Aplikace matematiky

Jiří Kafka

An example of a set of principles for the construction of linear finite difference operators

Aplikace matematiky, Vol. 10 (1965), No. 3, 235–241

Persistent URL: http://dml.cz/dmlcz/102956

AN EXAMPLE OF A SET OF PRINCIPLES FOR THE CONSTRUCTION OF LINEAR FINITE DIFFERENCE OPERATORS

JIŘÍ KAFKA

(to topic c)

1. Physical statement of problem. The partial differential equation which is to be solved by a numerical method is usually considered as given a priori. Inquiring into the origin of this partial differential equation, in most cases, it describes a physical system. The main task is not to solve the mathematical problem, but the physical one (on the other hand, the solution of the physical problem may be auxiliary to, e.g., the development of a new machine). The algorithm for the solution of the problem should therefore start from the physical statement of the given problem. For the purpose of further discussion, all principles of physics are to be used in their integral form (viz. in that form obtained by generalising experimental results), e.g. the Gauss' theorem of electro-statics is to be used when solving potential problems. The procedure contains four steps corresponding to four sections of the presented paper.

Let us take the diffusion equation as an example of a partial differential equation describing a lamellar field. The finite-difference method will be used to solve a mixed boundary and initial value problem.

The following notation is used:

Θ ... three-dimensional domain;
Ω ... surface (boundary of Θ);
t ... time;
$J_{(v)}$ ... component of the vector of mass flow density in the direction of the external normal to Ω.
V ... volume of Θ;
S ... area of Ω;
u ... concentration [amount of mass in unit volume (usually denoted by c)];
D ... diffusion coefficient;

The phenomena under consideration are governed by the law of conservation of matter

$$\iint_{\Theta} \{u(t_2) - u(t_1)\} \, \mathrm{d}V + \int_{t_1}^{t_2} \mathrm{d}t \iint_{\Omega} J_{(v)} \, \mathrm{d}S = 0 \,. \tag{1}$$

A second physical law, the diffusion law [5], also applies

$$J = - D \operatorname{grad} u \,. \tag{2}$$

In sections 2 to 4 we shall concentrate on one-dimensional diffusion. Diffusion can be considered as one-dimensional e.g. in the case of a narrow tube with impermeable walls. The cross-section of this tube will be denoted by S_0 (fig. 1).

2. The decomposition of the domain in which the solution is to be determined. The interval $\langle a, b\rangle$ on the x-axis is divided into L intervals with lengths Δx_l ($l=1, 2, \ldots L$); the time-interval $\langle t_0, t_K\rangle$ is divided into K intervals with lengths Δt_k ($k = 0, 1, \ldots \ldots K - 1$) (fig. 2). The domain in the xt-space is thus decomposed into rectangles which will be called "elementary domains". Their total number is $N = K \,.\, L$.

A weighed directed graph which will be called "the net" is now constructed. The boundaries of these elementary domains do not from the meshes of the net. In every elementary domain a point is chosen; these points are to be the vertices of the net. The position of the vertex is determined by the requirement of minimum remainder term (therefore it is placed in the centre of the Δx_l – interval) and by the requirement of stability (therefore it is placed on the end of the Δt_{k-1} – interval). The absolute values of non-diagonal elements of the super-matrix $\mathbf{H}$ in (9) or (10) are the weights of the edges. The connection matrix of this graph is then obtained by writing a "one" for every non-zero element of $\mathbf{H}$, and then zeros in the main diagonal.

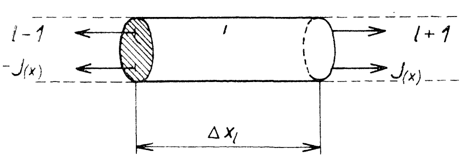

Fig. 1. One elementary interval in the diffusion tube.

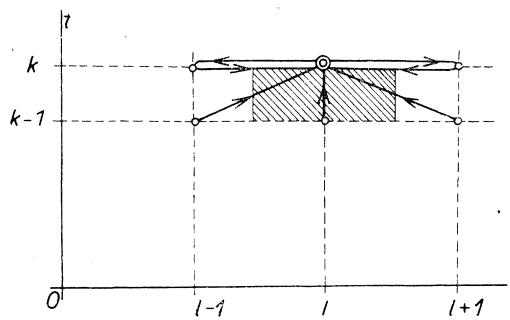

Fig. 2. The net in the neighbourhood of an internal vertex. One elementary domain is shadowed.

