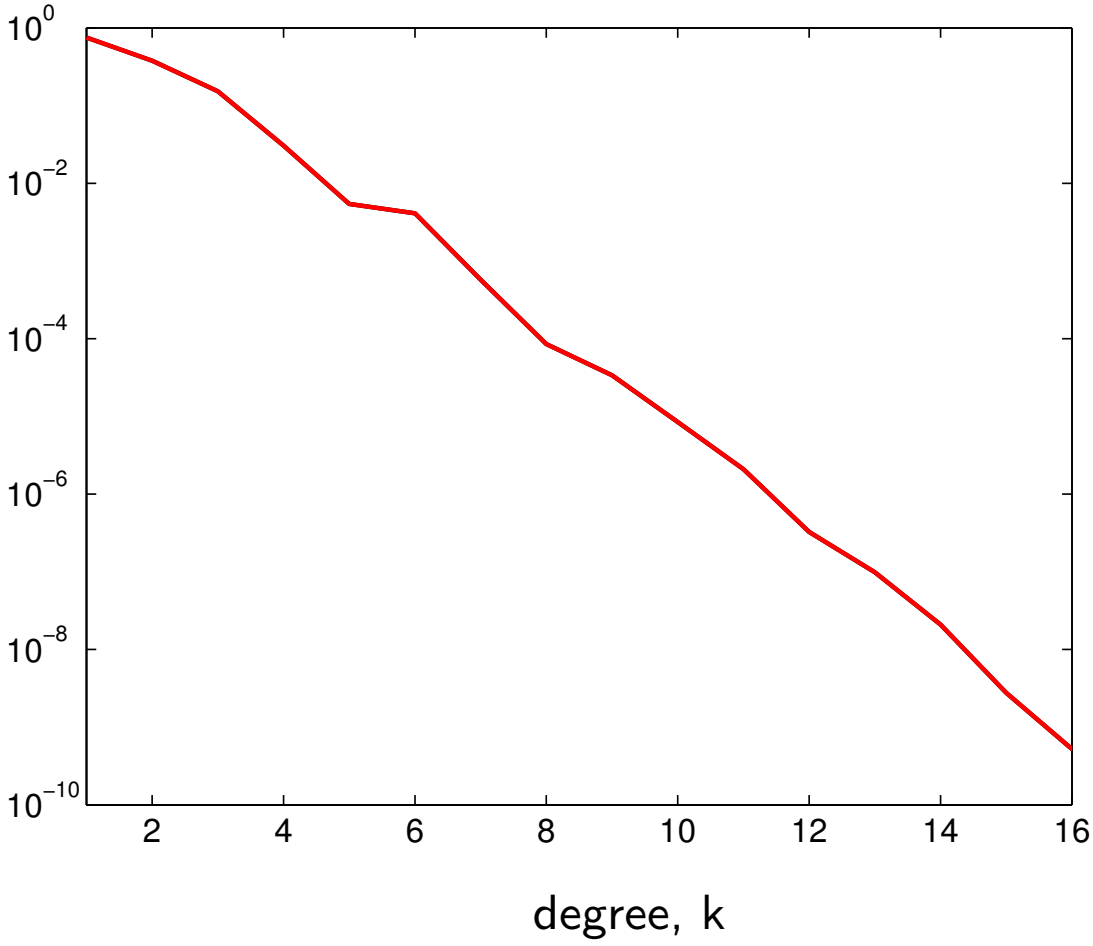
Approximation of Leja–Bagby points

- Leja–Bagby points can be easily approximated by discretizing the boundaries of Ω_A, Ω_B, and picking arbitrary $q_0 \in \Omega_A$, $-p_0 \in \Omega_B$.



Error from Leja–Bagby points

$$\frac{\max\{|\phi_k(z)| : z \in \Omega_A\}}{\min\{|\phi_k(z)| : z \in \Omega_B\}}$$



For this example, we estimate $\exp(-1/\mathrm{cap}(\Omega_A, \Omega_B)) \approx 0.24$.

Strategies for selecting shifts

Estimate spectra of A, B (or numerical ranges [Starke 1993], or pseudospectra) to obtain *disjoint* $\Omega_A, \Omega_B \subset \mathbf{C}$.

