PADÉ, STIELTJES, LANCZOS AND THE EXPONENTIAL SUPERCONVERGENCE OF FINITE-DIFFERENCE SCHEMES

VLADIMIR DRUSKIN* AND LEONID KNIZHNERMAN†

Abstract.
The standard three-point finite-difference discretization of the two-point problems for the second order ODE system is considered. We implement an approach commonly used for acceleration of iterative methods, i.e., we relate the approximate boundary Neumann-to-Dirichlet map to a rational approximation of the impedance function on the spectrum of the ODE matrix and optimize the error revisiting a connection between Padé approximations, the Stieltjes moment problem, Lanczos method and the inverse spectral problem for tridiagonal matrices. It increases from second to exponential the convergence order of the Neumann-to-Dirichlet map without increasing the stencil of the finite-difference scheme and loosing stability. We consider two model problems with constant coefficients, present numerical examples, then show an extension to a more general problem allowing variable (even rough) coefficients.

1. Introduction. The simplest and most common finite-difference (FD) discretization of the second derivatives of elliptic equations is the three-point stencil. Such an approximation applied to two- or three-dimensional Laplace equations yields five or seven point schemes respectively with the second order of convergence. Traditionally the convergence order is increased with the help of larger FD stencils, the exponential convergence can be only achieved with using even more computationally expensive spectral and pseudospectral methods. It is worth mentioning that so-called superconvergence, i.e., the *polynomial* increase of the convergence rate (typically by one or two orders) compared to the local truncation error is known for finite-element or FD approximations.

We show that a special distribution of grid steps of the standard three-point scheme yields the *exponential* (for some cases even *superexponential*) convergence rate of the Neumann-to-Dirichlet (ND) map.

We stress that the connection between FD schemes and Padé approximation of the square root used here is known; for example, it was employed for the solution of the first order hyperbolic problems [4, e.g.], where it produced the polynomial order of *global* convergence with the size of stencil proportional to the convergence order.

First we consider bounded and semiinfinite homogeneous domains, where we extend the approach of [2]. We demonstrate numerical examples exhibiting the exponential convergence for a semiinfinite domain and the superexponential convergence for a bounded domain. We conclude with a generalization for the Sturm-Liouville equation in the Krein-Kac [5] form with variable coefficients.

2. A problem with constant coefficients and its FD approximation. Let us consider the two-point problem on $[0, L]$, $L > 0$ (including $L = +\infty$ as the limit

* Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877-4108, U.S.A.
† Central Geophysical Expedition, Narodnogo Opolcheniya St., 40-3, Moscow 123298, Russia.

case)

$$(1) \qquad Aw - \frac{d^2 w}{dx^2} = 0, \qquad \frac{dw}{dx}|_{x=0} = -\varphi, \ w|_{x=L} = 0.$$

Here we for simplicity assume that A is a symmetric positive definite matrix, i.e., $w \in \mathbf{R}^n$, $A \in \mathbf{R}^{n\times n}$. However, all the results will be valid for self-adjoint (even unbounded) operators in Hilbert spaces provided the spectrum of the Neumann data is contained in a positive segment of the real axis. The ND can be defined as

$$w(0) = f(A)\varphi,$$

$f(\lambda)$ is commonly called the impedance function, simple calculations yield

$$(2) \qquad f(\lambda) = \frac{1 - e^{-2L\sqrt{\lambda}}}{\sqrt{\lambda}(1 + e^{-2L\sqrt{\lambda}})}.$$

The FD solution is defined at "potential" nodes x_i, $i = 1, \ldots, k+1$, $x_1 = 0$, $x_{i+1} > x_i$, the FD derivatives are defined at nodes $\hat{x}_i$, $i = 0, \ldots, k$, $\hat{x}_0 = 0$. We denote $h_i = x_{i+1} - x_i$, $\hat{h}_i = \hat{x}_i - \hat{x}_{i-1}$. Instead of (1) we will solve the FD problem

$$Aw_i - \frac{1}{\hat{h}_i}\left(\frac{w_{i+1} - w_i}{h_i} - \frac{w_i - w_{i-1}}{h_{i-1}}\right) = 0, \qquad i = 2, \ldots, k,$$

$$(3) \qquad Aw_1 - \frac{1}{\hat{h}_1}\left(\frac{w_2 - w_1}{h_1}\right) = \frac{1}{\hat{h}_1}\varphi, \qquad w_{k+1} = 0.$$

