Varin V.P.

An analysis of a vacuum diode model

ISSN 2071-2898

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Российской академии наук

V.P. Varin

An analysis of a vacuum diode model

Препринт №

Москва 2012

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ОРДЕНА ЛЕНИНА ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
ИМЕНИ М.В. КЕЛДЫША

V.P. Varin

AN ANALYSIS OF A VACUUM DIODE MODEL

Moscow, 2012

УДК 521.1+531.314

V.P. Varin. An analysis of a vacuum diode model. Preprint of the Keldysh Institute of Applied Mathematics of RAS, Moscow, 2012.

One of the vacuum diode models is written as a singular boundary value problem for 2 ODEs of the 2nd order with one parameter. The parameter is unknown and must be found along with the initial conditions at the origin to satisfy the boundary conditions at the end of the interval. The problem presents considerable difficulties in its theoretical and, especially, numerical study. Here we give a complete analytical and numerical solution of this problem.

В.П. Варин. Анализ модели вакуумного диода. Препринт Института прикладной математики им. М.В. Келдыша РАН, Москва, 2012.

Одна из моделей вакуумного диода представлена сингулярной краевой задачей для 2 ОДУ 2-го порядка с одним параметром. Параметр является неизвестной величиной и подлежит определению вместе с начальными данными в нуле так, чтобы выполнялись краевые условия на конце интервала. Проблема представляет значительные трудности как теоретического, так и, особенно, численного характера. Здесь мы даем полное аналитическое и численное решение этой задачи.

Работа поддержана Российским фондом фундаментальных исследований, грант 11-01-00023.

E-mail: varin@keldysh.ru
http: www.keldysh.ru
(see electronic library/ catalogue of publications/ preprints).

§ 1. Introduction

In this paper we analyse a vacuum diode model which appears in the problem of magnetic insulation in the study of plasma. The model is written as a boundary value problem for two second order ODEs:

$$
\begin{aligned}
\frac{d^2f}{dx^2}(x) &= j\frac{1+f(x)}{\sqrt{(1+f(x))^2-1-a^2(x)}},\\
\frac{d^2a}{dx^2}(x) &= j\frac{a(x)}{\sqrt{(1+f(x))^2-1-a^2(x)}},
\end{aligned}
\tag{1}
$$

with the initial and boundary conditions

$$
\begin{aligned}
&f(0)=0, \quad a(0)=0, \quad f'(0)=0 \quad a'(0)=C,\\
&f(1)=f_1, \quad a(1)=a_1.
\end{aligned}
\tag{2}
$$

Here f is the electrostatic potential, a is the magnetic potential, j is the current (the parameter in the problem). The problem is considered on the interval $x \in [0,1]$. Given the positive values of both potentials $f(1)=f_1$ and $a(1)=a_1$ at the end of the interval $x=1$, the initial value $a'(0)=C>0$ and the parameter $j>0$ must be found such that the boundary conditions (2) be satisfied.

The author of this paper became familiar with this problem at a conference, in a private communication, as it is stated above, i.e., in a purely mathematical form. We will keep it this way and abstain here from any physical interpretation of the problem and the obtained results. For the derivation of the original problem, its rather rich history, and physical meaning and consequences of our study, we refer the reader to the paper [1] with the nested references there.

We use Latin instead of Greek letters for typesetting and portability reasons ($f=\varphi$, and $C=\beta$, $j=j_x$ in [1]). The author have found the paper [1] on the internet only after this research was almost completed. This is another reason for keeping our notation.

The singular boundary value problem (BVP) (1)–(2) was never fully investigated. In particular, it is unknown which values f_1 and a_1 can be or which cannot be attained. The same is true for the parameters C and j, i.e., the boundary of admissible values was never specified. For example: whether the values $C=3$ and $j=1$ are good for any positive boundary values f_1 and a_1? (The answer is no.)

It is also not exactly clear how this problem should be treated numerically. Since the equations (1) are singular at the origin, i.e., the left hand sides of the equations (1) become infinity, no standard algorithm for numerical integration of ODEs is immediately suitable. Some attempts to solve the problem (1)–(2) with the shooting technique seem dubious at best if the problem of singularity is not addressed.

Further, suppose we can solve the problem numerically for some boundary values. How can we guarantee that our algorithm would not fail for some other boundary values? Or, if it fails, then where? And how many solutions can exist for the given boundary values? These are some of the questions that were never even posed for this problem.

As it will become apparent, the difficulty with the problem (1)–(2) lies with the fact that it cannot be studied satisfactorily either analytically or by purely numerical means. In this paper we will combine both approaches and give a complete solution to the above problem. By complete we mean that we can answer any reasonable question about the solutions to the problem (1)–(2), and have an effective way to compute solutions numerically with (in principle) an arbitrary precision.

This study can be summarized as follows.

We give an analytical solution to the initial value problem (IVP), i.e., we integrate the equations (1) explicitly. As it turned out, most of this work was already accomplished in the paper [1]. But here we move further and write the solution in a simple and convenient way using elliptic integrals.

Both IVP and BVP for the equations (1) are reduced to some nonlinear equations expressed in canonical elliptic integrals. Using this representation, we give an explicit analytical description of the domain of *admissible values* $A = \{C, j\}$, and of the domain of *attainable values* $B = \{f_1, a_1\}$, i.e., a solution to the IVP can reach the end of the interval $x = 1$ if and only if (iff) C, j are in A; and a solution to the BVP exists iff f_1, a_1 are in B.

Along the way, both domains A and B are split into sectors where different sets of elliptic integrals express the solutions. These integrals depend, in general, on complex arguments, so their numerical evaluation is still a problem (successfully solved).

Thus the IVP (1)–(2) gives a map $M\colon A \to B$, where $B = M(A)$ by definition. We study this map analytically and numerically.

First, we develop an effective way to compute the map M without the use of elliptic integrals, and without the solution of nonlinear equations with Newton iterations. This algorithm is used for the plots and for verification of the analytical solution.

