Zolotarev numbers

Heather Wilber
11 July 2019

1. Introduction. These notes are an introduction to Zolotarev numbers. We keep things concrete here and focus on the role of Zolotarev numbers in a single application: solving Sylvester equations with the ADI method. For in-depth details about the history of Zolotarev numbers in approximation theory, we refer to [1, 11]. Zolotarev's work is relevant in many other areas of numerical linear algebra that are not discussed in these notes.

2. The Sylvester equation. Let A, B, X, F be $n \times n$ matrices, and consider the matrix equation

$$AX - XB = F. \tag{1}$$

This is known as a Sylvester equation. In the special case where $B = A^*$ and X, F are Hermitian, (1) is called a Lyapunov matrix equation.[1] These matrix equations arise in many important applications, including the numerical solving of PDEs, control theory, signal processing, time series analysis, and in the study of structured matrices.

We wish to solve (1) for X in an efficient way. One can show that a unique solution exists whenever A and B have spectra (sets of eigenvalues) that are disjoint from one-another. Many numerical methods for solving (1) exist, and an excellent overview is given in [8]. The Bartels-Stewart method is a reliable and backward-stable direct method requiring $\mathcal{O}(n^3)$ operations. However, when A and B are sparse or otherwise structured, it can be advantageous to use iterative methods instead. This is especially true when F is a low rank matrix, since it often implies that X can be constructed in low-rank form. Iterative methods include projection-based methods, usually based on Krylov or rational-Krylov subspaces (these crucially depend on F being low rank), and ADI-based methods. We only discuss ADI here, as it is as a gateway to the Zolotarev numbers. More details about Krylov-based methods (and the connections between these methods and ADI) can be found in [8, Sec. 4.4.2].

3. The ADI method. The ADI algorithm is an iterative method that numerically solves (1) by alternately updating the column and row spaces of an approximate solution [5,6]. One ADI iteration consists of the following two steps:

1. Solve for $X^{(j+1/2)}$, where

$$\left(A - \beta_{j+1} I\right) X^{(j+1/2)} = X^{(j)} \left(B - \beta_{j+1} I\right) + F. \tag{2}$$

2. Solve for $X^{(j+1)}$, where

$$X^{(j+1)} \left(B - \alpha_{j+1} I\right) = \left(A - \alpha_{j+1} I\right) X^{(j+1/2)} - F. \tag{3}$$

An initial guess, $X^{(0)}$, is required to begin the iterations. The construction of $X^{(k)}$ requires selecting a set of k 2-tuples, $\{(\alpha_j, \beta_j)\}_{j=1}^{k}$, referred to as *shift parameters*.

3.1. Deriving the ADI iteration. To derive the ADI iteration from first principles, we first observe that for any pair of real numbers (α, β), it is true that

$$(A - \beta I)X(B - \alpha I) - (A - \alpha I)X(B - \beta I) = (\beta - \alpha)F. \tag{4}$$

This leads to a natural iteration:

$$X^{(j+1)}(B - \alpha I) = (\beta - \alpha)(A - \beta I)^{-1}F + (A - \beta I)^{-1}(A - \alpha I)X^{(j)}(B - \beta I). \tag{5}$$

[1] The terms 'Sylvester' and 'Lyapunov' are often used in a more general sense in dynamical systems and control theory: a thorough discussion is given in [8].

Now we observe that

$$(\beta-\alpha)(A-\beta I)^{-1} = -I + (A-\alpha I)(A-\beta I)^{-1}, \tag{6}$$

and substitute this into (5). As a result, we have that

$$X^{(j+1)}(B-\alpha I) = -F + (A-\alpha I)(A-\beta I)^{-1}F + (A-\beta I)^{-1}(A-\alpha I)X^{(j)}(B-\beta I). \tag{7}$$

Now, we choose $X^{(j+1/2)}$ so that

$$(A-\beta I)X^{j+1/2} - F = X^{(j)}(B-\beta I).$$

By substituting this expression into (7) and observing that $(A-\alpha I)$ and $(A-\beta I)^{-1}$ commute, we recover the two-step ADI iteration shown in (2) and (3).

The ADI method was originally derived as an implicit-explicit scheme for numerically solving the heat equation. See Section 7 for some details from this perspective.

