involve

a journal of mathematics

msp

2010 vol. 3, no. 4

involve

pjm.math.berkeley.edu/involve

See inside back cover or http://pjm.math.berkeley.edu/involve for submission instructions.

The subscription price for 2010 is US $100/year for the electronic version, and $120/year (+$20 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW™ from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

http://msp.org/

A NON-PROFIT CORPORATION

Typeset in LaTeX

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.349

Identification of localized structure in a nonlinear damped harmonic oscillator using Hamilton's principle

Thomas Vogel and Ryan Rogers

(Communicated by Zuhair Nashed)

In the mid-seventeenth century Isaac Newton formalized the language necessary to describe the evolution of physical systems. Newton argued that the evolution of the state of a process can be described entirely in terms of the forces involved with the process. About a century and a half later, William Hamilton was able to establish the whole of Newtonian mechanics without ever using the concept of force. Rather, Hamilton argued that a physical system will evolve in such a way as to extremize the integral of the difference between the kinetic and potential energies. This paradigmatic reformulation allows for a type of reverse engineering of physical systems. This paper will use the Hamiltonian formulation of a nonlinear damped harmonic oscillator with third and fifth order nonlinearities to establish the existence of localized solutions of the governing model. These localized solutions are commonly known in mathematical physics as solitons. The data obtained from the variational method will be used to numerically integrate the equation of motion, and find the exact solution numerically.

1. Introduction

For those that are familiar with mathematical modeling, it is common knowledge that most physical systems can be modeled by considering the sum of the forces acting on the system. These various forces appear explicitly in what is commonly referred to as *the equation of motion* or *the governing system*. These individual terms can be interpreted as forces, accelerations, potentials, external damping, dispersive effects, etc. It is most often the case that this equation is a differential equation. This equation can then be solved using the seemingly endless supply of tricks and techniques that have been developed over recent centuries by mathematicians and others. This approach to model construction was largely formalized by Isaac Newton and his contemporaries in the mid-seventeenth century. Newton offered

MSC2000: primary 49S05; secondary 49M99, 34K35.
Keywords: variational, Hamilton's principle, solitons, embedded solitons.

a way to write down physical observations in a concise language which allowed for an accurate prediction of how the process would evolve in time. All one has to do is establish the net forces involved with the process, and solve the associated equation. This development, of course, ran concurrently with the advent of modern calculus. Newton's approach in describing physical processes by examining the net external forces reigned supreme in the collective conscious of scientists, engineers, and natural philosophers for over a century and a half.

William Hamilton, a physicist, later proposed in [1834; 1835] a completely different view of describing physical systems. While the Newtonian school held that a system evolves according to external influences, Hamilton argued that a physical system will evolve in such a way as to extremize certain mathematical quantities. Hamilton's formulation enabled him to establish the whole of Newtonian physics without ever using the concept of force. Instead, Hamilton correctly argued that a physical process will evolve in such a way as to extremize the integral of the difference between the kinetic and potential energies. Hamilton dubbed this quantity *the action*. With this, Hamilton offered humankind a paradigm shift in considering the evolution of physical processes.

Hamilton was able to develop his force-independent theory of physical systems thanks to the mathematics which had been developed over almost two centuries since Newton presented his force-driven theory of the universe. The mathematical roots of Hamilton's theory can be traced back to what is known as the *brachistochrone problem* [Nesbet 2003]. In 1696, Johann Bernoulli proposed this problem, which can be restated thus: *If two points are connected by a wire whose shape is given by an unknown function $y(x)$ in a vertical plane, what shape function minimizes the time of descent of a bead sliding without friction from the higher to the lower point*? This problem was addressed by several of Bernoulli's contemporaries, and what arose from these investigations was a new type of calculus, known today as the calculus of variations and developed into a full mathematical theory by Euler around 1744 (see [Rouse Ball 1901], for instance). The mathematics developed by Euler was extended by Joseph-Louis Lagrange (1736–1813), who discovered that Euler's equation for minimizing a functional integral (later to be named the Euler–Lagrange equation) could be expressed in a compact way by simply using integration by parts (see [O'Connor and Robertson 1999], for instance). It was Lagrange who introduced the integrand of the functional appropriate to mechanics, that is, the difference between potential and kinetic energies.

Around the time William Hamilton was developing his new style of physics, a man named John Scott Russell had made the first observation of a phenomena that would one day become relevant in constructing a global telecommunications network. As a civil engineer, Russell was quite active in the advancement of naval-vessel architecture. Russell revolutionized naval architecture by creating a new

system of hull construction, he was the first person to offer steam carriage service between Paisley and Glasgow in 1834, and he is also responsible for some of the first recorded experimental data of the Doppler effect of the sound frequency shift of passing trains [Eilbeck 2007]. One day in 1834, Russell was running an experiment in order to establish a conversion factor between steam power and horse power. As part of his experimental apparatus, he tethered horses to a boat. At some point in the experiment, things went wrong: the ropes binding the horses to the boat snapped. Russell curiously watched as a huge swell of water formed around the hull of the stalled boat. Suddenly this mound of water sprang forward and began propagating down the Union Canal. Russell and his trusty steed followed the traveling water crest until it dissipated in the standing water several kilometers away from the boat. What was so astonishing to Russell was the absolute lack of attenuation or dissipation of the water wave over such a long distance. Water wave dynamics, as understood in the nineteenth century, did not allow for such *waves of permanent form*, as Russell named them. (Today, mathematical solutions of this type of are known as solitons, a name arising arose from the concept of a solitary wave.) Much skepticism surrounded Russell's claim, and he dedicated his remaining days to recreating the phenomena he observed that fateful day along the Union Canal.

The disbelief of Russell's contemporaries lies with the formulation of the model of water wave equations at a relatively shallow depth. It was not until 1895 that two mathematicians, D. J. Korteweg and G. de Vries, successfully constructed a mathematical model which affirmed Russell's observation sixty years earlier. Korteweg and de Vries derived what is known today as the KdV equation:

$$v_t + vv_x + v_{xxx} = 0. \tag{1}$$

Correctly describing the behavior of shallow water waves [Debnath 1997], this is a nonlinear evolution equation (subscript x and t denote differentiation with respect to the space and time coordinates). Russell's contemporaries' disbelief was due to the fact that their models of the behavior of water wave dynamics were linear. Due to the complexity of solving nonlinear evolution equations, research in the area of solitons stalled until the 1960s. Its renaissance was, of course, due to the advent of the modern computer.

In the mid 1960s Martin David Kruskal ran the first numerical simulation of interacting solitons in the KdV equation. This early work contributed much to our current understanding of these rather exotic types of mathematical constructs. What Kruskal was able to demonstrate not only advanced applied mathematics, but forced mathematicians and physicists to reconsider the very notion of what is meant by interacting waves. Kruskal found that when two solitons interact, they will exhibit some level of interference much like any wave phenomena which

is observed in nature, though they do not linearly superimpose on one another. The difference is that upon interacting, the solitons will return to their original state. That is to say, once the interaction had taken place, the soliton reestablishes its original shape, velocity, and other governing physical characteristics with a possible phase shift being the only observable consequence of the interaction. Understanding these types of mathematical constructs has lead to some of the more profound advancements in the last few decades. Most notable of these advances are fiber optic and wireless communication over a global network. This paper will introduce a novel way to establish such localized structure (i.e., solitons) without the difficulties encountered by techniques which require working with the equation from the Newtonian perspective.