To resolve the problem of singularity, we expand the solution to the IVP in some convergent series at the origin, or, alternatively, we use analytical solution to obtain initial values at some point $x = x_0 > 0$, where the solution is analytical. Here we use a very simple system of ODEs instead of (1). Then we integrate this system numerically for the remaining part of the solution, i.e., for $x \in [x_0,\, 1]$, with a high precision Runge-Kutta algorithm. After the solution is obtained, we can verify it (and refine it if needed) with an arbitrary precision using its analytical representation.

Numerical experiments revealed that the map M is bijective and non-degenerate, i.e., the inverse map $M^{-1}\colon B \to A$ exists. In particular, there are no bifurcations and no multiple solutions for the IVP or BVP. We give a rigorous proof of this fact, and prove that the map M is a diffeomorphism. This means that both IVP and BVP can be solved for any admissible values. We demonstrate on some examples how to do this both ways with a guaranteed success, since each problem is reduced to a solution of a nonlinear equation on an interval where the solution exists.

§ 2. Analytical solution of the problem

Let us find first integrals of the system (1). We multiply the first and the second equations (1) by $df(x)/dx$ and $da(x)/dx$ respectively; then we integrate both equations. After subtracting the second integral from the first, we obtain

$$C_1 = 2\,j\sqrt{(1+f(x))^2 - 1 - a^2(x)} - \left(\frac{d}{dx}f(x)\right)^2 + \left(\frac{d}{dx}a(x)\right)^2, \qquad (3)$$

where C_1 is the constant of integration. Using boundary conditions (2), we see that $C_1 = C^2$, but for now, it is arbitrary.

Then we isolate the denominators in both equations (1), equate them and integrate this equation. We obtain the second first integral

$$C_2 = \frac{d}{dx}a(x)\,f(x) - a(x)\frac{d}{dx}\,f(x) + \frac{d}{dx}a(x), \qquad (4)$$

where C_2 is the constant of integration. Using boundary conditions (2), we see that $C_2 = C$, but for now, it is arbitrary.

Now we introduce an intuitive change of variables that (as it happened) was invented a long time ago (see [1] and nested references there). We put

$$f(x) = (r(x)+1)\cosh t(x) - 1, \quad a(x) = (r(x)+1)\sinh t(x). \qquad (5)$$

Here we do not make any assumptions with respect to the new functions $r(x)$ and $t(x)$. All their properties will be derived from the equations.

The Jacobian of the map $(f, a) \to (r, t)$ given by the equations (5) is equal to $r + 1$, and so the map is reversable for $r > -1$.

It is immediately seen that the initial conditions (2) translate

$$r(0)=0, \quad t(0)=0, \quad r'(0)=0 \quad t'(0)=C. \tag{6}$$

The inverse change of variables is given by the formulas

$$r(x)=\sqrt{(1+f(x))^2-a^2(x)}-1, \quad t(x)=\frac{1}{2}\log\left(\frac{1+f(x)+a(x)}{1+f(x)-a(x)}\right). \tag{7}$$

Note that $0\le t(x)=\operatorname{arctanh}(a(x)/(1+f(x)))$ in (7), and so must be $a(x)<1+f(x)$.

The expression under the square root in the equations (1) must not be negative for small $x>0$, otherwise no real solutions to the problem (1)–(2) exist. This implies $r(x)>0$ for at least small $x>0$. The denominator vanishes at the origin, and it can vanish again at some $x_*>0$. This would mean that the solution to the system (1) can not be continued beyond x_*.

Now we use the substitution (5) for the equations (1) and, after eliminating hyperbolic sines and cosines, we obtain

$$\begin{gathered}\frac{d^2}{dx^2}r\left(x\right)+\left(\frac{d}{dx}t\left(x\right)\right)^2\left(r(x)+1\right)=j\,\frac{\left(r\left(x\right)+1\right)}{\sqrt{r(x)(r(x)+2)}},\\ \left(\frac{d^2}{dx^2}t(x)\right)\left(r(x)+1\right)+2\left(\frac{d}{dx}t(x)\right)\left(\frac{d}{dx}r(x)\right)=0.\end{gathered} \tag{8}$$

From the second equation in (8), we find

$$t(x)=C_3\int_0^x\frac{1}{(r(s)+1)^2}\,ds+C_4. \tag{9}$$

The boundary conditions (2) imply $C_3=C$ and $C_4=0$, but we need the general solution for now. Substituting the equation (9) into the first equation in (8), we obtain

$$(r(x)+1)^3\frac{d^2}{dx^2}r(x)+C_3^2=j\,\frac{(r(x)+1)^4}{\sqrt{r(x)(r(x)+2)}}. \tag{10}$$

Now we make another intuitive change of variable in the equation (10)

$$r(x)=\sqrt{1+w^2(x)}-1. \tag{11}$$

In the paper [1], the function $w^2(x)$ was called the *effective potential.* We can assume $w(x)\ge 0$. At the origin, $w(0)=0$. As it was explained, the solution to the system (1) exists until the next zero $w(x_*)=0$, or up to $x=\infty$.

After the substitution (11), the equation (10) transforms into

$$(w(x)+w^3(x))\frac{d^2}{dx^2}w(x)+\left(\frac{d}{dx}w(x)\right)^2-2jw(x)-jw^3(x)+C_3^2=\frac{j}{w(x)}, \tag{12}$$

and the first integral (3) transforms into

$$\left(\frac{d}{dx}w(x)\right)^2=\frac{2\,j\,w^3(x)+2\,j\,w(x)+C_3^2-C_1\,(1+w^2(x))}{w^2(x)}. \tag{13}$$

The first integral (4) simplifies to $C_2=C_3$, and so it is not very useful.

The equation (12) is integrable, and it can be used for the general solution of the problem, which we do not need. So from now on $C_1=C^2$ and $C_3=C$.

Thus the equation (13) simplifies to

$$\left(\frac{d}{dx}w(x)\right)^2=2\,j\left(w(x)+\frac{1}{w(x)}\right)-C^2, \tag{14}$$

and the equation (12) (with the use of (14)) simplifies to

$$\frac{d^2}{dx^2}w(x)=j\left(1-\frac{1}{w^2(x)}\right). \tag{15}$$

The equation (9) (that we need in a differential form) transforms into

$$\frac{d}{dx}t(x)=\frac{C}{1+w^2(x)}. \tag{16}$$

It is clear that the equation (14) is the first integral of the equation (15), and so C is the constant of integration as well as initial condition in (2).