3. The construction of the finite difference operator at interior and boundary points. The principle (1) is applied to every elementary domain. Both integrals are approximated by quadrature formulae written with their remainder terms in footnotes[2]) and [3]). The first integral is computed according to the tangent rule[2]). The expression for the remainder term $\mathscr{R}_1$ is valid if the function $u(x, t)$ fullfils a Lipschitz condition and the following condition of second order[1])

$$\left|\Delta_{x^2}^2 u(x, t)\right| \leqq M_x^{(2)}(\Delta x)^2 \quad x \in \langle a, b) \tag{3}$$

Footnotes [1]) [2]) [3]) see next page.

with $M_x^{(2)}$ independent of x and Δx. The result of the operation is the following finite difference formula

$$(4) \qquad (u_{lk} - u_{l,k-1})\, S_0\, \Delta x_l + S_0 \mathscr{R}_1 + \int_{t_{k-1}}^{t_k} (- J_{(x)l-1,l}(t) + \\ + J_{(x)l,l+1}(t))\, S_0\, \mathrm{d}t = 0\,.$$

This equation is divided by S_0 and the integration along the t-axis is carried out according to the generalised trapezium rule[3]) (in this case the condition $|\Delta_{t^2}^2 J_{(x)}(x, t)| \leqq M_t^{(2)}(\Delta t)^2$ is assumed), $\sigma \in \langle 0, 1 \rangle$ is an arbitrary parameter,

$$(5) \qquad (u_{lk} - u_{l,k-1})\, \Delta x_l + \mathscr{R}_1 - ((1 - \sigma)\, J_{(x)l-1,lk} + \\ + \sigma J_{(x)l-1,l,k-1})\, \Delta t_{k-1} - \mathscr{R}_{2,l-1,l} + \\ + ((1 - \sigma)\, J_{(x)l,l+1,k} + \sigma J_{(x)l,l+1,k-1})\, \Delta t_{k-1} + \mathscr{R}_{2,l,l+1} = 0\,.$$

The gradient in (2) is approximated by means of the central difference formula[4]) (in addition to (3), the condition $|\Delta_{x^3}^3 u(x, t)| \leqq M_x^{(3)}(\Delta x)^3$ is required)

$$(6) \qquad J_{(x)l,l+1,k} = - (u_{l+1,k} - u_{lk})\, D_{l,l+1,k}/\Delta x_{l,l+1} - \mathscr{R}_3\,.$$

This expression is inserted into (5) and the resulting equation divided by Δt_{k-1}

$$(7) \qquad u_{lk}\Delta x_l/\Delta t_{k-1} + (1 - \sigma)\, \{(u_{lk} - u_{l-1,k})\, D_{l-1,lk}/\Delta_{xl-1,l} + \\ + (u_{lk} - u_{l+1,k})\, D_{l,l+1,k}/\Delta x_{l,l+1}\} + C_{l,l-1,k-1} u_{l-1,k-1} + \\ + C_{ll,k-1} u_{l,k-1} + C_{l,l+1,k-1} u_{l+1,k-1} + \mathscr{R}_4 = 0 \quad (l = 2, 3, \ldots L - 1)\,.$$

The remainder term $\mathscr{R}_4$ will be analysed in section 4. The symmetrical three-diagonal matrix $\mathbf{C}_{k-1}$ has in the main diagonal the elements $C_{\beta\beta k-1} = - \Delta x_\beta/\Delta t_{k-1} - \Sigma_{\gamma=1(\gamma\neq\beta)}^{L}\,.$ $.\, C_{\beta\gamma,k-1} = - \Delta x_\beta/\Delta t_{k-1} + \sigma(D_{\beta-1,\beta,k-1}/\Delta x_{\beta-1,\beta} + D_{\beta,\beta+1,k-1}/\Delta x_{\beta,\beta+1})$, in the

---

[1]) $\Delta_x u(x, t) = u(x + \Delta x, t) - u(x, t)$; $\Delta_t u(x, t) = u(x, t + \Delta t) - u(x, t)$.