2. Hamilton's principle

As mentioned in the previous section, solitons do not exist in linear equations. They only occur in nonlinear differential equations. Throughout the last several decades many techniques have been developed in establishing solutions to nonlinear differential equations [Debnath 1997; Drazin and Johnson 1989]. These techniques are characterized by their limited reach in solving large classes of problems. They are also characterized by being rather complicated. The most famous of these techniques is what is known as the inverse scattering technique. During the late 1960s Kruskal continued his work with solitons and developed what amounts to an analogous form of a Fourier transform for nonlinear differential equations. This work was developed by Kruskal in conjunction with three other mathematicians: Clifford Gardner, John Greene, and Robert Miura [Gardner et al. 1967]. What came out of this work is what is known as the inverse scattering technique (IST). The early development of the IST had a shortcoming: it only applied to integrable equations. In 1972 four young mathematicians, Mark Ablowitz, David Kaup, Alan Newell, and Harvey Segur, established what is known as the AKNS theory [Ablowitz et al. 1974]. This was an extension of the IST that ultimately allowed for analysis of nonintegrable systems. While these techniques have shaped modern applied mathematics, they are complicated and require a great deal of specialized understanding to be used effectively [Drazin and Johnson 1989]. Additionally, IST considers solutions to the nonlinear evolution equation as it is postulated in the Newtonian sense via considering the net forces acting on the system. So let us begin by considering some nonlinear differential operator, Φ, for which there exists some v satisfying

$$\Phi[v] = 0. \tag{2}$$

This represents the Newtonian formulation of the physical system. Depending on the structure of Φ, solutions to (2) most likely cannot be found directly. In fact,

it is difficult to make any generalization about (2) without assuming some sort of additional structure on Φ. Instead of restricting Φ to a certain class of nonlinear differential operators, consider a paradigmatic reformulation of (2). Suppose this nonlinear operator is the derivative (in some sense) of some associated *energy functional* L, a fact we write as

$$\Phi[v] = \nabla(L[v]). \tag{3}$$

Equation (2) may now be written in terms of the energy functional:

$$\nabla(L[v]) = 0. \tag{4}$$

This establishes a duality in which solutions to (2) are seen as the critical points of the functional $L[\,\cdot\,]$. This is the heart of Hamilton's principle. This approach enabled Hamilton to describe the entirety of Newtonian mechanics without having to consider the evolution of a system in terms of external forces. In modern mathematics this energy functional is termed the *Lagrangian*, and its formulation depends on the physical system of interest.

Suppose the expression of the Lagrangian is known, and that it is a functional of the variable $v(t)$, which itself may be a scalar, vector or tensor quantity. In the present work, we shall only consider a one-dimensional scalar case, where the integration variable is t. If we let $\mathcal{D}$ be the domain of support of the function v, the action, $S[v]$, is defined by

$$S = \int_{\mathcal{D}} L[v]\,dt. \tag{5}$$

Hamilton's principle states that the evolution of a dynamical system between two specific states is an extremum of the action functional given by (5). More formally, Hamilton's principle states that the solution to a given dynamical system $v(t)$ is determined by (6) for any bounded variation $\delta v(t)$, provided that this variation vanishes at any and all end points of the domain $\mathcal{D}$ [Kaup and Vogel 2007]. Note that this also defines the quantity $(\delta L/\delta v)(t)$, which is called the (first) variational derivative of L:

$$\lim_{\epsilon\to 0} \frac{S[v(t)+\epsilon\delta v(t)] - S[v(t)]}{\epsilon} = \int_{\mathcal{D}} \frac{\delta L}{\delta v}[v(t)]\,\delta v(t)\,dt = 0. \tag{6}$$

In terms of the nonlinear differential operator Φ, this establishes a connection between the governing equation of motion and the first variational derivative of the Lagrangian:

$$\Phi[\,\cdot\,] = \frac{\delta L[\,\cdot\,]}{\delta v}. \tag{7}$$

This paradigm shift offered by Hamilton allows for a rather novel approach to approximating solutions of evolution equations for which a Lagrangian can be

established. Suppose the physical characteristics (geometric or otherwise) of a particular type of solution to the equation of motion given by (2) are known. For instance, an ordinary soliton could be described in terms of a traveling "lump" having some associated amplitude and width. Of course, depending on the governing system, the solution could have other identifying characteristics such as position, velocity, chirp, phase, etc. An *ansatz*, or tentative functional form, can then be constructed in terms of parameters representing those physical characteristics. Let $v_0(t; q_i)$ be the ansatz, where the q_i from a finite collection of parameters representing the physical characteristics in question, and on which v_0 is dependent; these parameters could also depend on other independent variables, such as t. With the functional form of v_0 fixed, we can vary the q_i, and this variation gives a set of equations expressing the extremum principle:

$$\frac{\partial S}{\partial q_i} = \frac{\partial}{\partial q_i} \int_{\mathcal{D}} L[v_0] dt = 0. \tag{8}$$

(This notation presumes the structure of the q_i is constant. If the parameters are assumed to be dependent on time, the partial derivative would become a functional derivative.) Once this is done, we have the q_i determined in the sense that we have the (algebraic or differential) equations whose solutions represent a best fit for the parameter values according to Hamilton's principle. The nonlinear differential equation considered for analysis in this paper is

$$v'' + \kappa v' + \varphi v + v^3 + \omega v^5 = 0, \tag{9}$$

where $\kappa, \varphi, \omega \in \mathbb{R}$ and $v(\xi) : \mathbb{R} \to \mathbb{R}$. The prime indicates the differentiation with respect to the independent variable ξ.

3. Ordinary solitons

This paper will establish the existence of two different types of solitons for (9). The first type is known as *ordinary solitons*: localized solutions that occur in a region of the extrinsic parameter space, in this case $(\kappa, \varphi, \omega)$-space, for which the linear eigenmodes are exponential. The most frequent type of localized structure identified in nonlinear evolution equations are those solutions for which $v(\xi) \to 0$ as $\xi \to \pm\infty$. Ordinary solitons for which $v(\xi) \to 0$ as $\xi \to \pm\infty$ are referred to as *bright solitons*, while those with the asymptotic behavior $v(\xi) \to c$ where $c \in \mathbb{R}$ as $\xi \to \pm\infty$ are called *dark*. We will only consider bright solitons in the current work. Requiring a vanishing amplitude for very large values of ξ means that the eigenvalues of the linearized problem must remain real (otherwise, there would oscillatory behavior in the eigenmodes). The eigenvalues of the linearization of

(9) are given by

$$\lambda_{\pm} = \frac{-\kappa \pm \sqrt{\kappa^2 - 4\varphi}}{2}. \tag{10}$$

In order to keep them real-valued, it will be necessary to impose on the extrinsic parameter space the condition that $\kappa^2 - 4\varphi > 0$.

To begin the process of using Hamilton's principle to identify ordinary solitons, it is necessary to have the Lagrangian associated with (9) and an ansatz representing the geometry of the desired solution. A combination of inspection and trial and error shows that the Lagrangian from which (9) arises is given by

$$L(\xi, v, v') = \frac{e^{\kappa\xi}}{2}\left(\varphi v^2 - (v')^2 + \frac{v^4}{2} + \frac{\omega v^6}{3}\right); \tag{11}$$

this is the L that which recovers the equation of motion (9) under the associated Euler–Lagrange equation $L_v - (d/d\xi)L_{v'} = 0$. The ansatz for the soliton will be taken as

$$v_0(\xi; a, \rho) = a \exp\left(-\frac{\xi^2}{\rho^2}\right), \tag{12}$$

that is, a Gaussian function of amplitude a and core width ρ. There are two good reasons for choosing a trial function such as this: (i) it offers a relatively good geometric description of an ordinary soliton; and (ii) the Lagrangian evaluated at the ansatz is easy to integrate over $\mathbb{R}$. While other functional forms such $\operatorname{sech}^2(\xi)$ have similar geometric properties, it may become quite difficult to calculate the action. It could become necessary, for instance, to lift the integration into the complex plane to calculate the associated action.