- Asymptotically optimal shifts (Leja–Bagby or Fejér–Walsh points)
- Optimal shifts for a bounding rectangle [Istace & Thiran 1993, 1995]
- Penzl's shifting strategy [Penzl, 2000]
  - Collect Ritz values for A, A^{-1}; choose shifts via Bagby ordering.
  - Caveat: Ritz values not typically distributed like optimal points.
  - $|\phi_k(z)|$ may be large at points in the interior of the spectrum.

Strategies for selecting shifts

Estimate spectra of A, B (or numerical ranges [Starke 1993], or pseudospectra) to obtain *disjoint* $\Omega_A, \Omega_B \subset \mathbf{C}$.

- Asymptotically optimal shifts (Leja–Bagby or Fejér–Walsh points)
- Optimal shifts for a bounding rectangle [Istace & Thiran 1993, 1995]
- Penzl's shifting strategy [Penzl, 2000]
  - Collect Ritz values for A, A^{-1}; choose shifts via Bagby ordering.
  - Caveat: Ritz values not typically distributed like optimal points.
  - $|\phi_k(z)|$ may be large at points in the interior of the spectrum.
- Use global optimization to approximate optimal shifts for Ω [Sabino 2006]
  - Nelder–Mead direct search method via MATLAB's `fminsearch`
  - Sometimes better parameters are determined by restricting the search to the real line, rather than the entire complex plane.
  - Scalar objective function cheap to evaluate compared to matrix operations.
  - For the modest number of shifts practical in large-scale computations, these parameters can often out-perform those that are only *asymptotically* optimal.

Practical aspects of shift selection

- A modest improvement in convergence rate gives considerable speed-up:

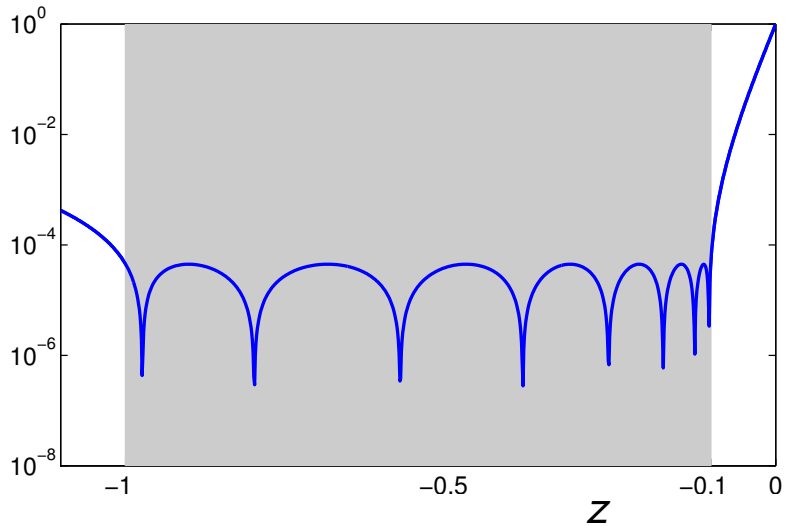
  If $\rho_1^{m_1} = \rho_2^{m_2} = \tau$ with $\rho_1 = 1 - \varepsilon_1$ and $\rho_2 = 1 - \varepsilon_2$, then

  $$\frac{m_1}{m_2} \approx \frac{\varepsilon_2}{\varepsilon_1}.$$

  For example, if $\rho_1 = 0.999$ and $\rho_2 = 0.998$, then $m_2 \approx m_1/2$.

- Due to the equioscillation behavior of ADI rational functions, it is better to overestimate than underestimate the spectrum.

$|\phi(z)|$



$\Omega_A = -\Omega_B = [-1, -1/10]$

Modified Low-Rank Smith algorithm

- An variation of ADI/cyclic Smith [Penzl 2000] that computes the approximate solution X_k in low-rank form and progressively compresses X_k to maintain low rank [Antoulas, Gugercin, Sorensen 2003].

Modified Low-Rank Smith algorithm

- An variation of ADI/cyclic Smith [Penzl 2000] that computes the approximate solution X_k in low-rank form and progressively compresses X_k to maintain low rank [Antoulas, Gugercin, Sorensen 2003].