[2]) $\int_a^b f(x)\, \mathrm{d}x = (b - a) f(\frac{1}{2}(a + b)) + \mathscr{R}_1$; $\mathscr{R}_1 = (b - a)\, \frac{1}{24}(b - a)^2 f''(\xi)$; $\xi \in \langle a - \frac{1}{2}(b - a), b + \frac{1}{2}(b - a)\rangle$. If the condition (3) is fullfilled, then the remainder term $\mathscr{R}_1$ may be approximated as $(b - a)\, \frac{1}{24}\Delta^2 f(a - \frac{1}{2}(b - a))$.

[3]) $\int_a^b f(x)\, \mathrm{d}x = (b - a)\, ((1 - \sigma) f(b) + \sigma f(a)) + \mathscr{R}_2$; $\mathscr{R}_2 = (b - a)\, \{(\sigma - \frac{1}{2})\, \Delta f(a) - - \frac{1}{12}(b - a)^2 f''(\xi)\}$; $\xi \in \langle a, b \rangle$. If the function $f(x)$ fullfils the condition (3), then the remainder term $\mathscr{R}_2$ can be approximated as $\mathscr{R}_2 = (b - a)\, \{(\sigma - \frac{1}{2})\, \Delta f(a) - \frac{1}{12}\Delta^2 f(a)\}$.

[4]) $(\mathrm{d}f(x)/\mathrm{d}x)_{x=b} = (f(b + \frac{1}{2}\Delta x) - f(b - \frac{1}{2}\Delta x))/\Delta x + \mathscr{R}_3$; $\mathscr{R}_3 = -\frac{1}{24}(\Delta x)^2 f'''(\xi)$; $\xi \in \in \langle b - \Delta x, b + \Delta x \rangle$. If the function $f(x)$ fullfils the following condition $|\Delta^3 f(x)| \leqq M^{(3)}(\Delta x)^3$, then the remainder term may be approximated as $\mathscr{R}_3 = -\frac{1}{24}\Delta^3 f(b - \frac{3}{2}\Delta x)/\Delta x$.

lower diagonal the elements $C_{\beta,\beta-1,k-1} = -\sigma D_{\beta-1,\beta,k-1}/\Delta x_{\beta-1,\beta}$ and in the upper diagonal the elements $C_{\beta,\beta+1,k-1} = -\sigma D_{\beta,\beta+1,k-1}/\Delta x_{\beta,\beta+1}$. According to the type of the chosen quadrature formulae, various types of finite difference formulae can be obtained; e.g. for $\sigma = 0$, (7) becomes the four-point implicit formula. The latter can be solved by means of an electrical network designed by G. Liebmann [2] [3] [4].

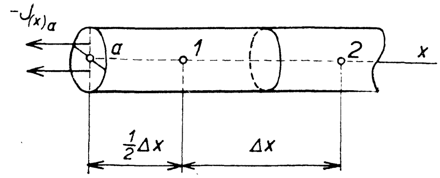

Fig. 3. The boundary condition on the left end of the tube.

If the boundary conditions of the second kind are prescribed (i.e. if the flux densities $J_{(x)a}$ and $J_{(x)b}$ are given), the finite difference equation at the points $l = 1$ and $l = L$ (vertices neighbouring to the boundary) reads

$$\begin{aligned}(8)\qquad & u_{1,k}\Delta x_1/\Delta t_{k-1} + (1-\sigma)(u_{1k} - u_{2k})\, D_{1,2,k}/\Delta x_{1,2} + \\ & + C_{1,1,k-1}u_{1,k-1} + C_{1,2,k-1}u_{2,k-1} - f_{1,k} = 0\,, \\ & u_{Lk}\Delta x_L/\Delta t_{k-1} + (1-\sigma)(u_{Lk} - u_{L-1,k})\, D_{L-1,Lk}/\Delta x_{L-1,L} + \\ & + C_{L,L-1,k-1}u_{L-1,k-1} + C_{LL,k-1}u_{L,k-1} - f_{Lk} = 0\end{aligned}$$

where $f_{1,k} = \int_{t_{k-1}}^{t_k} J_{(x)a}(t)\,\mathrm{d}t/\Delta t_{k-1}$ and $f_{Lk} = -\int_{t_{k-1}}^{t_k} J_{(x)b}(t)\,\mathrm{d}t/\Delta t_{k-1}$. It will be useful to set $f_{2,k} = f_{3,k} = \ldots = f_{L-1,k} = 0$. The finite difference equation at the points $l = 1$ and $l = L - 1$ is written whithout its remainder term. Since the boundary values are given exactly, they do not contribute to the remainder term and the remainder term of this equation is of the same order as the remainder term $\mathscr{R}_4$ of equation (7) discussed in section 4.