Calculating the action as defined in (5), where the function $v(\xi)$ is evaluated at the ansatz ($v = v_0$) gives rise to the action

$$\begin{aligned} &S(a, \rho) \\ &= -\frac{a^2 e^{\frac{\rho^2\kappa^2}{24}}\sqrt{\pi}}{144\rho}\left(-18a^2 e^{\frac{\rho^2\kappa^2}{48}}\rho^2 + 9\sqrt{2}e^{\frac{\rho^2\kappa^2}{12}}(4 + \rho^2(\kappa^2 - 4\varphi)) - 4\sqrt{6}a^4\rho^2\omega\right). \end{aligned} \tag{13}$$

As discussed in Section 2, Hamilton's principle states that solutions of (9) will evolve in such a way as to extremize the action. While Hamilton's principle cannot be satisfied in general by using the trial function, it is possible to establish what values of variational parameters bring the ansatz *closest* to the exact solution. In general these parameters could vary with respect to some independent variable (such as time). If this were the case, the action would be varied with respect to a and ρ by way of the functional derivative (i.e., $S_{q_i} - (d/d\xi)S_{q_i'}$, where the q_i represents the variational parameters). Since the current analysis presumes the structure of the variational parameters to be constant, varying the action amounts to taking the partial derivative with respect to the variational parameters.

Varying the action with respect to a and ρ gives, respectively,

$$S_a = -\frac{a e^{\frac{\rho^2\kappa^2}{24}}\sqrt{\pi}}{24\rho}\left(-12a^2 e^{\frac{\rho^2\kappa^2}{48}}\rho^2 + 3\sqrt{2}e^{\frac{\rho^2\kappa^2}{12}}(4+\rho^2(\kappa^2-4\varphi)) - 4\sqrt{6}a^4\rho^2\omega\right), \quad (14)$$

$$S_\rho = -\frac{a^2 e^{\frac{\rho^2\kappa^2}{24}}\sqrt{\pi}\left(+27\sqrt{2}e^{\frac{\rho^2\kappa^2}{12}}(-16+8\rho^2(\kappa^2-2\varphi)+\rho^4(\kappa^4-4\kappa^2\varphi))\right)}{1728\rho^2}$$
$$-\frac{a^2 e^{\frac{\rho^2\kappa^2}{24}}\sqrt{\pi}\left(-27a^2 e^{\frac{\rho^2\kappa^2}{48}}\rho^2(8+\rho^2\kappa^2) - 4\sqrt{6}a^4\rho^2(12+\rho^2\kappa^2)\omega\right)}{1728\rho^2}. \quad (15)$$

The variational solution space is five-dimensional — there are three extrinsic parameters κ, ω, and φ, plus two variational parameters ρ and a. We are interested in the points $(\kappa, \varphi, \omega; a, \rho)$ that represent best parameter fits for the ansatz (that is, satisfy $S_\rho = 0$ and $S_a = 0$) and also satisfy the condition $\kappa^2 - 4\varphi > 0$, which we established by considerations from the linear spectrum. It is clear from the expressions (14) and (15) for S_ρ and S_a that solutions must be obtained numerically.

Since there are too many degrees of freedom, we fix two of the parameters on any given run — we chose the variational amplitude a and the linear damping κ — and take a third parameter to be an independent variable. We chose for this role the core width, ρ, which is always positive. Thus we step through values of ρ and search numerically for values of φ and ω satisfying the Euler–Lagrange equations, i.e., making (14) and (15) vanish. We discard solutions that do not satisfy the condition $\kappa^2 - 4\varphi > 0$. Some results of this process can be seen in Figure 1. As expected, the variational method indicates the persistence of many solution curves satisfying the Euler–Lagrange equations. It is often the case that ordinary solitons in a nonlinear evolution equation occur in infinite families. The geometric characteristics of ordinary solitons (such as the amplitude) very often depend continuously on

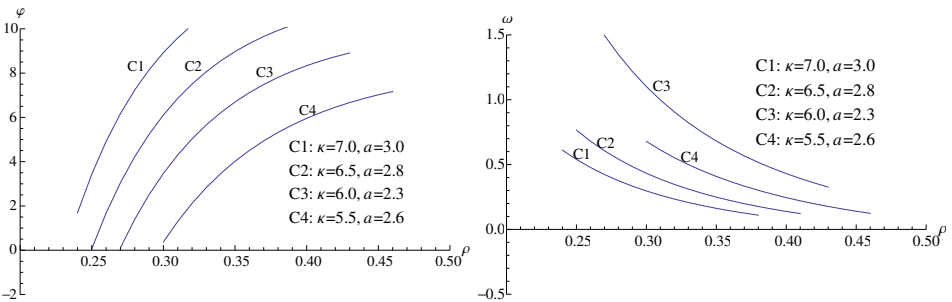


Figure 1. Projections in the (φ, ρ)- and (ω, ρ)-planes of some curves in parameter space satisfying $S_a = 0$ and $S_\rho = 0$. A point in parameter space is given by $(\kappa, \varphi, \omega; a, \rho)$; given a starting point, the numerical integration of (9) yields an "exact" solution.

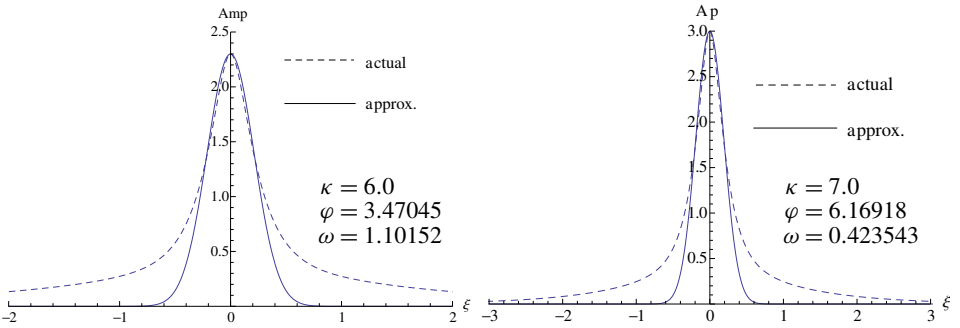


Figure 2. Ordinary solitons with the given extrinsic parameter values.

some extrinsic parameter in the governing model (for instance, the wave speed). Thus it is not surprising to find continuous curves in parameter space for which the variational method picks up localized structure.

With an abundance of data indicating the existence of localized structure, it is now time to numerically integrate (9) using this data, and find the exact (numerical) solutions. Integrating (9) requires two initial conditions: $v(0) = a$, where a is obtained from the variational data, and $v_\xi(0) = 0$ (from symmetry considerations). Figure 2 illustrates the result of this process for some selected solutions. Overlaid with the results of numerically integrating the model (9) is the ansatz evaluated at the variational solution data.