The equation (14) has the first integral

$$x(w)=\int_0^w\frac{\sqrt{s}}{\sqrt{2\,j-C^2\,s+2\,j\,s^2}}\,ds=\frac{1}{\sqrt{2\,j}}\int_0^w\frac{\sqrt{s}}{\sqrt{(k-s)(1/k-s)}}\,ds, \tag{17}$$

and the equation (16), with the help of (14), has the first integral

$$t(w)=\left(k+\frac{1}{k}\right)^{1/2}\int_0^w\frac{\sqrt{s}}{(1+s^2)\sqrt{(k-s)(1/k-s)}}\,ds, \tag{18}$$

where

$$k=\frac{4\,j}{C^2+\sqrt{C^4-16\,j^2}}. \tag{19}$$

The constant k will be used as modulus for elliptic integrals. Thus we can formulate

Theorem 1. *Either k is real, and then $0 < k \leq 1$ for $0 < j \leq C^2/4$, or k is complex, $|k| = 1$ for $j > C^2/4$. For $j \geq C^2/4$, both solutions (17) and (18) are continued indefinitely, i.e., w, x, and t are unbounded. If $j < C^2/4$, i.e., if $k < 1$, then the solutions (17) and (18) are valid for $0 \leq w \leq k$, i.e., w attains its maximal value $w = k$, and the solutions (17) and (18) are continued from this point as*

$$x_+(w) = 2\,x(k) - x(w) \quad \text{and} \quad t_+(w) = 2\,t(k) - t(w), \tag{20}$$

where w decreases form $w = k$ to 0, i.e., $0 \leq x \leq 2\,x(k)$ and $0 \leq t \leq 2\,t(k)$.

The proof is obvious.

The equation (18) and its counterpart in (20) reduce the solution of the BVP (1)–(2) to solution of one nonlinear equation on an interval. We find $t = t(1)$ and $w = w(1)$ at $x = 1$ by the formulas (7) and (11). Then we have to find the value of k; then we find j from the equation (17) or its counterpart in (20), and, finally, we find C from the equation (19). However, we do not know yet how to decide if k is real or complex, and which equations, (17) and (18), or (20) are needed. We solve these problems in Section 4.

For $k < 1$, we define the family of curves L_g as

$$L_g = \{C, j\}\colon\ x(k) = g, \quad 1/2 \leq g \leq \infty. \tag{21}$$

It is clear that the boundary of the domain of admissible values A for the IVP is given by the curve $L_{1/2}$. On the curve L_1, the maximal value of $w = k$ is attained at $x = 1$, and so it is the boundary of the subdomain in A where we have to switch from the integrals (17), (18) to the integrals (20) (or vise versa). It will be proved that $L_\infty = \{j = C^2/4\}$.

We can plot the image of the curve $L_{1/2}$, i.e., the boundary of the domain B of attainable values, without knowing anything yet about the curve $L_{1/2}$. We denote $M(L_g) = N_g$. Since $w = 0$ on the boundary of B, then its parametric representation is given by $f_1 = \cosh(t) - 1$, $a_1 = \sinh(t)$. Eliminating t, this writes $N_{1/2} = \{a_1 = \sqrt{f_1\,(f_1 + 2)}\}$. This curve lies below its asymptote $a_1 = f_1 + 1$ (see (7)), and so the domain B is below the curve $N_{1/2}$ (unless it is between $N_{1/2}$ and its asymptote, which is not true).

There is not much that can be done further without the computation of the integrals (analytical and numerical). These integrals can be evaluated numerically for the given values of C, j, and w without much trouble. But to solve the IVP or BVP, we have to apply Newton iterations to the nonlinear equations depending on these integrals, which is very costly. And we have to decide which of these integrals to use. All these technical problems will be solved with the use of elliptic integrals. However, it is useful to have an alternative way to solve the problem without the use of quadratures altogether. It is done in the next section.

§ 3. Numerical solution of IVP and BVP

The equations (14) (or (15)) and (16) can be used to obtain power expansions of solutions $w = w(x)$ and $t = t(x)$ at the origin, and initially it was done in this way. However, it proved to be simpler to obtain power expansions $x = x(w)$ and $t = t(w)$ from the integrals (17), (18). The latter expansions converge faster, and they are more compact. Thus we obtain

$$\begin{aligned} x(w) &= \frac{w^{3/2}}{\sqrt{2\,j}} \left(\frac{2}{3} + \frac{h\,w}{5} + \frac{w^2}{28}\,(3\,h^2 - 4) + \frac{h\,w^3}{72}\,(5\,h^2 - 12) + \ldots \right), \\ t(w) &= h^{1/2}\,w^{3/2} \left(\frac{2}{3} + \frac{h\,w}{5} + \frac{3\,w^2}{28}\,(h^2 - 4) + \frac{5\,h\,w^3}{72}(h^2 - 4) + \ldots \right), \end{aligned} \tag{22}$$

where $h = k + 1/k = C^2/(2\,j)$.

We developed the series (22) up to the terms w^{20}. Written in Horner form (both in h and w), these series are used to obtain initial values x_0 and t_0 at a fixed value w_0. Then the ODEs (15) and (16) are integrated together as was described in the introduction. The equation (16) is used in the system of ODEs in order to avoid the computation of quadratures, and the equation (15) is used instead of (14), since it does not require switching the branches (see Theorem 1). The remaining initial value $w'(x_0)$ is computed by the equation (14).

Given the values C and j, the ODEs (15) and (16) are integrated over x from $x = x_0$ until $x = 1$, if it is possible, or until the next value $w_* = 0$, for which $x(w_*) < 1$. In the first case, the IVP is solved, i.e., the point $(C, j) \in A$; otherwise, the point $(C, j) \notin A$ and discarded.

The values f_1 and a_1 are found uniquely by the values $w(1) > 0$ and $t(1) > 0$ with the formulas (11) and (5). Fig. 1 shows some typical examples of the computed functions when the initial values are below ($C = 1$, $j = 0.14$) or higher ($C = 3$, $j = 3$) than the curve $\{j = C^2/4\}$.