3.2. The ADI error equation. The convergence of the ADI method depends on the choice of shift parameters, as well as properties of A and B. We can understand this more clearly by taking a look at the ADI error equation. The error satisfies the following equation:

$$X_{j+1} - X = \tag{8}$$

$$(A-\alpha_j I)(A-\beta_j I)^{-1}(X_j - X)(B-\beta_j I)(B-\alpha_j I)^{-1} = \tag{9}$$

$$\left[\prod_{k=0}^{j}(A-\alpha_k I)(A-\beta_k I)^{-1}\right](X_0 - X)\left[\prod_{k=0}^{j}(B-\beta_k I)(B-\alpha_k I)^{-1}\right]. \tag{10}$$

We can write this more succinctly as

$$X - X^{(k)} = r_k(A)(X - X^{(0)})r_k(B)^{-1}, \qquad r_k(z) = \prod_{j=1}^{k}\frac{(z-\alpha_j)}{(z-\beta_j)}, \qquad k \geq 1, \tag{11}$$

where r_k is a degree (k,k) rational function with zeros at $\{\alpha_j\}_{j=1}^{k}$ and poles at $\{\beta_j\}_{j=1}^{k}$. If we choose $X^{(0)} = 0$, then we have that

$$\|X - X^{(k)}\|_2 \leq \|r_k(A)r_k(B)^{-1}\|_2\|X\|_2. \tag{12}$$

The big question in ADI is how to pick the shift parameters, $\{\alpha_j\}_{j=1}^{k}$ and and $\{\beta_j\}_{j=1}^{k}$ so that $\|r_k(A)r_k(B)^{-1}\|_2$ is as small as possible, since this will minimize the bound on the error. As we will see, a famous problem of Zolotarev's can give us the answer.

4. Rational approximation and the ADI error equation. Let's look at a simple example. Suppose that A and B are both normal matrices (a matrix X is normal if $X^*X = XX^*$), and that their eigenvalues are denoted by the sets $\lambda(A)$ and $\lambda(B)$, respectively.

Since A is normal, it can be written as $A = U\Lambda_A U^*$, where U is unitary and Λ_A is diagonal with nonzero entries given by the eigenvalues of A. Notice that for the polynomial $p_2(z) = z^2$, $p_2(A) = U\Lambda_A U^* U\Lambda_A U^* = U\Lambda_A^2 U^* = Up_2(\Lambda_A)U^*$. It is not hard to see that for any polynomial $p(z)$, $p(A) = Up(\Lambda_A)U^*$. Similarly, for our rational function $r_k(z)$ in (11), we have that $r(A) = Ur(\Lambda_A)U^*$. In the same way, if $B = V\Lambda_B V^*$, then $r_k(B) = Vr_k(\Lambda_B)V^*$. Now consider (12). We have that

$$\|X - X^{(k)}\|_2 \leq \|r_k(A)r_k(B)^{-1}\|_2\|X\|_2 \tag{13}$$

$$\leq \|Ur_k(\Lambda_A)U^*\|_2\|Vr_k(B)^{-1}V^*\|_2\|X\|_2 \tag{14}$$

$$\leq \|r_k(\Lambda_A)\|_2\|r_k(\Lambda_B)^{-1}\|_2\|X\|_2, \tag{15}$$

where we have used the fact that $\|\cdot\|_2$ is unitarily invariant. We can reformulate the problem of finding the shift parameters as a rational approximation problem. To minimize the bound in (15), we need to find a rational function of degree (k,k) that is as small as possible on the set $\lambda(A)$ and as large as possible on the set $\lambda(B)$. Formally, we seek the rational function associated with the following number:

$$Z_k(\lambda(A),\lambda(B)) := \inf_{r\in\mathcal{R}^k} \frac{\sup_{z\in\lambda(A)}|r(z)|}{\inf_{z\in\lambda(B)}|r(z)|}, \tag{16}$$

where $\mathcal{R}^k$ is the space of all rational functions with numerators of degree $\leq k$ and denominators of degree $\leq k$. The number Z_k is referred to as the kth Zolotarev number associated with $\lambda(A)$ and $\lambda(B)$, and the rational function that achieves Z_k is called a Zolotarev rational function. The names are in honor of Y.I. Zolotarev, a student of Chebyshev who first posed (and subsequently solved) a version of the extremal approximation problem shown in (16)[2] [13]. In his original paper, Zolotarev posed 4 approximation problems. The one in (16) is referred to as Zolotarev's third problem.[3]

Notice that the ADI error $\|X - X^{(k)}\|_2$ can be bound in terms of the Zolotarev number $Z_k(\lambda(A),\lambda(B))$. If we take r_k to be the Zolotarev rational function, then we have that $\|r_k(A)\|_2 \leq \sup_{z\in\lambda(A)}|r(z)|$, and $\|r_k(B)^{-1}\|_2 \leq \inf_{z\in\lambda(B)}|r(z)|$. From (15), it follows that

$$\|X - X^{(k)}\|_2 \leq Z_k(\lambda(A),\lambda(B))\|X\|_2. \tag{17}$$

By studying the behavior of the Zolotarev numbers, we can understand the convergence behavior of the ADI algorithm.