4. Embedded solitons

Embedded solitons get their name from the peculiar place in which they reside in the linear spectrum. Recall that for the case of ordinary solitons, asymptotic considerations lead us to require (in a very natural way) that the linear eigenmodes remain exponential. As it turns out, localized structure can exist in the spectrum of radiation modes. This breed of soliton is one of the relatively new types discovered in the last decade and a half [Yang et al. 1999; Champneys et al. 2001; Kaup and Malomed 2003]. It is possible that there are discrete values in the parameter space in which the nonlinearity is capable of "switching off" the radiation present in the linear eigenmodes. For this reason, these types of solitons became known as embedded solitons since they exist for parameter values embedded in the spectrum of radiation modes. A comprehensive explanation for the existence of such anomalous solutions is offered in [Champneys et al. 2001]. Thus we will restrict our attention to the region of parameter space where the linear eigenvalues determined by (10) are complex-valued. Once again, this solution will be established by way of considering a variational approximation. The ansatz is modified to allow for an additive radiation term:

$$v_0(\xi; a, \rho, \alpha) = a \exp\left(-\frac{\xi^2}{\rho^2}\right) + \alpha \cos(\psi\xi). \tag{16}$$

This variational trial function has been used in identifying embedded solitons in other systems [Kaup and Malomed 2003]. The $\cos(\psi\xi)$ structure is intentionally chosen to adhere to the symmetry of the core of the variational ansatz. The parameter ψ appearing in the phase of the radiation is *not* an additional parameter; it is a constant that will be determined in terms of variational and extrinsic parameters.

As indicated in Section 2, the variational trial function is then used to establish the action. This becomes quite tricky with an ansatz of the form (16). In general, this action integral will not converge. Upon inserting (16) into (11), some of terms can be integrated over all space, while others cannot. Here is the trick. To begin, the $\exp(\kappa\xi)$ factor arising in the Lagrangian from the damping is combined with the Gaussian structures by completing the square (the particulars of the substitution will vary from term to term). Then an effort is made, using various trigonometric identities, to isolate terms which are pure radiation. Such terms don't converge in the strict sense, but if the action is considered to be an *averaged* integral, the radiation can be considered to have a net zero contribution over all space. This approach was established in [Kaup and Malomed 2003]. Throughout the process of applying trig identities, terms which are not pure radiation (and are divergent) are generated. This is where the extra degree of freedom, ψ, in the phase of the radiation comes into play: ψ is established in such a way as to cause the remaining divergent terms to vanish. For this particular situation, ψ was calculated to be

$$\psi = \sqrt{\varphi + \tfrac{3}{8}\alpha^2 + \tfrac{1}{3}\alpha^4\omega}. \tag{17}$$

Upon taking the variational ansatz determined by (16) with a phase adjustment given by (17) and averaging out radiation terms, the effective action is found to be

$$\begin{aligned}
S(a,\rho,\alpha) &= \frac{a\sqrt{\pi}}{1440\rho} \\
&\times\Big(180a^3e^{\rho^2\kappa^2/16}\rho^2 - 90\sqrt{2}ae^{\rho^2\kappa^2/8}(4+\rho^2(\kappa^2-4\varphi)) + 40\sqrt{6}a^5e^{\rho^2\kappa^2/24}\rho^2\omega \\
&\quad + 288\sqrt{5}a^4e^{\rho^2(\kappa^2-\psi^2)/20}\rho^2\alpha\omega\cos P/10 + 480\sqrt{3}a^2e^{\rho^2(\kappa^2-\psi^2)/12}\rho^2\alpha\cos P/6 \\
&\quad + 900a^3e^{\rho^2(\kappa^2-4\psi^2)/16}\rho^2\alpha^2\omega\big(e^{\rho^2\psi^2/4} + \cos P/4\big) \\
&\quad + 1440e^{\rho^2(\kappa^2-\psi^2)/4}\rho^2\alpha\varphi\cos P/2 + \sqrt{2}ae^{\rho^2(\kappa^2-4\psi^2)/8}\rho^2\alpha^2\big(e^{\rho^2\psi^2/2} + \cos P/2\big) \\
&\quad + 400\sqrt{3}a^2e^{\rho^2(\kappa^2-9\psi^2)/12}\rho^2\alpha^3\omega\big(3e^{2\rho^2\psi^2/3}\cos P/6 + \cos P/2\big) \\
&\quad + 225\sqrt{2}ae^{\rho^2(\kappa^2-16\psi^2)/8}\rho^2\alpha^4\omega\big(3e^{2\rho^2\psi^2} + 4e^{3\rho^2\psi^2/2}\cos P/2 + \cos P\big) \\
&\quad + 360e^{\rho^2(\kappa^2-9\psi^2)/4}\rho^2\alpha^3\big(3e^{2\rho^2\psi^2}\cos P/2 + \cos 3P/2\big)\big) \\
&\quad + 90e^{\rho^2(\kappa^2-25\psi^2)/4}\rho^2\alpha^5\omega\big(10e^{6\rho^2\psi^2}\cos P/2 + 5e^{4\rho^2\psi^2}\cos 3P/2 + \cos 5P/2\big) \\
&\quad - 1440e^{\rho^2(\kappa^2-\psi^2)/}4\rho^2\alpha\psi\big(\psi\cos P/2 + \kappa\sin P/2\big)\Big),
\end{aligned}$$

where we have introduced the shorthand $P = \rho^2\kappa\psi$.

The action is then varied with respect to the three variational parameters a, ρ, and α. Since the embedded soliton itself has no radiation present (it is a purely localized solution) α will be taken to be zero *after* varying the action with respect to each parameter. This approach is discussed in [Kaup and Malomed 2003]. This results in the following three associated Euler–Lagrange equations:

$$(S_a)_{\alpha=0} = -24\sqrt{2}ae^{\rho^2\kappa^2/8} + 24a^3e^{\rho^2\kappa^2/16}\rho^2 - 6\sqrt{2}ae^{\rho^2\kappa^2/8}\rho^2\kappa^2 + 24\sqrt{2}ae^{\rho^2\kappa^2/8}\rho^2\varphi + 8\sqrt{6}a^5e^{\rho^2\kappa^2/24}\rho^2\omega, \quad (18)$$

$$(S_\rho)_{\alpha=0} = -27a^2e^{\rho^2\kappa^2/48}\rho^2(8+\rho^2\kappa^2) + 27\sqrt{2}e^{\rho^2\kappa^2/12}\big(-16+8\rho^2(\kappa^2-2\varphi)+\rho^4(\kappa^4-4\kappa^2\varphi)\big) - 4\sqrt{6}a^4\rho^2(12+\rho^2\kappa^2)\omega, \quad (19)$$

$$(S_\alpha)_{\alpha=0} = 3\sqrt{5}a^4\omega\cos\big(\tfrac{1}{10}\rho^2\kappa\sqrt{\varphi}\big) + 5\sqrt{3}a^2e^{\rho^2(\kappa^2-\varphi)/30}\cos\big(\tfrac{1}{6}\rho^2\kappa\sqrt{\varphi}\big) - 15e^{\rho^2(\kappa^2-\varphi)/5}\kappa\sqrt{\varphi}\sin\big(\tfrac{1}{2}\rho^2\kappa\sqrt{\varphi}\big). \quad (20)$$

These equations are then solved numerically in a similar fashion to the Euler–Lagrange equations in Section 3. Though this time there are three algebraic constraints. Possible solutions to these equations are also vetted according to the condition $\kappa^2 - 4\varphi < 0$ to ensure there is radiation present in the linear eigenmodes. Unlike the variational solution data obtained in Section 3, embedded solitons do not normally occur as continuous curves in parameter space. Rather, embedded solitons usually exist for a discrete value in the parameter space. We found one such solution, shown in Figures 3 and 4.