For $t = \text{const}$ the equations (7) and (8) represent a set of L simultaneous linear equations

$$(9)\qquad \sum_{\gamma=1}^{L} A_{\beta\gamma k}u_{\gamma k} + \sum_{\gamma=1}^{L} C_{\beta\gamma,k-1}u_{\gamma,k-1} = f_{\beta k} \quad (\beta = 1, 2, \ldots L) \quad \text{or}$$

$$\mathbf{A}_k\mathbf{u}_k + \mathbf{C}_{k-1}\mathbf{u}_{k-1} = \mathbf{f}_k$$

where $\mathbf{u}_k$ and $\mathbf{f}_k$ are column vectors and $\mathbf{A}_k$ is a three-diagonal symmetrical matrix. The elements of the main diagonal of $\mathbf{A}_k$ can be expressed as $A_{\beta\beta k} = \Delta x_\beta/\Delta t_{k-1} - \sum_{\gamma=1(\gamma\neq\beta)}^{L} A_{\beta\gamma k}$. There are K of such sets of equations and the matrix $\mathbf{H}$ of the complete set of $N = K\,.\,L$ simultaneous equations is composed of submatrices as shown below. The inverse of $\mathbf{H}$ is denoted by $\mathbf{G}$ to indicate its relation to the Green's function of the given boundary-value problem. The matrix $\mathbf{G}$ is written only for $\mathbf{A}_k$ and $\mathbf{C}_{k-1}$ independent of time (see (13))

$$(10)\qquad \mathbf{H} = \begin{bmatrix} \mathbf{A}_1 & & & & \\ \mathbf{C}_1 & \mathbf{A}_2 & & & \\ & \mathbf{C}_2 & \mathbf{A}_3 & & \\ & & \dots & \dots & \\ & & & \mathbf{C}_{K-1} & \mathbf{A}_K \end{bmatrix},$$

$$\mathbf{G} = \mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} & & & & \\ \mathbf{BA}^{-1} & \mathbf{A}^{-1} & & & \\ \mathbf{B}^2\mathbf{A}^{-1} & \mathbf{BA}^{-1} & \mathbf{A}^{-1} & & \\ \dots & \dots & \dots & & \\ \mathbf{B}^{K-1}\mathbf{A}^{-1} & \mathbf{B}^{K-2}\mathbf{A}^{-1} & \dots & \mathbf{BA}^{-1} & \mathbf{A}^{-1} \end{bmatrix}$$

where $\mathbf{B} = -\mathbf{A}^{-1}\mathbf{C}$. The form of $\mathbf{G} = \mathbf{H}^{-1}$ can be deduced in the following manner:

The simultaneous equations (9) are solved by means of the inverse of the matrix $\mathbf{A}_k$

$$(11)\qquad \boldsymbol{u}_k = \mathbf{A}_k^{-1}\boldsymbol{f}_k - \mathbf{A}_k^{-1}\mathbf{C}_{k-1}\boldsymbol{u}_{k-1} = \mathbf{A}_k^{-1}\boldsymbol{f}_k + \mathbf{B}_{k-1}\boldsymbol{u}_{k-1}\,.$$

This operation is repeated for $t = t_{k-1}$, and $\boldsymbol{u}_{k-1}$ is inserted into (11)

$$(12)\qquad \boldsymbol{u}_k = \mathbf{A}_k^{-1}\boldsymbol{f}_k + \mathbf{B}_{k-1}\mathbf{A}_{k-1}^{-1}\boldsymbol{f}_k + \mathbf{B}_{k-1}\mathbf{B}_{k-2}\boldsymbol{u}_{k-2}\,.$$