Given: parameters $\{p_j\}_{j=0}^{k-1}$, $\{q_j\}_{j=0}^{k-1}$, number s of applications/pair

Set $X_0 = 0$.
for $m = 0, 1, 2, \ldots$ until convergence
  for $j = 0, \ldots, k-1$ (loop over parameter pairs)
    Factor $A - q_j$, $B + p_j$, if necessary.
    for $i = 1, \ldots, s$ (s ADI iterations per pair)
      Perform ADI step with parameters (p_j, q_j) to obtain $X_{mks+js+i}$.
      Check residual.
    end
    Compress $X_{mks+(j+1)s}$ to reduce rank according to tolerance τ.
      (sequential Karhunen-Loeve algorithm; see [Baker 2004]).
  end
end

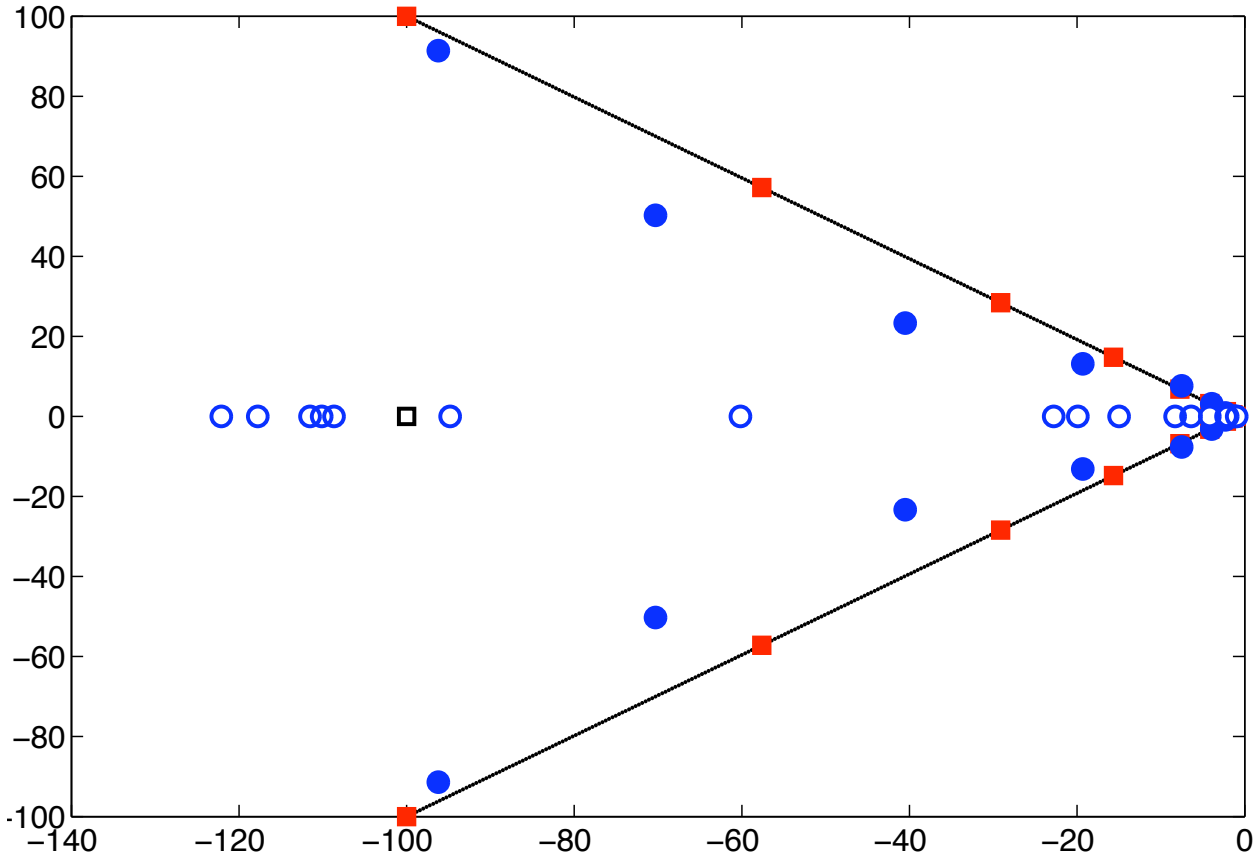
Practical considerations for a Modified Low-Rank Smith code