The plots of the functions $f(x)$ and $a(x)$ look benign, but with the expansions (22), we find

$$f(x) = \frac{3}{4}\,6^{1/3}\,j^{2/3}\,x^{4/3} + \frac{C^2\,x^2}{20} + \frac{9}{5600}\,6^{2/3}\,\frac{C^4 + 25\,j^2}{j^{2/3}}\,x^{8/3} + o(x^3),$$

$$a(x) = C\,\left(x + \frac{3}{28}\,6^{1/3}\,j^{2/3}\,x^{7/3} + \frac{1}{60}\,C^2\,x^3 + o(x^3) \right),$$

so the second derivative $f''(x) \to +\infty$ as $x \to +0$, and numerical integration can not start at the origin.

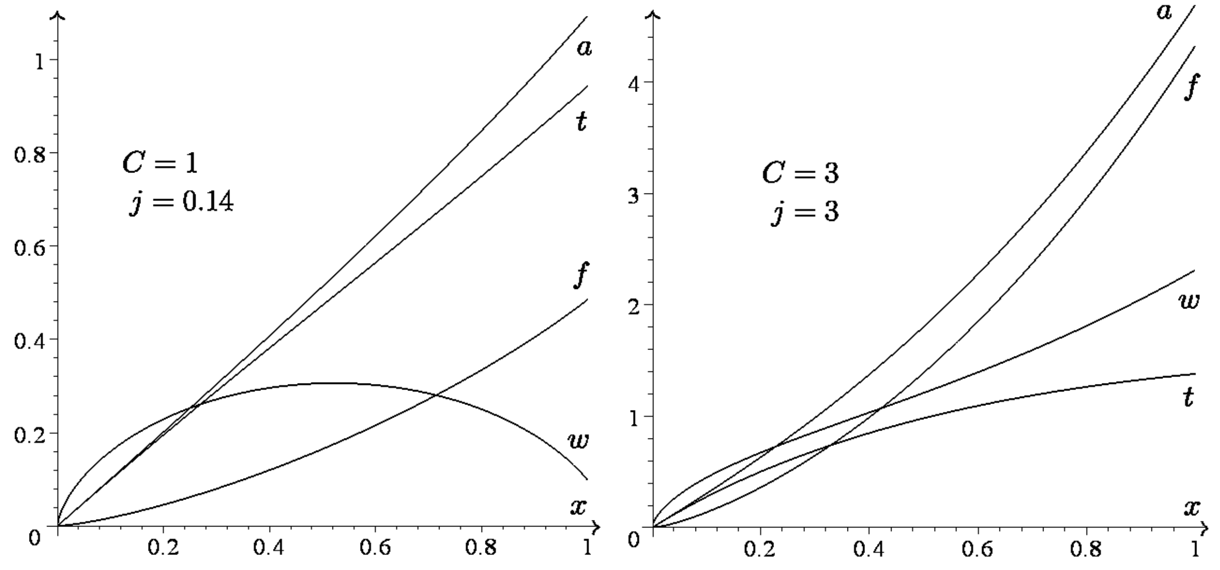

Fig 1. Typical plots of the functions $w(x)$, $t(x)$, $f(x)$, and $a(x)$.

We selected a mesh of points in the rectangle $(C,\, j) \in [0..10, 0..20]$, with the stepsize 1/10 in C, and 1/5 in j (10000 points). For each $(C,\, j)$ in the mesh, we either found the boundary values f_1 and a_1 as described above, or computations were aborted if $w'(x) < 0$ for $x < 1/2$ (so the program does not have to encounter another singularity). These computations were also performed for some fixed values of C for a denser mesh. The results are shown on Fig. 2.

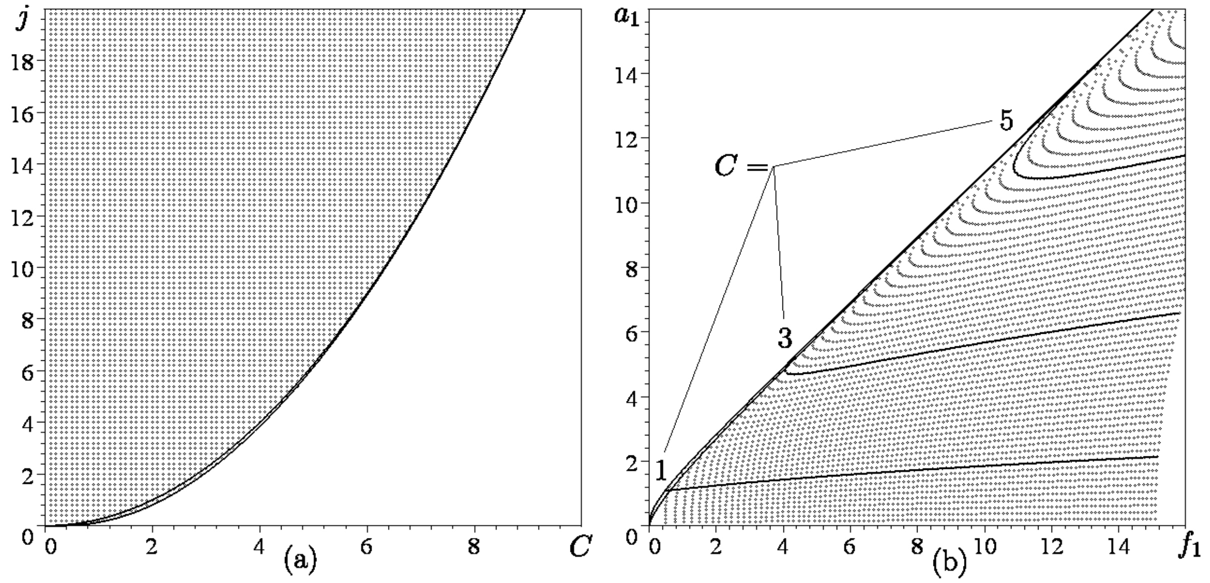

Fig 2. Domains A, B, and the map $M\colon A \to B$.