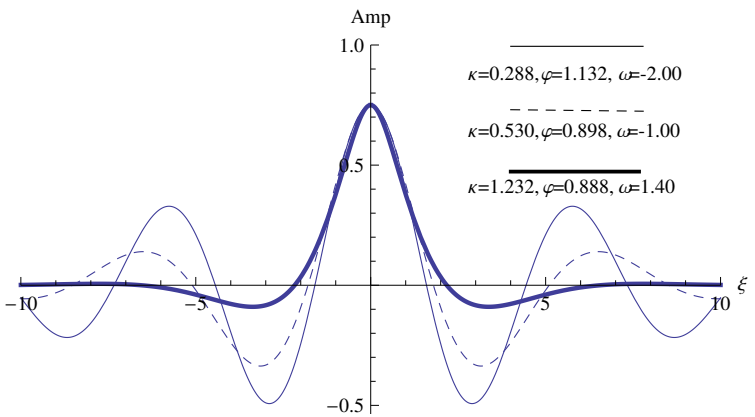


Figure 3. Delocalized solitons: results of numerically integrating (9) with (κ, ϕ, ω) values near those of the embedded soliton. The radiation dissipates as the parameter values approach the limit.

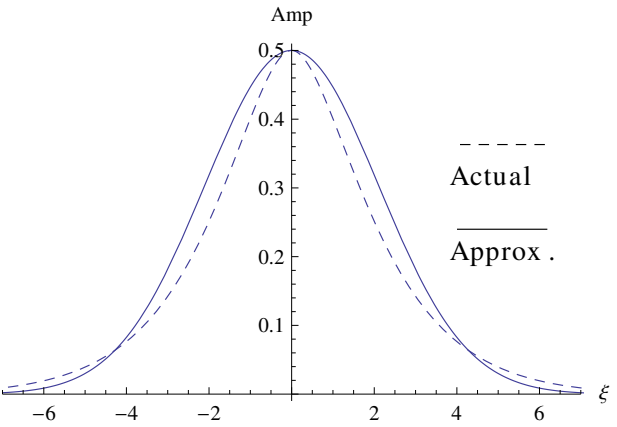


Figure 4. Exact solution (obtained via numerical integration) versus the variational trial function for extrinsic parameter values $\kappa = 1.1595$, $\varphi = 0.366215$, and $\omega = -0.101605$.

This figure also offers an overlay of the variational ansatz evaluated at the parameter values obtained by solving (18)–(20). Also included in Figure 3 is the result of integrating the equation of motion (9) *near* values for which the embedded soliton exists. These are commonly referred to as delocalized solitons. The closer the parameter values get to that of the embedded soliton the closer the amplitude of the radiation gets to zero.

5. Conclusions

The results obtained in this paper demonstrate the effectiveness and relative accuracy of using Hamilton's principle to establish localized structure in nonlinear evolution equations. These techniques are not limited to nonlinear equations. They can be implemented on just about any type of differential equation whether it is nonlinear or linear, partial or ordinary. As long as the Lagrangian can be established and there is some general understanding of the geometric characteristics of the desired solution, a variational method can be implemented. It has been shown [Kaup and Vogel 2007] that the variational method can fail to give reasonably accurate results in situations such as tracking soliton versus soliton interactions in a governing system. The approach outlined in this paper has the advantage of being able to establish solutions with relative ease when compared to some of the more complicated approaches available (e.g., inverse scattering techniques, calculating homoclinic orbits in phase space, etc.). In fact this methodology is accessible enough that advanced undergraduate students (such as the coauthor of this paper) with a good deal of mathematical maturity can use it in their own research projects.

References

[Ablowitz et al. 1974] M. J. Ablowitz, D. J. Kaup, A. C. Newell, and H. Segur, "The inverse scattering transform-Fourier analysis for nonlinear problems", *Stud. Appl. Math.* **53**:4 (1974), 249–315. MR 0450815 56 #9108 Zbl 0408.35068

[Champneys et al. 2001] A. R. Champneys, B. A. Malomed, J. Yang, and D. J. Kaup, "Embedded solitons: solitary waves in resonance with the linear spectrum", *Phys. D* **152/153** (2001), 340–354. MR 2002e:35198 Zbl 0976.35087

[Debnath 1997] L. Debnath, *Nonlinear partial differential equations for scientists and engineers*, Birkhäuser, Boston, 1997. MR 98f:35001 Zbl 0892.35001

[Drazin and Johnson 1989] P. G. Drazin and R. S. Johnson, *Solitons: an introduction*, Cambridge University Press, Cambridge, 1989. MR 90j:35166 Zbl 0661.35001

[Eilbeck 2007] C. Eilbeck, "John Scott Russell and the solitary wave", web page, Heriot-Watt University, 2007, available at http://www.ma.hw.ac.uk/~chris/scott_russell.html.

[Gardner et al. 1967] C. S. Gardner, J. D. Greene, M. D. Kruskal, and R. M. Miura, "Method for solving the Korteweg–de Vries equation", *Phys. Rev. Lett.* **19**:19 (1967), 1095–1097. Zbl 1103.35360

[Hamilton 1834] W. R. Hamilton, "On a general method in dynamics", *Phil. Trans. Royal Soc. London II* (1834), 247–308. Reprinted in his *Mathematical papers, II: dynamics*, edited by A. W. Conway and A. J. McConnell, Cambridge University Press, 1940.

[Hamilton 1835] W. R. Hamilton, "Second essay on a general method in dynamics", *Phil. Trans. Royal Soc. London I* (1835). Reprinted in his *Mathematical papers, II: dynamics*, edited by A. W. Conway and A. J. McConnell, Cambridge University Press, 1940.

[Kaup and Malomed 2003] D. J. Kaup and B. A. Malomed, "Embedded solitons in Lagrangian and semi-Lagrangian systems", *Phys. D* **184**:1-4 (2003), 153–161. MR 2004k:35311 Zbl 1037.35065

[Kaup and Vogel 2007] D. J. Kaup and T. K. Vogel, "Quantitative measurement of variational approximations", *Phys. Lett. A* **362** (2007), 289–297.

[Nesbet 2003] R. K. Nesbet, *Variational principles and methods in theoretical physics and chemistry*, Cambridge University Press, New York, 2003.

[O'Connor and Robertson 1999] J. J. O'Connor and E. F. Robertson, "Joseph-Louis Lagrange", page at the MacTutor web site, 1999, available at http://www-history.mcs.st-andrews.ac.uk/Biographies/Lagrange.html.

[Rouse Ball 1901] W. W. Rouse Ball, *A short account of the history of mathematics*, Macmillan, New York, 1901.

[Yang et al. 1999] J. Yang, B. A. Malomed, and D. J. Kaup, "Embedded solitons in second-harmonic-generating systems", *Phys. Rev. Lett.* **83**:10 (1999), 1958–1961.

Received: 2009-10-01 Revised: Accepted: 2010-11-12

tvogel@stetson.edu *Department of Mathematics and Computer Science, Stetson University, 421 N. Woodland Blvd., Unit 8332, DeLand, FL 32723, United States*

rrogers@stetson.edu *Department of Mathematics and Computer Science, Stetson University, 421 N. Woodland Blvd., Unit 8332, DeLand, FL 32723, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.363

Chaos and equicontinuity

Scott Larson

(Communicated by Zuhair Nashed)

Chaos theory examines the iterates of continuous functions to draw conclusions about long-term behavior. As this relatively new theory has evolved, one difficulty still present is the absence of universally agreed upon definitions. On the other hand, function spaces and equicontinuity are well established concepts with mathematical definitions that are universally accepted. We will present some theorems that display the natural connections between chaos and equicontinuity.