- Number of parameter to use and their values
- Order in which parameters are applied: Bagby ordering
- Number of consecutive times to apply each pair of parameters
- Use of real versus complex arithmetic
- Shift application: direct or inexact Krylov (potential for recycling, etc.)
  Effect of shift choice on fill-in, convergence rate
  — Heuristic for balancing number of shifts versus cost of factorization
- Accuracy of SVD compression
  — New error bound on accuracy of truncated SVD approximation
- Residual computation

For details on each of these areas, see:
J. N. Sabino, *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*, Rice University CAAM Report 06-08, 2006.

Thesis includes the following numerical examples (and many more).

Illustration of several shift designs: ν domain



· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts ■ Leja–Bagby shifts □ optimal for rectangle

Illustration of several shift designs: v domain

Spectral radius of the ADI iteration matrix

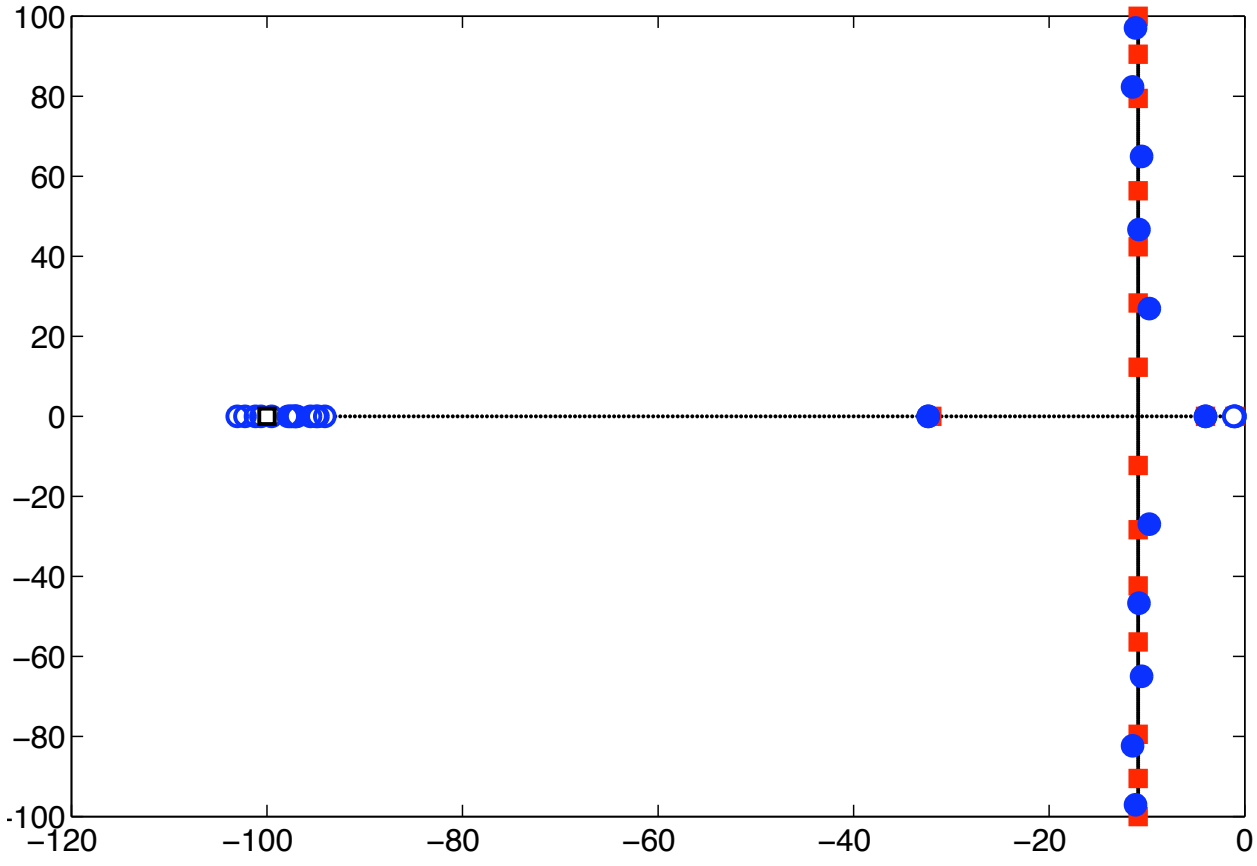
$$\phi_k(A) = \prod_{j=0}^{k-1} (A - p_j)(A + p_j)^{-1}$$

with k shifts for the strategies shown on the previous slide.