For example, if we let A be the three-point approximation of the differential operator $-\frac{\partial^2}{\partial y^2} + k(y)I$, then (3) becomes the well-known five-point FD approximation of elliptic equations in $\mathbf{R}^2$. However, in standard FD schemes x_i and $\hat{x}_i$ are related, for example, usually $\hat{h}_i = (h_i + h_{i-1})/2$, but we allow both h_i and $\hat{h}_i$ to be arbitrary positive numbers to increase the degree of freedom for the optimization.

We will optimize the error of computing the ND using (3), i.e., $\delta_k = \|f(A)\varphi - w_1\|$. The crucial fact is that

$$(4) \qquad w_1 = f_k(A)\varphi,$$

where the discrete impedance function $f_k(\lambda)$ is a rational function of λ depending on $h_i, \hat{h}_i$ as parameters.

3. Rational representation of the finite-difference Neumann-to-Dirichlet map. We denote $u_i = \sqrt{\hat{h}_i} w_i$ and transform the FD problem (3) to the block-tridiagonal form

$$(5) \qquad Au_i - (\beta_{i-1}u_{i-1} + \alpha_i u_i + \beta_i u_{i+1}) = \delta_1^i \hat{h}_1^{-1/2}\varphi, \qquad i = 1, \ldots, k,$$

where

$$(6) \qquad \beta_0 = 0, \qquad \alpha_1 = -\frac{1}{\hat{h}_1 h_1}, \qquad \beta_1 = \frac{1}{h_1\sqrt{\hat{h}_1\hat{h}_2}},$$

$$\beta_i = \frac{1}{h_i\sqrt{\hat{h}_i\hat{h}_{i+1}}}, \qquad \alpha_i = -\hat{h}_i^{-1}\left(\frac{1}{h_i} + \frac{1}{h_{i-1}}\right), \qquad i = 2, \ldots, k.$$

Let z_j, λ_j be A's eigenvectors and eigenvalues respectively, $\varphi_j = z_j^T \varphi$. Then $c_i^j = z_j^T u_i$ satisfy the equations

$$\lambda_j c_i^j - \left(\beta_{i-1} c_{i-1}^j + \alpha_i c_i^j + \beta_i c_{i+1}^j\right) = \delta_1^i \hat{h}_1^{-1/2} \varphi_j, \qquad i = 1, \ldots, k, \tag{7}$$

where δ_1^i is the Kronecker symbol. From (7) we obtain a matrix expression for c_1^j as follows:

$$c_1^j = \hat{h}_1^{-1/2} \varphi_j e_1^T (\lambda_j I - H)^{-1} e_1, \tag{8}$$

where $H \in \mathbf{R}^{k\times k}$ is a symmetric matrix with the main diagonal (α_i) and the subdiagonal (β_i). Let θ_i be H's eigenvalues and s_i be the first components of the corresponding eigenvectors. Then (8) can be rewritten using eigendecomposition as

$$c_1^j = \hat{h}_1^{-1/2} \varphi_j \sum_{i=1}^{k} \frac{s_i^2}{\lambda_j - \theta_i}. \tag{9}$$

The spectral formula (9) allows us to rewrite w_1 as a matrix function:

$$\hat{h}_1^{1/2} w_1 = u_1 = \hat{h}_1^{-1/2} \sum_{i=1}^{k} s_i^2 (A - I\theta_i)^{-1} \varphi.$$

Finally, using (4) and denoting $y_i = s_i^2 / \hat{h}_1$, we obtain

$$f_k(\lambda) = \sum_{i=1}^{k} \frac{y_i}{\lambda - \theta_i}. \tag{10}$$

Now we estimate the approximation error. Again using eigendecomposition, we obtain

$$\delta_k = \sqrt{\sum_{j=1}^{n} \varphi_j^2 \left[f_k(\lambda_j) - f(\lambda_j)\right]^2} \leq \max_{\lambda \in [\lambda_{min}, \lambda_{max}]} |f_k(\lambda) - f(\lambda)| \, \|\varphi\|, \tag{11}$$

where $[\lambda_{min}, \lambda_{max}] \supset \mathrm{sp}(A)$, $0 < \lambda_{min} < \lambda_{max}$.