1. Introduction

Modern dynamical systems theory began in 1890, when Henri Poincaré was studying the three-body problem. He discovered the existence of aperiodic points that approach neither infinity nor a fixed point. Although this chaotic behavior was observed in 1890, it was not until around 1960 that chaos was formally studied. The invention of the electronic computer made studying chaos possible, by allowing one to iterate a simple function many times.

It can be said that chaos occurs when a deterministic system appears to be random. Edward Lorenz observed this phenomenon in 1961, when small round off errors led to unexpected results. The sensitive dependence on initial conditions caused his deterministic system to appear random. This can be examined using chaos theory.

There is no universal agreement on what the precise definition of a chaotic system should be. Many authors use different definitions to describe similar concepts [Devaney 1986; Li and Yorke 1975; Martelli 1999; Robinson 1999]. It is easy to see how this could be a problem and it would be useful to unify the terminology. A good starting point would be relating chaos to the classic idea of equicontinuity.

Well over one hundred years ago, equicontinuous families of functions were introduced in [Arzelà 1895; Ascoli 1884]. Equicontinuity allowed Giulio Ascoli and Cesare Arzelà to understand the behavior of families of continuous functions.

MSC2000: 37D45, 54C05, 54C35.
Keywords: equicontinuity, chaos.

Since chaos describes the behavior of a family of iterates of a continuous function, it seems natural to examine chaotic systems in terms of equicontinuity.

2. Equicontinuity

Ascoli and Arzelá used equicontinuity to explain which families of continuous functions will be "well behaved". Their results now have many important applications throughout mathematics. We will first introduce their classic definition.

A function space is defined to be the set of functions from a space X into a space Y, denoted by Y^X. For our purposes we restrict our function space to the set of continuous functions from X to Y, denoted by $\mathscr{C}(X, Y)$.

Definition 2.1. Let (Y, d) be a metric space. Let $\mathscr{F}$ be a subset of the function space $\mathscr{C}(X, Y)$. If $x_0 \in X$, the set $\mathscr{F}$ of functions is said to be *equicontinuous at* x_0 if given $\epsilon > 0$, there is a neighborhood U of x_0 such that for all $x \in U$ and all $f \in \mathscr{F}$,

$$d(f(x), f(x_0)) < \epsilon.$$

If the set $\mathscr{F}$ is equicontinuous at x_0 for each $x_0 \in X$, it is said simply to be *equicontinuous*.

Many times in topology, it is important to know when a set will be compact. Informally speaking, if a family of continuous functions is compact, then it will be well behaved. More precisely, if a sequence from a compact family of continuous functions converges in the supremum metric, then it must converge to a continuous function. A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. As Ascoli and Arzelá determined, equicontinuity is the additional property needed to assure that a closed and bounded subset of a family of continuous functions will be compact.

Theorem 2.2 (Arzelà–Ascoli theorem). *Let X be a compact space, and consider $\mathscr{C}(X, \mathbb{R}^n)$ in the* sup *metric. A subset $\mathscr{F}$ of $\mathscr{C}(X, \mathbb{R}^n)$ is compact if and only if it is closed, bounded, and equicontinuous.*

3. Chaos

Chaos theory describes behavior that is the diametric opposite of well behaved. The first time the word *chaos* was used to describe this mathematical phenomenon was in [Li and Yorke 1975], where the authors described when the iterates of a continuous function on an interval of the real line exhibit chaotic behavior. The definition of a chaotic function in that paper does not conveniently generalize, so most subsequent authors have chosen to use a different definition of chaos.

One commonly cited definition of a chaotic function, given in Robert Devaney's book [1986], involves two important characteristics: topological transitivity and

sensitive dependence on initial conditions. But as for chaos itself, multiple definitions exist for these two ideas (see [Devaney 1986; Martelli 1999; Robinson 1999; 2004]). We will adopt those used in [Robinson 2004].

Definition 3.1. $f: X \to X$ is said to be *topologically transitive* if there exists $x \in X$ such that $\{f^n(x) \mid n \in \mathbb{Z}^+\}$ is dense in X.

Definition 3.2. Let f be a map on a metric space X. The map has *sensitive dependence on initial conditions at* x_0, provided that there exists $\epsilon > 0$ such that, for any $\delta > 0$, there exists a y_0 such that $d(x_0, y_0) < \delta$ and a $n > 0$ such that

$$d(f^n(x_0), f^n(y_0)) \geq \epsilon.$$

The map has *sensitive dependence on initial conditions on a set* A, provided that it has sensitive dependence on initial conditions at every points $x_0 \in A$.

These are the two characteristic properties of a chaotic function, according to [Robinson 1999].

Definition 3.3. A subset $S \subseteq X$ is said to be *invariant* under f provided $f(S) = S$.

Definition 3.4. A map f on a metric space X is said to be *chaotic on an invariant set* Y provided (i) f restricted to Y is topologically transitive and (ii) f restricted to Y has sensitive dependence on initial conditions.

This idea of sensitive dependence on initial conditions appears to be related to an equicontinuous family of functions. The following section will explore the connections between chaos and equicontinuity.

4. Connections

If a family of iterates of a continuous function is equicontinuous at a point, then all iterates of nearby points will remain close together. This idea appears contrary to sensitive dependence on initial conditions, so the following theorem is natural.

Theorem 4.1 [Henry and Trapp 1998]. *Let X be a metric space, $f: X \to X$ be a continuous function, $x \in X$, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Then f has sensitive dependence on initial conditions at x, if and only if $\mathcal{F}$ is not equicontinuous at x.*

Proof. Suppose that $\mathcal{F}$ is not equicontinuous at $x \in X$. Then there exists an $\epsilon > 0$ such that, for any $\delta > 0$, there exists a y such that $d(x, y) < \delta$ and an $n > 0$ such that $d(f^n(x), f^n(y)) \geq \epsilon$. Therefore, $\mathcal{F}$ being equicontinuous at x is the negation of f having sensitive dependence on initial conditions at x. □

It becomes clear that equicontinuity is closely related to chaos. There are additional natural connections, so we continue to show the various forms in which equicontinuity appears. One definition that appears throughout texts on chaos is stability.

Definition 4.2. A point p is *Lyapunov stable* provided given any $\epsilon > 0$ there is a $\delta > 0$ such that if $|x - p| < \delta$ then $\left|f^j(x) - f^j(p)\right| < \epsilon$, for all $j \geq 0$.

Theorem 4.3. *Let X be a metric space, $f : X \to X$ be a continuous function, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Then $\mathcal{F}$ is equicontinuous at $x \in X$ if and only if x is Lyapunov stable.*

Proof. First suppose that $\mathcal{F}$ is equicontinuous at $x \in X$. Then for all $\epsilon > 0$ there exists $\delta > 0$ such that $d(x, y) < \delta$ implies that

$$d(f^n(x), f^n(y)) < \epsilon \quad \text{for all } n \in \mathbb{N}.$$

Without loss of generality, we may assume $\delta \leq \epsilon$. Thus, $d(x, y) < \epsilon$. Therefore,

$$d(f^n(x), f^n(y)) < \epsilon \quad \text{for all } n \geq 0.$$

So f is Lyapunov stable at x. Now if f is Lyapunov stable at x, then $\mathcal{F}$ is clearly equicontinuous at x. □

Definition 4.4 (Stable fixed point [Henry and Trapp 1998]). Let p be a fixed point of f. We call p a stable fixed point provided there is a neighborhood U of p such that

$$\lim_{n\to\infty} \operatorname{diam}(f^n(U)) = 0.$$

Definition 4.5. A point p is called *periodic* if $f^n(p) = p$ for some $n \in \mathbb{N}$. The smallest such n is the *period* of p.