shifts	rectangle	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Leja–Bagby
2	0.96	0.52	0.75	0.74 (3)
4	0.92	0.22	0.27	0.31 (5)
8	0.85	0.043	0.041	0.061 (9)
16	0.73	0.0017	0.0016	0.0038 (17)

Leja–Bagby approach uses an extra shift to maintain complex conjugates.

Illustration of several shift designs: t domain



· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts ■ Leja–Bagby shifts □ optimal for rectangle

Illustration of several shift designs: t domain

Spectral radius of the ADI iteration matrix

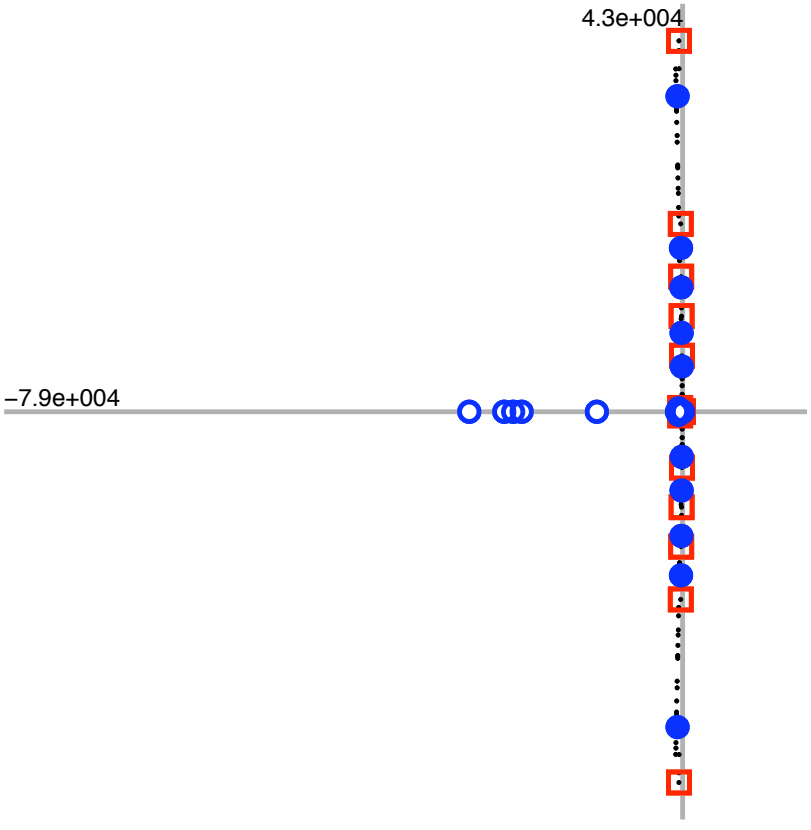
$$\phi_k(A) = \prod_{j=0}^{k-1} (A - p_j)(A + p_j)^{-1}$$

with k shifts for the strategies shown on the previous slide.

shifts	rectangle	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Leja–Bagby
2	0.96	0.87	0.88	0.94 (3)
4	0.92	0.71	0.75	0.75 (5)
8	0.85	0.47	0.25	0.40 (9)
16	0.73	0.20	0.051	0.062 (17)

Leja–Bagby approach uses an extra shift to maintain complex conjugates.