Let now Δt_k, $D_{\beta\gamma k}$ (and therefore also $\mathbf{A}_k$, $\mathbf{C}_{k-1}$, $\mathbf{B}_{k-1}$) be independent of time, $\Delta t_k = \Delta t$, $D_{\beta\gamma k} = D_{\beta\gamma}$, etc. The initial conditions $\boldsymbol{u}_0$ belong to the right hand side of (9) for $k = 1$ which reads $\boldsymbol{f}_1 - \mathbf{C}\boldsymbol{u}_0$. The following expression can be deduced by induction

$$(13)\qquad \boldsymbol{u}_k = \sum_{\lambda=0}^{k-1} \mathbf{B}^{\lambda}\mathbf{A}^{-1}\boldsymbol{f}_{k-\lambda} + \mathbf{B}^k\boldsymbol{u}_0 = \sum_{\lambda=0}^{k-2} \mathbf{B}^{\lambda}\mathbf{A}^{-1}\boldsymbol{f}_{k-\lambda} + \mathbf{B}^{k-1}\mathbf{A}^{-1}(\boldsymbol{f}_1 - \mathbf{C}\boldsymbol{u}_0)\,.$$

The expression for $\mathbf{G}$ in (10) is thus obtained.

4. Remainder terms and errors. The remainder term is deduced only for a net with regular meshes ($\Delta x_l = \Delta x$; $\Delta t_{k-1} = \Delta t$) and for $D_{l,l+1,k} = \text{const} = D$. Introducing the dimensionless quantity $\alpha = D\,\Delta t/(\Delta x)^2$, (7) is transformed into

$$(14)\qquad \Delta_t u_{l,k-1} - \alpha(1-\sigma)\,\Delta_{x^2}^2 u_{l-1,k} - \alpha\sigma\,\Delta_{x^2}^2 u_{l-1,k-1} + \mathscr{R}_5 = 0$$
$$(l = 2, 3, \dots L-1)\,.$$

Inserting the remainder terms of (4), (5) and (6), e.g.

$$\mathscr{R}_2 = \{(\sigma - \tfrac{1}{2})\,\Delta_t J_{(x)l-1,l,k-1} - \tfrac{1}{12}\Delta_{t^2}^2 J_{(x)l-1,l,k-1}\}\,\Delta t =$$
$$= \{-(\sigma - \tfrac{1}{2})\,\Delta_{xt}^2 u_{l-1,k-1} + \tfrac{1}{12}\Delta_{xt^2}^3 u_{l-1,k-1}\}\,D\,\Delta t/\Delta x -$$
$$- \{(\sigma - \tfrac{1}{2})\,\Delta_t\mathscr{R}_3 - \tfrac{1}{12}\Delta_{t^2}^2\mathscr{R}_3\}\,\Delta t$$

into (7) and supposing[5]) the approximate validity of (14), we have

$$(15) \quad \mathscr{R}_5 = \mathscr{R}_4\,\Delta t/\Delta x = \alpha\{\tfrac{1}{12} + (\tfrac{1}{2} - \sigma)\,\alpha\}\,\{(1-\sigma)\,\Delta^4_{x^4}u_{l-2,k} + \sigma\,\Delta^4_{x^4}u_{l-2,k-1}\} + \\ + \tfrac{1}{24}\alpha^2\{2\alpha(1-\sigma)^2\,\Delta^6_{x^6}u_{l-3,k+1} + (1-\sigma)(4\sigma\alpha + \sigma - \tfrac{1}{2})\,\Delta^6_{x^6}u_{l-3,k} + \\ + \sigma(2\sigma\alpha + \sigma - \tfrac{1}{2})\,\Delta^6_{x^6}u_{l-3,k-1}\} - \tfrac{1}{288}\alpha^3\{(1-\sigma)^2\,\Delta^8_{x^8}u_{l-4,k+1} + \\ + 2\sigma(1-\sigma)\,\Delta^8_{x^8}u_{l-4,k} + \sigma^2\,\Delta^8_{x^8}u_{l-4,k-1}\}\,.$$

The term $u_{lk} - u_{l,k-1}$ is exact, as only the time-dependence is concerned (we have not differentiated with respect to time). As stated in [1], the stability condition (in our case $1 + 4\alpha(\frac{1}{2} - \sigma) \geqq 0$) can be deduced from a physical principle (in our case, from the following theorem: "In the absence of diaphragms, diffusion from a locality of low concentration to one of high concentration cannot occur".).