Fig. 2 requires some commentaries. On Fig. 2 (a), there are two curves in addition to the mesh points. The lower curve is $L_{1/2}$, i.e., the boundary of the domain A, that we borrowed from the next section. We could do without, since the wrong points are discarded automatically. The upper curve is $\{j = C^2/4\}$, that we promised to prove is L_∞. These two curves are mapped on Fig. 2 (b) into $N_{1/2}$, i.e., the boundary of the domain B, that we already know explicitly, and into N_∞, that will be given in the next section. So the curves of the family (21), and of its image in B, are packed very closely. The three curves on Fig. 2 (b) correspond to $C = 1$, $C = 3$, and $C = 5$ on Fig. 2 (a). Only a part of the mapped mesh is shown on Fig. 2 (b), and the image of the edge of the mesh, i.e., $\{j = 20\}$, is also visible.

Now we have enough material to prove the following

Theorem 2. *The map $M\colon A \to B$ is a local diffeomorphism.*

Proof. The integrals (17), (18) depend analytically on C and j, hence the series (22) as well. We take a point $p = (C, j) \in A$, and a sufficiently small w, that is fixed. Then the series (22) give a local diffeomorphism of a neighborhood of p into the neighborhood of the point $(x(w), t(w)) \in \mathbb{R}$. To prove the latter statement, we compute the Jacobian matrix of the map $(C, j) \to (h, j)$, and its determinant, which equals $d_0 = \sqrt{2\,h/j}$; then we compute the Jacobian matrix of the map $(h, j) \to (x(w), t(w))$ given by the series (22), and its determinant d_1. Then we expand the Jacobian $d = d_0\,d_1$ in power series in w. After this rather bulky calculation, we obtain

$$d = \frac{1}{9\,j^2}\,w^3 + \frac{2\,h}{15\,j^2}\,w^4 + \frac{2\,(36\,h^2 - 25)}{525\,j^2}\,w^5 + \ldots \tag{23}$$

Hence d is positive for small w.

The solutions to ODEs (15) and (16) depend analytically on initial values and parameters from the neighborhood of the point $x(w)$ until the end of the interval $x = 1$, which they reach, since $p \in A$. So we have another diffeomorphism due to standard theorems on existence and uniqueness of solutions to ODEs, and their analytical dependence on initial values and parameters. Thus M is a superposition of three diffeomorphisms, hence the statement of the theorem. ■

Theorem 2 can not guarantee the absence of multiple solutions to the BVP. On the other hand, Fig. 2 strongly suggests that the map M is bijective. The only problem here presents the bundle of curves (21) and its image, that may hide some surprises. We can blow up these regions with a suitable change of variables and investigate further, but it is better to postpone this until the next section.

Now we give some necessary technical details related to computations.

Unless stated otherwise, all computations are performed with the standard double float arithmetic (about 16 decimal places). For all points in Fig. 2, we take $w_0 = 0.1$, compute x_0 and t_0 as described above, then integrate the system (15), (16) over the interval $[x_0, 1]$ with a fixed stepsize $(1 - x_0)/1024$ using a Runge-Kutta integrator of the 8th order. It seems like a lot of computations, but it took about 6 seconds of CPU time on a personal computer (6002 points are visible on the left of Fig. 2).

We can estimate the accuracy of computations in the IVP by the final result, i.e., by the values f_1 and a_1, and how they vary depending on the settings of the algorithm. The series (22) were developed with a safety margin aiming at computations with extended precision. And a Runge-Kutta integrator of the 8th order is very accurate for regular functions. As an example, we computed the values f_1 and a_1 for C and j given on Fig. 1 for three different settings:

a) $w_0 = 1/8$, stepsize $(1 - x_0)/1024$,
b) $w_0 = 1/8$, stepsize $(1 - x_0)/4096$,
c) $w_0 = 1/16$, stepsize $(1 - x_0)/4096$.

The results are presented in Table 1.

Table 1.

C, j	f_1	a_1	$\Delta(a, b, c)$
$C = 1$, $j = 0.14$	0.48552808946968	1.09412288422309	7×10^{-13}
$C = 3$, $j = 3$	4.32596753511663	4.69258380040852	5×10^{-15}

Here $\Delta(a, b, c)$ is the maximal discrepancy in the values f_1, a_1 with the three settings a)–c). Later we will use independent error estimates based on exact solutions.

Now we turn to the solution of the BVP. As it was mentioned, the equations (15) and (16) do not care where the point is located: as long as the point is inside the domain A, we can reach the end of the interval $x = 1$. According to Fig. 2, we can expect the same for the solution of the BVP, i.e., if a point is taken inside the domain B, we can find a unique solution inside the domain A (it is not proven yet). But we have to pay a price: we need to solve a system of two equations with Newton iterations instead of only one equation, as Theorem 1 suggests.

However, technically, it is not a problem. If a point in A is taken close enough to the solution, then the Newton iterations converge quadratically. So the modus operandi should be: consult the table corresponding to Fig. 2 (b) for the closest match for the given boundary values f_1, a_1; then find the corresponding best approximation to C, j in the table of Fig. 2 (a); then apply Newton iterations to the system $\{t(1) = t(f_1, a_1),\ w(1) = w(f_1, a_1)\}$, i.e., use the shooting technique, with the algorithm described above.

We performed these computations for three points in various places of the domain B. The results are presented in Table 2.

Table 2.

f_1, a_1	C	j
$f_1 = 1$, $a_1 = 1$	0.87984323357317	0.53372981508599
$f_1 = 8$, $a_1 = 3$	1.72888497635265	8.93296912808290
$f_1 = 0.3$, $a_1 = 0.8$	0.75910968591211	0.07610136648478

Only the third line in Table 2 required some adjustment of the first approximation, since this point falls into the bundle of curves (21), and the mesh there is not dense enough. Other two cases converged in 4–5 iterations.

Numerically, the problem (1)–(2) can be considered as solved, since it can be tabulated with arbitrary accuracy. However, several important questions remained unanswered.