Two theorems relating this definition to equicontinuity are proved in [Henry and Trapp 1998]. We say that p is a stable periodic point if and only if p is a stable fixed point of f^n for some $n \in \mathbb{N}$.

Theorem 4.6 [Henry and Trapp 1998]. *Let p be a fixed point of f. If p is stable then the iterates of f are equicontinuous at p.*

Theorem 4.7 [Henry and Trapp 1998]. *If p is a stable periodic point, then f has equicontinuous iterates at p.*

Another definition that appears extensively in chaos theory is that of an ω-limit set.

Definition 4.8. A point y is an *ω-limit point of x for f* provided there exists a sequence of n_k going to infinity as k goes to infinity such that

$$\lim_{k\to\infty} d(f^{n_k}(x), y) = 0.$$

The set of all ω-limit points of x for f is called the *ω-limit set of x* and is denoted by $\omega(x)$ or $\omega(x, f)$.

A useful characterization of an ω-limit set is given by the following theorem.

Theorem 4.9 [Robinson 1999]. *Let $f: X \to X$ be a continuous function on a metric space X. Then for any $x \in X$, $\omega(x) = \bigcap_{N\geq 0} \mathrm{cl}(\bigcup_{n\geq N}\{f^n(x)\})$.*

Notice that if $\omega(x) = X$, then $\{f^n(x) \mid n \in \mathbb{Z}^+\}$ is dense in X. This is useful for showing a system is topologically transitive. We will now prove a theorem that allows us to connect ω-limit sets to equicontinuity.

Theorem 4.10. *Let X be a metric space, $f: X \to X$ be a continuous function, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Now suppose that $x \in X$ is a point at which $\mathcal{F}$ is equicontinuous. If $y \in X$ and $x \in \mathrm{cl}(\bigcup_{n\geq 0}\{f^n(y)\})$, then $\mathcal{F}$ is equicontinuous at y.*

Proof. Suppose $x \in X$ is a point at which $\mathcal{F}$ is equicontinuous and

$$x \in \mathrm{cl}\Big(\bigcup_{n\geq 0}\{f^n(y)\}\Big).$$

Let $\epsilon > 0$ be given. By equicontinuity of $\mathcal{F}$, there exists a $\delta_1 > 0$ such that $d(x, \alpha) < \delta_1$ implies that $d(f^n(x), f^n(\alpha)) < \epsilon/2$, for all $n \in \mathbb{N}$. Without loss of generality, we may take $\delta_1 < \epsilon$. Now let $x \in \mathrm{cl}(\bigcup_{n\geq 0}\{f^n(y)\})$. So there exists an $m \in \mathbb{N}$ such that $d(x, f^m(y)) < \delta_1/2$. Since any finite set of continuous functions is equicontinuous, there exists $\delta_2 > 0$ such that $d(y, \beta) < \delta_2$ implies that $d(f^i(y), f^i(\beta)) < \delta_1/2$ for $1 \leq i \leq m$. Thus

$$d(x, f^m(\beta)) \leq d(x, f^m(y)) + d(f^m(y), f^m(\beta)) < \delta_1/2 + \delta_1/2 = \delta_1.$$

Hence for $j > m$,

$$d(f^j(y), f^j(\beta)) \leq d(f^j(y), f^{j-m}(x)) + d(f^{j-m}(x), f^j(\beta)) < \epsilon/2 + \epsilon/2 = \epsilon.$$

So $d(y, \beta) < \delta_2$ implies that for all $n \in \mathbb{N}$,

$$d(f^n(y), f^n(\beta)) \leq \max\{d(f^i(y), f^i(\beta)), d(f^j(y), f^j(\beta)) \mid 1 \leq i \leq m,\ j > m\} < \epsilon. \quad \square$$

Corollary 4.11. *Let X be a metric space, $f: X \to X$ be a continuous function, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Now suppose that $x \in X$ is a point at which $\mathcal{F}$ is equicontinuous. If $x \in \omega(y)$ then $\mathcal{F}$ is equicontinuous at $y \in X$.*

Proof. Let $\mathcal{F}$ be equicontinuous at $x \in X$ and suppose $x \in \omega(y)$. But

$$x \in \omega(y) \subseteq \mathrm{cl}\Big(\bigcup_{n\geq 0}\{f^n(y)\}\Big),$$

so $\mathcal{F}$ is equicontinuous at x. $\square$

The previous theorems have shown how chaos can be recast in terms of equicontinuity. This is possible because of the intimate connection between chaos and equicontinuity as the following theorem and its corollaries show.

Theorem 4.12. *Let X be a metric space, $f: X \to X$ be a continuous function, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Suppose there exists a point $x \in X$ such that $\mathcal{F}$ is equicontinuous at x. If $\omega(y) = X$, then $\mathcal{F}$ is equicontinuous on $\{f^n(y) \mid n \in \mathbb{Z}^+\}$.*

Proof. If $\omega(y) = X$, then $x \in \omega(y)$. Since x is a point at which $\mathcal{F}$ is equicontinuous, y must also be a point at which $\mathcal{F}$ is equicontinuous. But notice that for any $n \in \mathbb{N}$, $\omega(f^n(y)) = \omega(y) = X$. So for all $n \in \mathbb{N}$, $f^n(y)$ is also a point at which $\mathcal{F}$ is equicontinuous. Therefore, $\mathcal{F}$ is equicontinuous on $\{f^n(y) \mid n \in \mathbb{Z}^+\}$. □

Corollary 4.13 [Kolyada 2004]. *Let X be a metric space, $f: X \to X$ be a continuous onto function, and $\mathcal{F} = \{f^n \mid n \in \mathbb{Z}^+\}$. Suppose that there exists a point $x \in X$ such that $\omega(x) = X$. Then $\mathcal{F}$ is equicontinuous on a dense subset of X, if and only if f is not chaotic on X.*

Proof. First suppose that f is not chaotic on X. Since $\omega(x) = X$, f is topologically transitive on X. Thus f must not have sensitive dependence on initial conditions on X. Hence there exists a point at which $\mathcal{F}$ is equicontinuous. Since $\omega(x) = X$, $\mathcal{F}$ is equicontinuous on $\{f^n(x) \mid n \in \mathbb{Z}^+\}$. But $\{f^n(x) \mid n \in \mathbb{Z}^+\}$ is dense in X.

Now suppose that $\mathcal{F}$ is not equicontinuous on a dense subset of X. Since $\omega(x) = X$, there are no points in X at which $\mathcal{F}$ is equicontinuous. Thus f has sensitive dependence on initial conditions on X. Therefore f is chaotic on X. □

Definition 4.14 (Minimal Set [Robinson 1999]). A set S is a minimal set for f provided (i) S is a closed, nonempty, invariant set and (ii) if B is a closed, nonempty, invariant subset of S, then $B = S$.