5. Properties of Zolotarev numbers.

In general, it is extremely difficult to solve Zolotarev's third problem for arbitrary sets $\lambda(A)$ and $\lambda(B)$. It is particularly hard to solve such a problem for sets of discrete points, such as the spectra of two matrices. To make the problem easier, we assume that $\lambda(A) \subset E$ and $\lambda(B) \subset G$, where E and G are disjoint sets in the complex plane. We can then study $Z_k(E,G)$ in lieu of $Z_k(\lambda(A),\lambda(B))$.

5.1. Separation and decay rates.

The first thing we might notice about $Z_k(E,G)$ is that it decays rapidly with k whenever E and G are well-separated from one-another, and it decays more slowly with k if E and G are very close together. We can see this with a simple example. Choose E as the interval $[10,20]$, and set $G = [-20,-10]$, $G' = [-10,0]$. In Figure 2, the blue dots show $Z_k(E,G)$, plotted on a log scale against $k = 1,\ldots,10$, and the red dots show $Z_k(E,G')$.

Intuitively, $Z_k(E,G')$ decays more slowly because it is more difficult to make a rational function small on E and large on G when E and G are closer together. In Figure 2, we plot r_k for (E,G) in blue and for (E,G') in red when $k=4$ to demonstrate this.

As an extreme example, consider the sets shown in Figure 3. If r_k is large on G, it is because there are a lot of poles on (or near) G. These, 'blow up' the value of r_k on G. Since portions of E are nearby to G, the size of r_k on E is affected by the poles and cannot be too small. We can try to counteract the influence of the poles by placing zeros on (or near) E, but of course this then affects the size of r_k on G. As a result, we require many zeros and poles in order to 'tack down' or 'blow up' r_k appropriately.

[2] Zolotarev was not thinking about matrix equationsor ADI, but rather a generalization of a well-known problem of Chebyshev in approximation theory. He originally posed the problem not for two sets of eigenvalues, but for the sets E and $-E$, where E is a finite interval on the positive real line.

[3] It is worth noting that in some instances, people use the terms 'Zolotarev number' or 'Zolotarev function' for quantities/functions associated with Zolotarev's other problems.

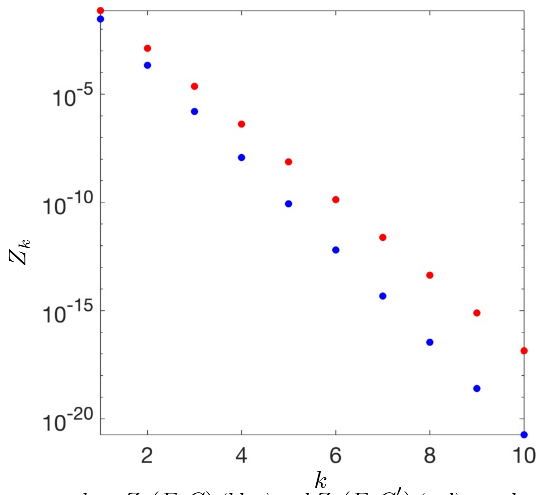

FIG. 1. *The Zolotarev numbers $Z_k(E, G)$ (blue) and $Z_k(E, G')$ (red) are plotted on a log scale against the index k. Here, $E = [10, 20]$, $G = [-20, -10]$ and $G' = [-10, 0]$. The closer the intervals are together, the slower the Z_k decays.*