§ 4. Solution of the problem in elliptic integrals

The integrals (17) and (18) are not very suitable for numerical evaluation for a number of reasons, that we would not discuss here. The canonical elliptic integrals, on the other hand, are computed almost as easily as an elementary function. Unfortunately, the integrals (17) and (18) can be expressed through the elliptic ones in a number of ways, some of which are outright ugly and probably useless. We believe, we found one of the simplest representations:

$$x(w) = \left(\frac{2}{k\,j}\right)^{1/2} \left(F\left(\sqrt{w/k}, k\right) - E\left(\sqrt{w/k}, k\right)\right), \tag{24}$$

$$t(w) = 2\,\mathrm{Im}\left(\sqrt{1+k^2}\,\Pi\left(\sqrt{w/k}, i\,k, k\right)\right), \tag{25}$$

where $i^2 = -1$. Here F, E, and Π are incomplete (or complete, depending on arguments) elliptic integrals of the first, second, and third kind respectively in Legendre normal form [2, page 859].

We do not need to prove the formulas (24) and (25), since the integrals (17) and (18) are derived backwards from these formulas much easier than vise versa.

Both integrals (24) and (25) give real values for any admissible values of parameters, i.e., if $k \in \mathbb{R}$, then $0 < w \le k < 1$; and if $k \in \mathbb{C}$, then $w > 0$ is arbitrary, but $k = \exp(i\,s)$, $0 < s < \pi/2$, since $h > 0$, and $\overline{k}$ is as good as k. Note also that the Im() function appeared in (25) because we omitted the complex conjugate part of the formula.

We verified the determinant (23) with the formulas (24) and (25), where the two diffeomorphisms now are $(C, j) \to (k, j)$ and $(k, j) \to (x(w), t(w))$. Ironically, the computations with the series (22) are much simpler.

Now we can explore the bundle of curves (21), which we temporarily expand to $0 \le g \le \infty$.

For $w = k$, incomplete elliptic integrals in (24) and (25) become complete, and the curves L_g are written parametrically as

$$L_g = \left\{ C = \frac{2\sqrt{k^2+1}}{k\, g}\, (K(k) - E(k))\, ,\ j = \frac{2}{k\, g^2}\, (K(k) - E(k))^2 \right\}, \tag{26}$$

where $k \in [0, 1)$ is the parameter on the curve L_g.

The curves (26) have the same asymptotics at the origin

$$j = \frac{g}{\pi} C^3 - \frac{15}{2} \frac{g^3}{\pi^3} C^5 + 90 \frac{g^5}{\pi^5} C^7 + \dots, \tag{27}$$

which is obtained by excluding k from the respective series for C and j.

It is easy to prove that on L_g, both C and j are increasing functions of k. For this, we need to differentiate their expressions with respect to k, then convert the result into ordinary integrals (we omit the details).

We denote $q = k/(2\,(k^2+1))$. Since $j/C^2 = q < 1/4$ on the curves (26), they all lie below the curve $\{j = C^2/4\}$ and have the same asymptotics at infinity.

Since $q(k)$ is an increasing function of $k \in [0, 1)$, it follows that the curve L_u lies below the curve L_v (we write $L_u < L_v$), iff $u < v$. In particular, the curves (26) can not intersect except at the origin and at infinity.

Finally, for any point $p_0 = (C_0, j_0)$ below the curve $\{j = C^2/4\}$, we find $q_0 = j_0/C_0^2 < 1/4$, solve the equation $q(k) = q_0$ for k, then find g from either equation (26). Thus we proved

Theorem 3. *The family of curves* (26) *forms analytical foliation of the region* $(C, j) \in \{j < C^2/4;\ 0 < C,\ 0 < j\}$.

It follows that $L_\infty = \{j = C^2/4\}$, as we promised, and $L_0 = \{j = 0\}$.

As we are interested in curves (21), Theorem 3 applies there as well, i.e., exactly one curve of the family (21) passes through any point between the curves $L_{1/2}$ and L_∞.

The last paragraph before Theorem 3 gives the recipe for deciding where a point $p_0 = (C_0, j_0)$ lies:

(a) if $q_0 > 1/4$, then k is complex, and $p_0 > L_\infty$, i.e., the point is above the curve L_∞.

(b) if $q_0 < 1/4$, $g > 1$, then k is real, $k < 1$, and $L_1 < p_0 < L_\infty$.

(c) if $q_0 < 1/4$, $1/2 < g < 1$, then $k < 1$, and $L_{1/2} < p_0 < L_1$.

(d) if $q_0 < 1/4$, $g < 1/2$, then $p_0 < L_{1/2}$, i.e., $p_0 \notin A$.

Let us apply this knowledge to Fig. 1 and verify the corresponding Table 1 with explicit formulas. From now on, all computations are performed with extended precision (32 and more decimal places). We will discuss numerics at the end of this section, but here we only remark that all digits are correct in the results given below.

Example 1. We take $C_0 = 1$, $j_0 = 0.14$; hence $q_0 = 0.14$, and $k = 0.306263201628$. Then we find $g = 0.521944611384$ from (26). Hence we have the case (c), and we can verify that $g = x(k)$ by the formula (24). Since we are below the curve L_1, we have to use the formulas (20) (see Theorem 1). So we solve the equation $2\,g - x(w) = 1$ for $w \in [0, k]$, and find $w = 0.098431797830$. Now with (25), we find $t = 2\,t(k) - t(w) = 0.942833016664$. Knowing t and w, we find $f_1 = 0.48552808947026010217$ and $a_1 = 1.09412288422323116107$ as described in Section 3. ■

Example 2. We take $C_0 = 3$, $j_0 = 3$; hence $q_0 = 1/3$, and we have the case (a), i.e., k is complex, $k = 0.75 + i\,0.661437827766$. Since we are above the curve L_∞, we solve the equation $x(w) = 1$ and find $w = 2.312052651057$. Now with (25), we find $t = t(w) = 1.380558666227$. Hence, $f_1 = 4.32596753511667262691$ and $a_1 = 4.69258380040852416505$. ■

These results are in agreement with Table 1 to, in some cases, all decimal places.

From now on, we use the coordinates t, w in the domain B instead of f, a.