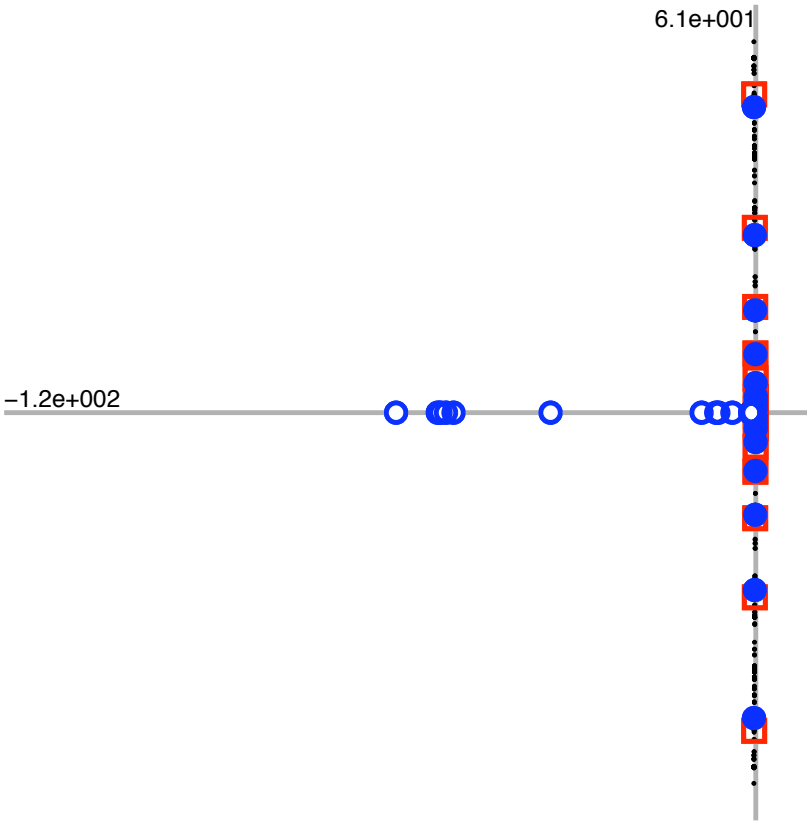
Shift designs for SLICOT examples: CD player



	spectral radius		
shifts	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Bagby
2	0.9963	0.9997	0.9999
4	0.9918	0.9991	0.9997
8	0.9799	0.9949	0.9990
16	0.9642	0.9866	0.9935

· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts □ Leja–Bagby shifts

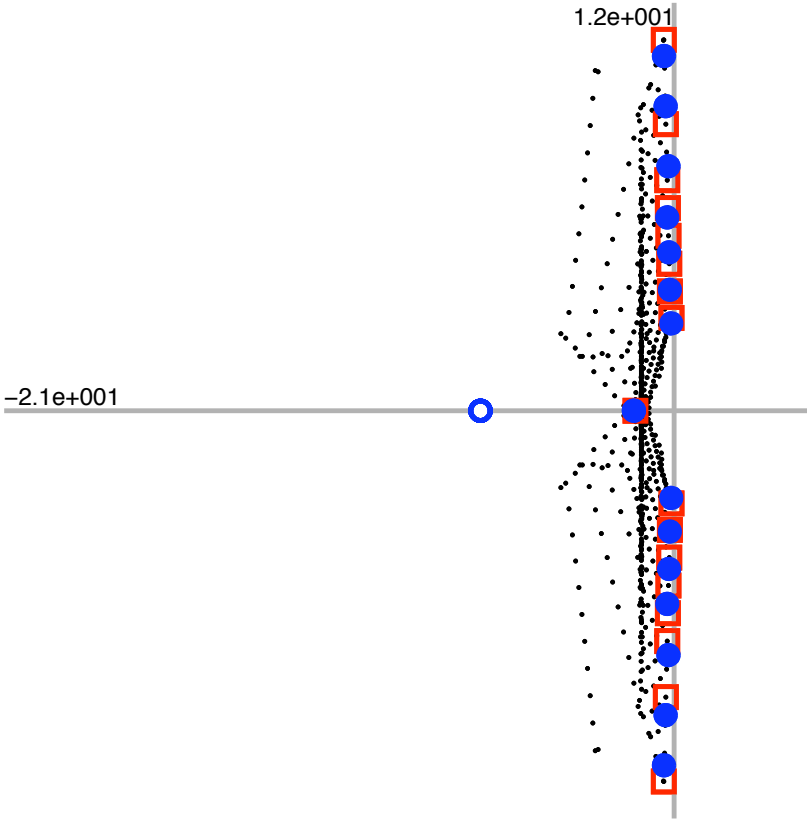
Shift designs for SLICOT examples: ISS



	spectral radius		
shifts	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Bagby
2	0.9959	0.9980	1.0000
4	0.9925	0.9997	1.0000
8	0.9834	0.9995	0.9999
16	0.9668	0.9983	0.9991

· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts □ Leja–Bagby shifts

Shift designs for SLICOT examples: Eady

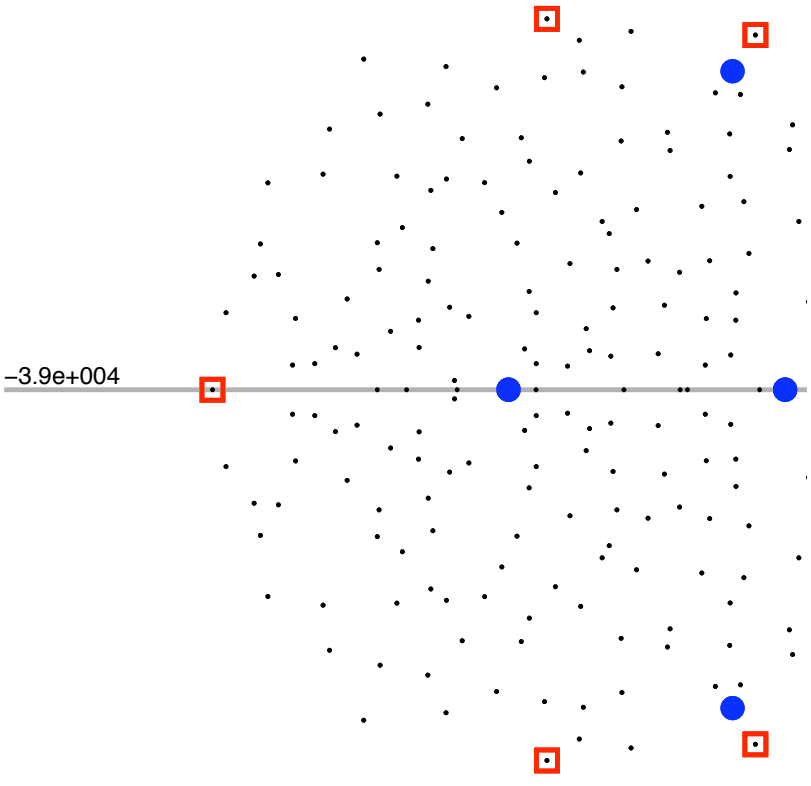


	spectral radius		
shifts	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Bagby
2	0.9555	0.9555	0.9837
4	0.9130	0.9132	0.9753
8	0.8335	0.9094	0.9479
16	0.6947	0.6447	0.7693

(n.b. significant departure from normality)

· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts □ Leja–Bagby shifts

Shift designs for SLICOT examples: random



	spectral radius		
shifts	NM ($\mathbf{R}$)	NM ($\mathbf{C}$)	Bagby
2	1.0000	1.0000	0.9298
4	0.9999	0.8414	0.8785
8	0.9999	0.2645	0.3020
16	0.9998	0.0069	0.0152

· spectrum of A ● complex Nelder–Mead shifts ○ real Nelder–Mead shifts □ Leja–Bagby shifts

Timings for Lyapunov solve, power grid model (nonsymmetric)

n	matrix dimension
$t_{\texttt{eigs}}$	time required to estimate the spectrum, seconds
t_{shifts}	time required to compute real Nelder–Mead shifts, seconds
t_{smith}	time required for the modified Smith method, seconds
k	number of distinct shifts selected
s	number of times each shift is applied
$\widehat{k}$	number of distinct shifts actually used
r	rank of computed solution

n	$t_{\texttt{eigs}}$	t_{shifts}	t_{smith}	k	s	$\widehat{k}$	r
7396	1	1.0	3	4	33	3	37
29796	6	1.0	17	4	58	3	59
67196	13	1.0	50	4	77	3	80
119596	25	1.0	121	4	105	3	103
269396	52	1.1	345	7	80	5	143
366796	77	5.3	591	5	135	4	165

Convergence criterion: relative residual norm $\leq 10^{-6}$
Sun Ultra 20, 2.2 GHz AMD Opteron 148; 3 GB RAM