4. The Padé–Chebyshev approximation. Let us first for simplicity consider the Padé approximation: find y_i, θ_i, $i = 1, \ldots, k$, such that

$$\left.\frac{d^i}{d\lambda^i} \left[f_k(\lambda) - f(\lambda)\right]\right|_{\lambda = b} = 0, \qquad i = 0, \ldots, 2k-1, \tag{12}$$

for some $b \in [\lambda_{min}, \lambda_{max}]$. The impedance function is a Stieltjes function,[1] namely, it can be written as

$$f(\lambda) = \int_{-\infty}^{0} (\lambda - x)^{-1} \, d\sigma(x),$$

[1] More precisely it is a Markov function of x, however since any Markov function can be transformed to a more familiar Stieltjes form, we will refer to f as to the Stieltjes function.

where $\sigma(x)$ is a positive measure on $]-\infty, 0]$ such that the integral converges. Similarly, f_k can be presented as

$$f_k(\lambda) = \int_{-\infty}^{0} (\lambda - x)^{-1} \, d\sigma_{f_k}(x),$$

where

$$\sigma_{f_k} = \sum_{i=1}^{k} y_i \delta(\lambda - \theta_i).$$

So, (12) is equivalent to the following moment problem: find y_i, θ_i, $i = 1, \ldots, k$, such that

$$\int_{-\infty}^{0} (b - x)^{-i-1} \, d\left[\sigma(x) - \sigma_{f_k}(x)\right] = 0, \qquad i = 0, \ldots, 2k-1. \tag{13}$$

Making the change of variables $t = (b - x)^{-1}$, we can rewrite (13) as a Stieltjes moment problem [1]:

$$\int_{0}^{b^{-1}} t^i \, d\left[t \cdot (\sigma - \sigma_{f_k}) \circ \left(b - t^{-1}\right)\right] = 0, \qquad i = 0, \ldots, 2k-1. \tag{14}$$

Problem (14) is solved by means of the underlying Lanczos process: we construct the Lanczos polynomials of t orthonormal with respect to the measure $t \cdot (\sigma - \sigma_{f_k}) \circ (b - t^{-1})$, starting with $q_1(t) = 1$. Then $y_j = {v_j}^2/\gamma_j$, $\theta_j = b - {\gamma_j}^{-1}$, where γ_j and v_j are respectively the eigenvalues and the first components of the suitably normalized correspondent eigenvectors of the Lanczos tridiagonal matrix. The diagonal Padé approximation of a Stieltjes function with a positive convergence radius converges exponentially. However, its approximation error grows rapidly away from b.

So, instead we use the Padé–Chebyshev approximation that has better approximation properties on a segment [1].

Let $[\lambda_1, \lambda_2]$ be a segment on the real axis, $0 < \lambda_1 < \lambda_2$. Then the Padé–Chebyshev approximant $f_k(\lambda)$ is defined by the conditions

$$\int_{\lambda_1}^{\lambda_2} \lambda^i \left[f_k(\lambda) - f(\lambda)\right] \rho(\lambda) \, d\lambda = 0, \qquad i = 0, 1, \ldots, 2k-1, \tag{15}$$

where $\rho(\lambda) = \left\{1 - \left[\frac{2\lambda - (\lambda_2 + \lambda_1)}{\lambda_2 - \lambda_1}\right]^2\right\}^{-1/2}$ is the Chebyshev spectral weight adjusted to $[\lambda_1, \lambda_2]$. To compute f_k, we can use a general scheme given in [1], however formula (10) presents a special case of Padé–Chebyshev approximant — a $[k-1/k]$ one, and $f(\lambda)$ is the Stieltjes (Markov) function. For this case, again similarly to the standard Padé, the algorithm is reduced to the Stieltjes moment problem and the Lanczos method. [1, 2].