Theorem 4. *The domain B is the first quadrant of the plane $\{t, w\}$. The map $M\colon A \to B$ is bijective.*

Proof. First, we prove that the boundaries are attainable. Obviously, $N_{1/2} = \{w = 0\}$. Following Example 1, we find where the line $C = \text{const}$ is attached to the boundary $N_{1/2}$ of the domain B (see Fig. 2). We solve the first equation in (26) where $g = 1/2$ with respect to $k \in (0, 1)$. The solution, obviously, exists and unique. The value j on the boundary $L_{1/2}$ is found from the second equation in (26) or from (19). Then $t = 2\,t(k)$ by the formula (25).

The boundary $\{t = 0\} = \{a_1 = 0\}$ can be attained in this way. We put $x = 1$, $k = i$, and an arbitrary $w > 0$ in the equation (24), which we solve then for $j = j(w)$. Since $0 = k + 1/k = C^2/(2\,j)$ for $k = i$, then $C = 0$. Now $t = 0$ is found from the equation (25).

According to the paragraph after Theorem 1, the solution of the BVP is reduced to solution of a nonlinear equation. But we must know first which equations to use, i.e., where a chosen point in B lies. So we need to find out how the three curves $L_{1/2}$, L_1, and L_∞ are mapped into the domain B.

The curve $N_{1/2}$ is the abscissa of the plane $\{t, w\}$, $t \geq 0$.

By virtue of (25), and since $w = k$ on L_1, the curve N_1 is given by

$$N_1 = \left\{t = 2\,\mathrm{Im}\left(\sqrt{1+w^2}\,\Pi\left(i\,w, w\right)\right)\right\},\ w \in [0,1). \tag{28}$$

On L_∞, $j = C^2/4$, hence $k = 1$, and the integrals (24) and (25) are expressed in elementary functions. The equation $x(w) = 1$ can be written in the form

$$\sqrt{w} = \tanh\left(\sqrt{w} + \frac{C\sqrt{2}}{4}\right), \tag{29}$$

which has a unique solution $w \in [0,1)$ for every $C \geq 0$. In addition, the equation (29) can always be solved by simple iterations, since tanh() is a contracting mapping. Thus the curve N_∞ is given by

$$N_\infty = \left\{t = \sqrt{2}\,\mathrm{arctanh}\,\sqrt{w} - \mathrm{arctanh}\left(\frac{\sqrt{2\,w}}{1+w}\right)\right\},\ w \in [0,1). \tag{30}$$

So both curves N_1 and N_∞ have the horizontal asymptote $w = 1$.

The series (22) with $h = 2$ give the power expansion for t in (30). And since the power expansion for t in (28) is given by the series

$$t = \pi\left(\frac{1}{2}\,w + \frac{1}{8}\,w^3 + \frac{13}{128}\,w^5 + \dots\right),$$

it follows that the curve N_∞ is higher than N_1, i.e., $N_\infty > N_1$ on the plane $\{t, w\}$ for small w and hence everywhere, since the images of the curves L_g can not intersect due to Theorem 2. It should have been expected, since the curve L_1 lies between the curves $L_{1/2}$ and L_∞ in A.

Now we are equipped to solve the BVP regardless of where the point $p = (t, w)$ in B is located, and, incidentally, to prove that the map M is bijective.

We denote the preimage of the point p as q, i.e., $M^{-1}(p) = q \in A$, provided q exists. We recall that r and t are found uniquely by f_1 and a_1 and the formulas (7) if $a_1 \leq \sqrt{f_1\,(f_1+2)}$, i.e., if $p \in B$. By r, we find $w = \sqrt{r\,(r+2)}$. If $w < 1$, then we compute two values t_1 and t_∞ by the formulas (28) and (30) respectively.

Thus, there are the following cases (see Fig. 3):

(1) if $w \geq 1$, then $p > N_\infty$, i.e., $q > L_\infty$.

(2) if $w < 1$ and $t < t_\infty$, then still $p > N_\infty$, i.e., $q > L_\infty$.

(3) if $w < 1$ and $t_\infty < t < t_1$, then $N_\infty > p > N_1$, i.e., $L_\infty > q > L_1$.

(4) if $w < 1$ and $t_1 < t$, then $N_1 > p$, i.e., $L_1 > q > L_{1/2}$.

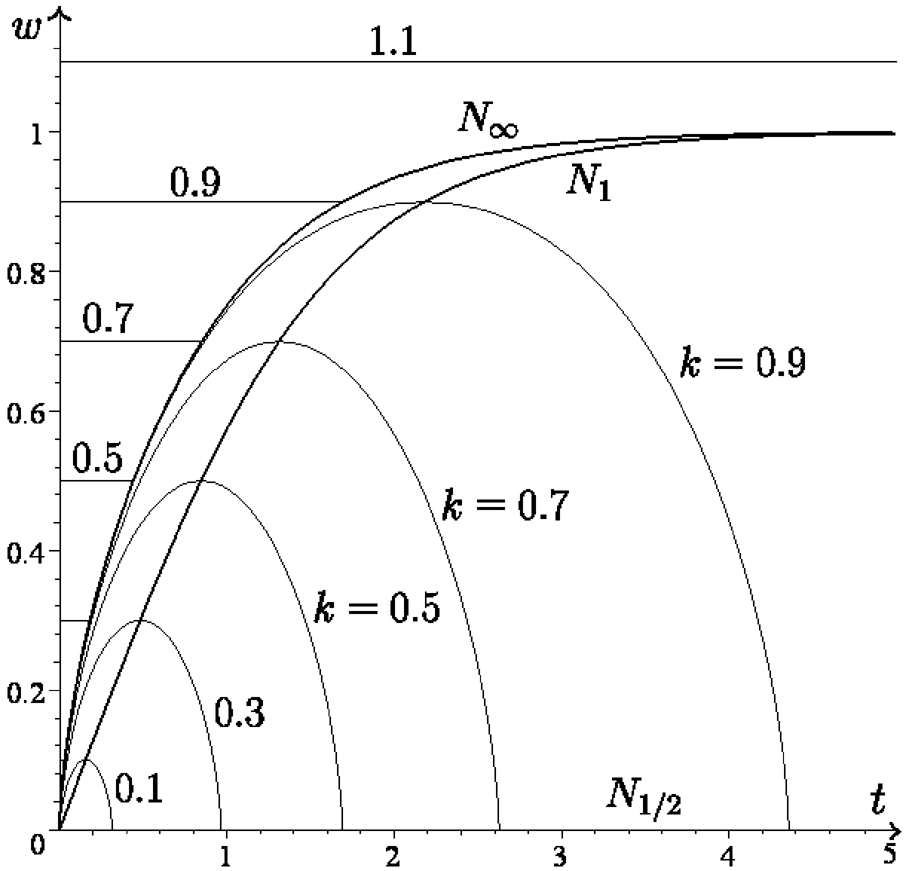

Fig 3. Global parametrization of the domain B in three sectors.