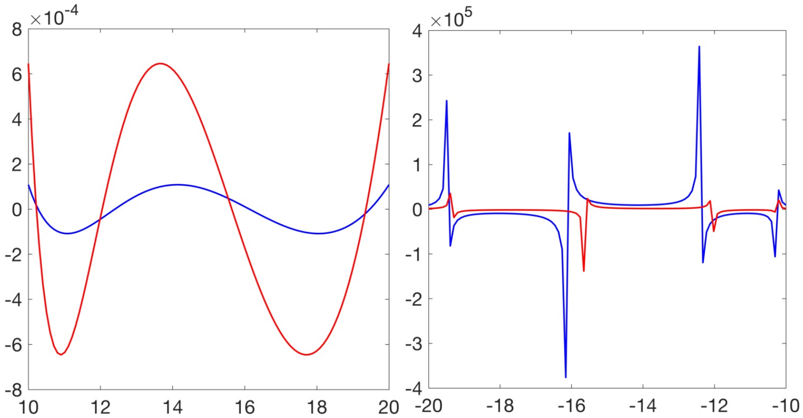

FIG. 2. *Left:The Zolotarev rational functions $r_4(z)$ associated with (E, G) (blue) and (E, G') (red) are plotted for $z \in E$. Right: The Zolotarev rational functions $r_4(z)$ associated with (E, G) (blue) and (E, G') (red) are plotted for $z \in G$ and $z \in G'$, respectively. For this plot, we have shifted the plot on G' and superimposed it on the plot for G.*

These examples show us that ADI will converge rapidly if E and G are sufficiently separated from one another.

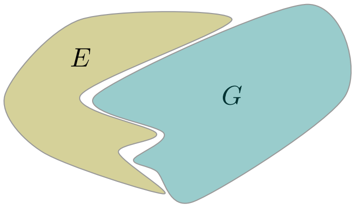

FIG. 3. *The sets E and G are very close together. This makes it hard to construct a rational function r_k that is small in magnitude on E and large in magnitude on G.*

5.2. Other useful properties. Several useful properties of Zolotarev numbers are given in [2].For any sets E and G that are disjoint and well-separated, the following hold:

1. $Z_0(E, G) = 1$,
2. $Z_k(E, G) = Z_k(G, E)$,
3. $Z_{k+1}(E, G) \leq Z_k(E, G)$,
4. $Z_{k_1+k_2}(E, G) \leq Z_{k_1}(E, G)Z_{k_2}(E, G)$,
5. $Z_k(E_1, G_1) \leq Z_k(E, G)$ whenever $E_1 \subseteq E, G_1 \subseteq G$,
6. $Z_k(T(E), T(G)) = Z_k(E, G)$, where T is a Möbius transformation.

5.3. Properties related to potential theory. TO DO

5.4. Known solutions to Zolotarev's third problem. We only know r_k and $Z_k(E, G)$ in certain special cases. The problem was first solved by Zolotarev for $E = [-\gamma, 1] \subset \mathbb{R}$ and $G = -E$. Details on the solution can be found in [1, 2, 11]. Here, we simply note that the value of $Z_k([-\gamma, 1], [1, \gamma])$ is given in [2, Thm. 3.1] as an infinite product. A convenient, tight bound is given in [2, Sec. 3]:

$$Z_k([-\gamma, 1], [1, \gamma]) \leq 4 \left[\exp\left(\frac{\pi^2}{2\log(4\gamma)}\right)\right]^{-2k}. \qquad (18)$$

The zeros and poles of r_k associated with $Z_k([-\gamma, 1], [1, \gamma])$ are given by elliptic integrals and can be computed explicitly with little difficulty in MATLAB (see the appendix in [4] for pseudocode; a MATLAB function is available in the freeLYAP repository.)

Solutions for many other sets E', G' can be derived from the above known solution due to the invariance of Z_k under Möbius transformations. For example, if $E' = [a, b] \subset \mathbb{R}$ and $G' = [c, d] \subset \mathbb{R}$, with $[a, b] \cap [c, d] = \emptyset$, then there is a Möbius transformation that maps $[a, b] \cup [c, d]$ to $[-\gamma, 1] \cup [1, \gamma]$. Likewise, we can use Möbius transformations to map arcs on the unit circle to intervals on the real line.

Less work has focused on solving Zolotarev's third problem for more general sets, but the solution is known when E and G are two disjoint disks in the complex plane. See [10] and [12] for details.

6. Using Zolotarev rational functions for ADI. The following criteria make ADI ideal for solving $AX - XB = F$:

1. A and B are normal matrices,
2. Shifted linear solves and matrix-vector multiplication involving A and B are cheap,
3. The spectra of A and B are contained in two disjoint and well-separated sets E and G,
4. A solution to Zolotarev's problem is known for the sets E and G.

If all four criteria are met, then we know how to pick optimal shift parameters for ADI, and we know convergence will be rapid. In fact, we have explicit bounds on the approximation error $\|X - X^{(k)}\|_2$ via (17). For this reason, we know exactly how many ADI iterations are required. In this setting, we view ADI not as an iterative algorithm, since we aren't monitoring a residual to determine convergence, but rather as a process for building up the approximation $X^{(k)}$ in k steps.