If $L = +\infty$, the spectrum is continuous, and from [3, theorem 1] we can obtain the exponential convergence of $f_k(\lambda)$ to $f(\lambda)$ for $\lambda \in \mathbf{C} \backslash]-\infty, 0]$. Intuitively, the best choice for $[\lambda_1, \lambda_2]$ would be $[\lambda_{min}, \lambda_{max}]$. However, as it was shown in [2], for large quotients $\lambda_{max}/\lambda_{min}$ a better logarithmical convergence rate is obtained if $\lambda_1 = \lambda_{min}$, $\lambda_2 \approx 4\lambda_{min}^{1/3}\lambda_{max}^{2/3}$. Then

$$\max_{\lambda\in[\lambda_{min},\lambda_{max}]} \left|f_k(\lambda) - \lambda^{-1/2}\right| \cong O\left[\exp\left(-4k\sqrt[6]{\frac{\lambda_{min}}{\lambda_{max}}}\right)\right]. \tag{16}$$

Estimate (16) exhibits the exponential convergence with the rate that is slightly depends on A's condition number.

If $L < +\infty$ the spectral measure is discrete. The logarithmical capacity of the discrete spectrum is equal to 0 and we arrive at the *superexponential* convergence [1, theorem 6.6.9]!

Now we discuss the well-posedness of the Padé–Chebyshev approximant.

THEOREM 4.1. *The poles θ_i of the Padé–Chebyshev approximant f_k defined by (15) are simple and negative, and its residues y_i are positive, $i = 1, 2, \ldots, k$.*

This theorem is closely related to theorem 5.2.1 of [1] for the standard Padé, and similarly, its proof is based on the equivalence to the Stieltjes moment problem [2]. Theorem 4.1 provides stability for the underlying FD problem due to a one-to-one correspondence between f_k and the strings with positive masses [5, theorem 11.1]. The string of k discrete masses $\hat{h}_i$ connected by springs with the stifnesses $1/h_i$ has the same matrix as our FD scheme.

5. Algorithm. Now we found $h_i, \hat{h}_i$ in four steps:

1°. Find $y_1, \ldots, y_k$, $\theta_1, \ldots, \theta_k$ from (15).

2°. Assuming the normalization $\sum_{i=1}^k s_i^2 = 1$, compute $\hat{h}_1 = \frac{1}{\sum_{i=1}^k y_i}$ and $s_i^2 = \hat{h}_1 y_i$.

3°. Solve the inverse spectral problem, i.e., from θ_i, s_i^2, $i = 1, \ldots, k$, obtain H.[2]

4°. Using (6), we obtain $h_1 = -\frac{1}{\hat{h}_1\alpha_1}$, then recursively calculate:

$$\hat{h}_i = \frac{1}{\beta_{i-1}^2 h_{i-1}^2 \hat{h}_{i-1}}, \qquad h_i = -\frac{1}{\alpha_i \hat{h}_i + 1/h_{i-1}}, \qquad i = 2, \ldots, k.$$

6. Numerical experiments. We consider two problems with $L = \infty$ and $L = 1$, for both cases $\lambda_{min} = 10^{-2}$ and $\lambda_{max} = 10^2$. In Figure 1 we present four different grids for different k for the both problems. In the logarithmic scale (the points $\hat{x}_0 = 0$ and $x_1 = 0$ can not be seen) the grids are quite regular, the derivative nodes $\hat{x}_i$ separate the potential nodes x_i and vice versa, the minimal grid step decreases versus k. In case of an unbounded domain the upper grid boundary moves away versus k, for the bounded case it quickly converges to the boundary $x = 1$.