Both cases (1) and (2) fall into the sector of the plane $\{t, w\}$ above the curve N_∞ but treated slightly differently. Namely, in the case (1), t is an unbounded strictly increasing function of $s \in [-\pi/2, 0)$, where $k = \exp(-i\,s)$. This is better seen from the integral (18), which we need to differentiate with respect to s (we omit the details). Thus k is found uniquely. Then j is found by the formula (24), where $x(w) = 1$, and, finally, C is found by the formula (19).

In the case (2), $0 \leq t \leq t_\infty$, where $t = t_\infty$ is reached at $k = 1$, or $s = 0$. This is the only difference from the case (1).

In the case (3), k is real, $w \leq k \leq 1$. If we use $h = k + 1/k$ as the parameter in the integral (18), then $t(w, h)$ is a strictly increasing function of $h \in [2, w + 1/w]$, which is found as in the case (1). Thus $t(w, k)$ is a strictly decreasing function of $k \in [w, 1]$. In addition, $t(w, 1) = t_\infty$, and $t(w, w) = t_1$. Thus k, and hence C and j are found uniquely as in the case (1) and (2).

In the case (4), we need to switch the integrals (see Theorem 1). On the curve N_1, the maximal value of $w = k$ is attained, so the equation $t = 2\,t_1(k) - t(w, k)$ should be used for $k \in [w, 1)$. Here $t_1(k)$ is found by the formula (28) for $w = k$, and $t(w, k)$ is found by the formula (25). For $k = w$, $t = t_1$, i.e., $p \in N_1$. Since $t_1(k) \to +\infty$ as $k \to 1$, the solution for k always exists and unique. Then j

is found from the equation $1 = 2\,x(k) - x(w)$, where $x(w)$ is computed by the formula (24), and, finally, C is found by (19). $\blacksquare$

We do not need to prove that the map $M^{-1}\colon B \to A$ is a diffeomorphism, since it is already done in Theorem 2. But now we can compute the Jacobian of the map M^{-1} explicitly, although by a sector. We verified that, indeed, the Jacobian does not vanish in each sector on Fig. 3. As an example, we give the expression for the Jacobian in the sector between the curves N_1 and N_∞:

$$\left|\frac{\partial(C,j)}{\partial(w,k)}\right| = \left(\frac{2\,(1-k^2)\,\sqrt{w}}{k^2\,\sqrt{(1-k\,w)\,(k-w)\,(1+k^2)}}\right)\left(F\left(\sqrt{\frac{w}{k}},k\right) - E\left(\sqrt{\frac{w}{k}},k\right)\right)^2 .$$

The Jacobian in Theorem 2, i.e., $|\partial(t,w)/\partial(C,j)|$, is computed by the formula $|\partial(t,w)/\partial(C,j)| = |\partial(t,w)/\partial(w,k)|/|\partial(C,j)/\partial(w,k)|$. The latter formula is too big to cite it here, but we verified this result numerically computing Jacobians with the algorithm of solution of the IVP given in Section 3.

Now we can verify Table 2 with the algorithm of solution of the BVP that does not rely on numerical integration of ODEs, and that is implemented with arbitrary precision. In the first line of Table 2 (case(1)), the error is less than 1×10^{-14}, and all decimal places are correct for C; in the second line (case(1)), the error is less than 2×10^{-11}; and in the third line (case(4)), the error is less than 1.5×10^{-12}.

Concluding this paper, we discuss, as we promised, some numerical aspects of our computations. First, we need to stress that computations with extended precision are not only useful, but sometimes necessary. Fig. 3 makes this obvious, since for big t, and w close to 1, there is a chance that we miss the right sector, and hence take wrong integrals for solution of the BVP. For example, for $t = 5$ we find (see the paragraph before Example 1):

$$\begin{aligned} w_\infty &= 0.99902367207111949779, \\ w_1 &= 0.99804830064554260930. \end{aligned}$$

For greater t, ordinary 16 decimal places quickly become inadequate. Luckily, there are exellent open source utilities for computations with arbitrary precision (see [3, 4] and references there).

Unlike this, open source utilities for computing elliptic integrals with complex arguments are very rare. Maxima, for example, can compute only the integrals of the first and second kind, but not elliptic Π, which is necessary here. A special note for Maple users. It is rather slow, and sometimes gives only half of the promised digits. Otherwise, it is quite satisfactory. We found the paper [5] on the internet, where this subject is treated extensively, and there are some very useful references. We adapted algorithms that we found for Fortran and some CAS.

Finally, computation of elliptic integrals is an iterative fast convergent process that gives a guaranteed result for real arguments. For complex ones it may be not so. Since there is a lot of square roots involved in the process, these integrals are prone to switch some complex branches without prior notice. So some tuning might be in order.

References

1. *N. Ben Abdallah, P. Degond, F. Méhats.* Mathematical models of magnetic insulation // Physics of Plasmas, V. 5, N. 5, (1998), p. 1522-1534.

2. *I.S. Gradshteyn, I.M. Ryzhik.* Table of Integrals, Series, and Products. Seventh Edition. Academic Press, Elsevier, 2007.

3. *D.H. Bailey.* A fortran-90 based multiprecision system // ACM Transect. on Math. Software, V. 21(4), (1995), p. 379-387.

4. *D.H. Bailey, J.M. Borwein, N.J. Calkin, R. Girgensohn, D. Russell, L.H. Moll, V.H. Moll.* Experimental Mathematics In Action. Canada, A.K. Peters, 2006.

5. *B.C. Carlson.* Toward Symbolic Integration of Elliptic Integrals // J. Symbolic Computation, V. 28, (1999), p. 739-753.