The propagation of errors[6]) is determined in accordance with Bruyevitch's theory of accuracy of computing mechanisms and linkages in the modified reading of [2] and [8] by using the $\mathbf{G}$ – matrix. The error $\vartheta(A_{\zeta\eta})$ of the element $A_{\zeta\eta}$ of $\mathbf{A}$ when used in the s-th time step causes an error $\vartheta(u_{\beta k})$ of the result $u_{\beta\gamma}$ in the k-th time step. The matrix $\mathbf{A}^{-1}$ is denoted by $\mathscr{F}$, the elements of $\mathbf{B}^k$ are denoted by $[B^k]_{\beta\gamma}$,

$$(16) \qquad \vartheta(u_{\beta k}) = \Big(\sum_{\gamma=1}^{L}[B^{k-s}]_{\beta\gamma}\,F_{\gamma\eta} - \sum_{\gamma=1}^{L}[B^{k-s}]_{\beta\gamma}\,F_{\gamma\zeta}\Big)\,J_{(x)\eta\zeta s}\vartheta(A_{\eta\zeta})/A_{\eta\zeta}$$
$$(k, s = 1, 2, \dots K;\ \beta, \eta = 1, 2, \dots L;\ \zeta = \eta - 1 \text{ or } \zeta = \eta + 1)\,.$$

The error of the result caused by errors $\vartheta(u_{l0})$ of the initial values and by errors $\vartheta(J_{(x)ak})$ and $\vartheta(J_{(x)bk})$ of the boundary values can be calculated by substituting $\vartheta(u_{l0})$ for u_{l0} and $\vartheta(J_{(x)ak})$ or $-\ \vartheta(J_{(x)bk})$ for f_{lk} into (13).

References

[1] *G. M. Dusinberre:* Numerical Analysis of Heat Flow. McGraw-Hill, New York 1949.

[2] *J. Kafka:* Synthesis of finite difference analogues[7]). Information processing machines (Stroje na zpracování informací) **10**, pp. 267—284. Nakladatelství Československé akademie věd, Praha 1964.

[3] *J. Kafka:* A finite difference analogue[7]) of the quasi-stationary electromagnetic field. To be published in Information processing machines **12**.

[4] *G. Liebmann:* The solution of transient heat flow and heat transfer problems by relaxation. Brit. J. Appl. Phys. **6** (1955), No. 4, pp. 129—134.

---

[5]) This means that the equivalence of the finite difference operators $\Delta_t = \alpha\{(1-\sigma)\,\mathbf{E}_t\Delta^2_{x^2} + \sigma\Delta^2_{x^2}\}$ is assumed, where $\mathbf{E}_t\,u(x, t) = u(x, t + \Delta t)$.

[6]) These errors may be round-off errors or inaccuracies of quantities calculated from experimental data, e.g. $\vartheta(A_{\beta\gamma})/A_{\beta\gamma}$ may vary from 0.01% up to 1%.

[7]) In [2] and [3] the term "finite difference analogue" is used as an abbreviation for "analogue computer designed for the solution of partial differential equations by means of finite difference methods".

[5] *R. A. Robinson - R. H. Stokes.:* Electrolytic solutions. Butterworths, London 1959.

[6] *В. К. Саульев:* Интегрирование уравнений параболического типа методом сеток (V. K. Saulyev: The Solution of Equations of the Parabolic Type by Finite Difference Methods). Государственное издательство физико-математической литературы, Москва 1960.

[7] *А. А. Самарский-А. Н. Тихонов:* Уравнения математической физики (A. A. Samarskii-A. N. Tykhonov: Partial Differential Equations of Mathematical Physics). Государственное издательство физико-математической литературы, Москва 1951.

[8] *M. L. Bychowski:* Die Genauigkeit elektrischer Netze, die zur Lösung von Laplace-Gleichungen vorwendet werden. Verlag Technik, Berlin 1952. *(Transloted from Russian.)*

Jiří Kafka, Matematická laboratoř ČVUT, Horská 4, Praha 2, ČSSR.