[2] Theorem 4.1 allows us to apply [6, theorem 7.2.2] about the existence and the uniqueness of the determination of the symmetric tridiagonal $k \times k$ matrix from real k-dimensional arrays of the simple eigenvalues θ_i and first components of the eigenvectors s_i. A simple recursive k-step algorithm from the proof of [6, theorem 7.2.2] can be used for this purpose; this algorithm is equivalent to the Lanczos recursion. To avoid loss of orthogonality of the Lanczos vectors in computer that causes some artificial irregularity in the grid structure, we use reorthogonalization.

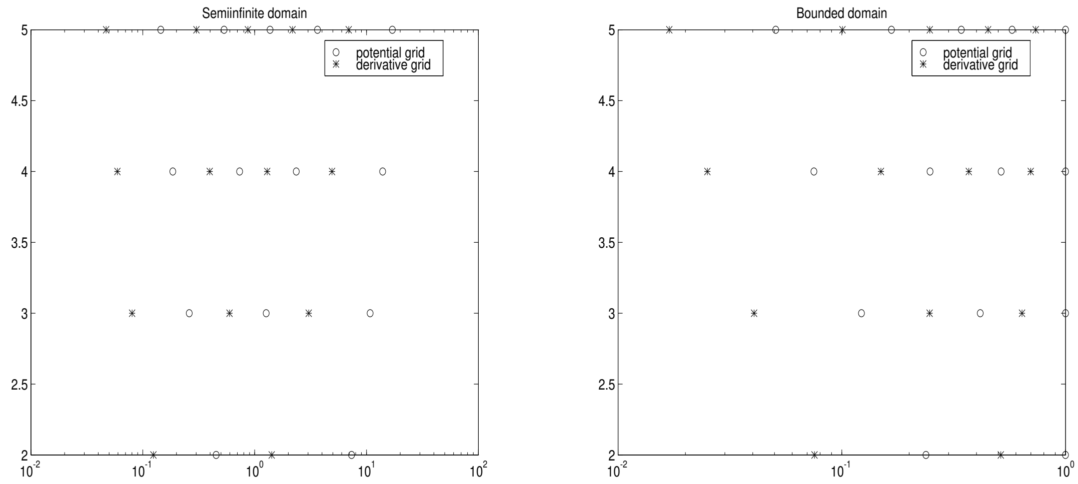

Figure 1. The four grids for $k = 2, 3, 4, 5$.

In Figure 2 we plot the maximal relative error of our approximation on the segment $[\lambda_{min}, \lambda_{max}]$, i.e., $\max_{\lambda\in[\lambda_{min},\lambda_{max}]} |f(\lambda)^{-1} f_k(\lambda) - 1|$. This graph is equivalent to the maximal possible relative grid error of the Dirichlet data. For $L = +\infty$ the convergence (as k increases) is exponential in a good agreement with (16). As it was pointed in section 3, for a bounded domain the error graph clearly exhibits the superexponential convergence.

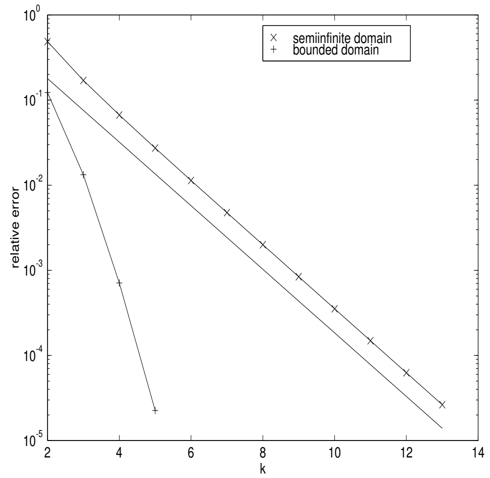

Figure 2. The maximal possible grid approximation error. The plain straight line is (up to a constant multiplier) the asymptotic estimate (16).