Even when all four criteria are not fully met, we may still be able to use ADI very effectively. For example, if the matrices A and B are near-normal, then we expect that ADI will still converge rapidly. If we do not know the exact solution to Zolotarev's problem for the sets E and G, but we have a way to construct an approximate solution, then we can use this approximate solution to find ADI shift parameters.

Truly iterative variants of ADI are sometimes used when only items 2 and 3 are met, especially when the boundaries of E and G are not known. In this setting, one uses a heuristic strategy to choose a collection of shift parameters. These may be a small set of shift parameters chosen a priori and then applied cyclically (sometimes this is referred to as the cyclic Smith's method [7, 9]), or they may be shift parameters adaptively computed on the fly at each iteration [3, 7, 8]. In this setting, one typically monitors the residual and stops iterating once it is small enough.

6.1. The low rank property. If, in addition to the above criteria, F is a low rank matrix, then we also know that the solution X is well-represented by a low rank matrix [2]. This means that methods such as factored ADI [3] and Krylov-based methods can be effective for constructing low rank approximations to X. More details about factored ADI can be found in [3, 12].

7. ADI and its connection to the heat equation. Originally, the ADI equations were discovered in [6] as an implicit-explicit numerical method for solving the heat equation. We give a very concise review and refer the reader to [6] for details. Consider the 2D diffusion equation

$$u_t = u_{xx} + u_{yy}, \quad u(x, y) \in [-1, 1]^2, \tag{19}$$

with Dirichlet boundary conditions. The Crank-Nicolson method is an implicit time-stepping scheme used to update the solution from u_n to u_{n+1} in the following way:

$$u_{n+1} - u_n = \Delta t(D_x^2 + D_y^2)(u_{n+1} + u_n),$$

where Δt is the timestep and D_x^2 is the second-order central finite difference operator applied in the x-direction. To make this method more computationally efficient, it was realized that one can break the update into two steps, solving implicitly only in either x or y at each step:

$$u_{n+1/2} - u_n = (\Delta t/2)(D_x u_{n+1/2} + D_y u_n), \tag{20}$$

$$u_{n+1} - u_{n+1/2} = (\Delta t/2)(D_x u_{n+1/2} + D_y u_{n+1}) \tag{21}$$

The great advantage of this approach is that each ADI step only requires the solving of a symmetric tridiagonal linear system. It can be shown that this method is unconditionally stable. The connection between ADI for solving the heat equation and ADI for solving linear matrix equations more generally comes from the fact that one can use the above approach once the solution to the heat equation reaches steady state. This leads to a numerical method for solving Poisson's equation. The time step (Δt) can then be replaced with arbitrary values (these become the shift parameters).

REFERENCES

[1] N. I. Achieser, *Theory of approximation*, Courier Corporation, 2013.

[2] B. Beckermann and A. Townsend, *On the singular values of matrices with displacement structure*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1227–1248.

[3] P. Benner, R.-C. Li, and N. Truhar, *On the ADI method for Sylvester equations*, J. Comput. Appl. Math., 233 (2009), pp. 1035–1045.

[4] D. Fortunato and A. Townsend, *Fast Poisson solvers for spectral methods*, arXiv preprint arXiv:1710.11259, (2017).

[5] A. Lu and E. L. Wachspress, *Solution of Lyapunov equations by alternating direction implicit iteration*, Computers & Mathematics with Applications, 21 (1991), pp. 43–58.

[6] D. W. Peaceman and H. H. Rachford, Jr, *The numerical solution of parabolic and elliptic differential equations*, J. Soc. for ind. Appl. Math., 3 (1955), pp. 28–41.

[7] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (1999), pp. 1401–1418.

[8] V. Simoncini, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441.

[9] R. Smith, *Matrix equation XA+BX=C*, SIAM J. Appl. Math., 16 (1968), pp. 198–201.

[10] G. Starke, *Near-circularity for the rational Zolotarev problem in the complex plane*, Journal of approx. theory, 70 (1992), pp. 115–130.

[11] J. Todd, *Applications of transformation theory: A legacy from Zolotarev (1847–1878)*, in Approximation theory and spline functions, Springer, 1984, pp. 207–245.

[12] A. Townsend and H. Wilber, *On the singular values of matrices with high displacement rank*, Linear Algebra Appl., 548 (2018), pp. 19–41.

[13] E. Zolotarev, *Application of elliptic functions to questions of functions deviating least and most from zero*, Zap. Imp. Akad. Nauk. St. Petersburg, 30 (1877), pp. 1–59.