7. The Sturm-Liouville equation in the Krein-Kac form. Multidimensional elliptic equations with coefficients depending only on one variable can be Laplace-transformed to the Sturm-Liouville equation, which in turn by coordinate transformation [5, § 14] can be reduced to the Krein-Kac string. We consider the Krein-Kac string equation (scalar) written in Stieltjes differentials [5] on the segment $[0, L]$, $L > 0$:

$$\lambda m(x) w - \frac{d^2 w}{dx^2} = 0, \qquad \left.\frac{dw}{dx}\right|_{x=0} = 1, \ w(L) = 0, \tag{17}$$

where $m(x)$ is a nonnegative mass distribution that can be obtained as the generalized derivative of a bounded nondecreasing (allowing discontinuities) function $M(x)$, such that $x = 0$ and $x = L$ are its points of increase [5, § 1], $\lambda \in \mathbf{C} \backslash]-\infty, 0]$. Then the impedance function of the ND of (17) can be defined as $f(\lambda) = w(0)$. If $m = 1$, then $f(\lambda)$ is defined by (2). We will keep the same notations for the potential grid, but we denote steps of the derivative grid $\hat{x}_i$ as $\dot{h}_i = \hat{x}_i - \hat{x}_{i-1}$ here. Then (17) can be approximated using the conservative three-point scheme as

$$
\begin{aligned}
\text{(18)} \qquad \lambda m_i w_i - \frac{1}{\dot{h}_i}\left(\frac{w_{i+1} - w_i}{h_i} - \frac{w_i - w_{i-1}}{h_{i-1}}\right) &= 0, \qquad i = 2, \ldots, k, \\
\lambda m_1 w_1 - \frac{1}{\dot{h}_1}\left(\frac{w_2 - w_1}{h_1}\right) &= \frac{1}{\dot{h}_1}, \qquad w_{k+1} = 0,
\end{aligned}
$$

where the grid mass m_i is the arithmetical average of $m(x)$ on $[\hat{x}_{i-1}, \hat{x}_i[$, i.e., $m_i = \frac{M(\hat{x}_i - 0) - M(\hat{x}_{i-1} - 0)}{\dot{h}_i}$. Substitution $A \to \lambda$, $\varphi \to 1$ and $\hat{h}_i \to m_i \dot{h}_i$ into (3) yields (18) and vice versa, so $w_1 = f_k(\lambda)$ as defined in (10) and we can extend the algorithm given in section 5 to (18). The Kac-Krein impedance function is a *Stieltjes* function [5, Remark 2.1], so theorem 4.1 is still valid and we again can use [3, theorem 1] if $L = \infty$ or [1, theorem 6.6.9] if $L < \infty$ showing respectively exponential or superexponential convergence of $f_k(\lambda)$ to $f(\lambda)$ for $\lambda \in \mathbf{C} \backslash]-\infty, 0]$. This extension shows that the exponential convergence of the three-point schemes can be achieved even for problems with rough coefficients. Such a surprising conclusion reflects the fact that irregularities of coefficients of differential operators do not affect the analyticity of the impedance functions.

We are indebted to David Ingerman for bringing our attention to [5].

REFERENCES

[1] G. A. Baker and P. Graves-Morris, *Padé Approximants*, Addison-Wesley Publishing Co., London et al., 1996.

[2] V. Druskin, L. Knizhnerman, *Gaussian spectral rules for the three-point second differences: I. A two-point positive definite problem in a semiinfinite domain*, Schlumberger-Doll Research, Research Note, November 1997, submitted to SIAM Num. An.

[3] A. A. Gonchar, E. A. Rakhmanov, and S. P. Suetin, *On the rate of convergence of Padé approximants of orthogonal expansions*, in Progress in approximation theory, A. A. Gonchar and E. B. Saff, Eds, Springer-Verlag, 1992, pp. 169–190.

[4] A. Iserles and G. Strang, *The optimal accuracy of difference schemes*, Trans. Amer. Math. Soc., 277 (1983), pp. 779–803.

[5] I. S. Kac and M. G. Krein, *On the spectral functions of the string*, Amer. Math. Soc. Transl., 103, (1974), pp. 19–102.

[6] B. N. Parlett, *The symmetric eigenvalue problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.