Lemma 4.15 [Robinson 1999]. *Let X be a metric space, $f: X \to X$ a continuous map, and $S \subseteq X$ a nonempty compact subset. Then, S is a minimal set if and only if $\omega(x) = S$ for all $x \in S$.*

Corollary 4.16 [Kolyada 2004]. *Let X be a metric space, $f: X \to X$ be a continuous function, S be a nonempty compact minimal subset of X, and $\mathcal{F} = \{f^n \mid n \in \mathbb{N}\}$. Then either f is chaotic on S or $\mathcal{F}$ is equicontinuous on S.*

Proof. First suppose that there is a point $x \in S$ such that $\mathcal{F}$ is equicontinuous at x. Then since S is minimal, $x \in \omega(y)$ for all $y \in S$. Thus $\mathcal{F}$ is equicontinuous on S. But if there are no points in S at which $\mathcal{F}$ is equicontinuous, then f is chaotic on S. □

5. Conclusion

Many definitions have been used to describe when a continuous function will be chaotic. Since we are relating chaos to equicontinuity, we propose to define chaos in terms of a family of continuous functions. Abstracting the connections from the previous section, we offer the following definition of a chaotic family of continuous functions.

Definition 5.1. Let (Y, d) be a metric space. Let $\mathscr{F}$ be a subset of the function space $\mathscr{C}(X, Y)$. Then $\mathscr{F}$ is *chaotic* if there exists $x \in X$ such that $\{f(x) \mid f \in \mathscr{F}\}$ is dense in Y and $\mathscr{F}$ is not equicontinuous at x.

Now suppose that we let X be a metric space, $f\colon X \to X$ be a continuous function, and $\mathscr{F} = \{f^n \mid n \in \mathbb{N}\}$. Then the definition of f being chaotic with our first definition of chaos is equivalent to $\mathscr{F}$ being chaotic with this definition.

It is our belief that chaos terminology should be unified. We have displayed the intrinsic relations of the classical concept of equicontinuity and the modern idea of chaos, hoping to help unify the definitions within chaos theory.

Acknowledgments

I thank my mentor, Dr. Don Reynolds, for his guidance, wisdom, and support. I also thank the Office of Research and Sponsored Programs from the University of Wisconsin – Eau Claire for providing financial support.

References

[Arzelà 1895] C. Arzelà, "Sulle funzioni di linee", *Mem. Accad. Sci. Bologna* (5) **5** (1895), 225–244. JFM 26.0454.01

[Ascoli 1884] G. Ascoli, "Le curve limite di una varietà data di curve", *Atti R. Accad. Lincei Cl. Mem. Sci. Fis. Mat. Nat.* **3** (1884), 521–586. JFM 16.0342.02

[Devaney 1986] R. L. Devaney, *An introduction to chaotic dynamical systems*, Benjamin/Cummings, Menlo Park, CA, 1986. MR 87e:58142 Zbl 0632.58005

[Henry and Trapp 1998] M. Henry and G. Trapp, "Equicontinuity: from stable points to chaotic functions", pp. 471–475 in *Proc. IASTED Conference on Modelling and Simulation* (Pittsburgh, PA), edited by M. H. Hamza, IASTED/ACTA Press, Anaheim, CA, 1998.

[Kolyada 2004] S. F. Kolyada, "Li–Yorke sensitivity and other concepts of chaos", *Ukrain. Math. J.* **56**:8 (2004), 1242–1257. MR 2005k:37028 Zbl 1075.37500

[Li and Yorke 1975] T. Y. Li and J. A. Yorke, "Period three implies chaos", *Amer. Math. Monthly* **82**:10 (1975), 985–992. MR 52 #5898 Zbl 0351.92021

[Martelli 1999] M. Martelli, *Introduction to discrete dynamical systems and chaos*, Interscience Series in Discrete Mathematics and Optimization, Wiley, New York, 1999. MR 2001k:37002 Zbl 1127.37300

[Robinson 1999] C. Robinson, *Dynamical systems: Stability, symbolic dynamics, and chaos*, 2nd ed., Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1999. MR 2001k:37003 Zbl 0914.58021

[Robinson 2004] R. C. Robinson, *An introduction to dynamical systems: continuous and discrete*, Prentice-Hall, Upper Saddle River, NJ, 2004. MR 2005k:37001 Zbl 1073.37001

Received: 2009-12-07 Accepted: 2010-11-12

larsonj@uwec.edu *Department of Mathematics, 508 Hibbard Humanities Hall, University of Wisconsin, Eau Claire, WI 54702-4004, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.371

Minimum rank, maximum nullity and zero forcing number for selected graph families

Edgard Almodovar, Laura DeLoss, Leslie Hogben, Kirsten Hogenson, Kaitlyn Murphy, Travis Peters and Camila A. Ramírez

(Communicated by Chi-Kwong Li)

The minimum rank of a simple graph G is defined to be the smallest possible rank over all symmetric real matrices whose ij-th entry (for $i \neq j$) is nonzero whenever $\{i, j\}$ is an edge in G and is zero otherwise. Maximum nullity is taken over the same set of matrices, and the sum of maximum nullity and minimum rank is the order of the graph. The zero forcing number is the minimum size of a zero forcing set of vertices and bounds the maximum nullity from above. This paper defines the graph families *ciclos* and *estrellas* and establishes the minimum rank and zero forcing number of several of these families. In particular, these families provide examples showing that the maximum nullity of a graph and its dual may differ, and similarly for the zero forcing number.

1. Introduction

All matrices discussed are real and symmetric; the set of $n \times n$ real symmetric matrices will be denoted by $S_n(\mathbb{R})$. A *graph* $G = (V_G, E_G)$ means a simple undirected graph (no loops, no multiple edges) with a finite nonempty set of vertices V_G and edge set E_G (an edge is a two-element subset of vertices). For $A \in S_n(\mathbb{R})$, the *graph* of A, denoted by $\mathcal{G}(A)$, is the graph with vertices $\{1, \ldots, n\}$ and edges $\{\{i, j\} : a_{ij} \neq 0,\ 1 \leq i < j \leq n\}$. The diagonal of A is ignored in determining $\mathcal{G}(A)$.

Let G be a graph. The *set of symmetric matrices described by* G is

$$\mathcal{S}(G) = \{A \in S_n(\mathbb{R}) : \mathcal{G}(A) = G\}.$$

The *maximum nullity of* G is

$$\mathrm{M}(G) = \max\{\operatorname{null} A : A \in \mathcal{S}(G)\},$$

MSC2000: 05C50, 15A03, 15A18.
Keywords: minimum rank, maximum nullity, zero forcing number, dual, ciclo, estrella.
Much of this work was done during the ISU Math REU 2009. Research of E. Almodovar, L. Hogben, K. Murphy, C. Ramírez was supported by DMS 0502354. Research of L. DeLoss, L. Hogben, K. Hogenson, C. Ramírez was supported by DMS 0750986.

and the *minimum rank* of G is

$$\operatorname{mr}(G) = \min\{\operatorname{rank} A : A \in \mathcal{S}(G)\}.$$

Clearly $\operatorname{mr}(G) + \operatorname{M}(G) = |G|$, where the *order* $|G|$ is the number of vertices of G. Extensive work has been done on the problem of determining minimum rank and maximum nullity of graphs. A variety of techniques have been developed to determine the minimum rank, and the minimum rank of numerous families of graphs has been determined, but in general the problem remains open. See [Fallat and Hogben 2007] for a survey of results and discussion of the motivation for the minimum rank problem.

The zero forcing number was introduced in [AIM 2008] and the associated terminology was extended in [Barioli et al. 2010; 2009; Edholm et al. 2010; Hogben 2010; Huang et al. 2010]. Let G be a graph with each vertex colored either white or black. Vertices change color according to the *color-change rule*: if u is a black vertex and exactly one neighbor w of u is white, then change the color of w to black. When the color-change rule is applied to u to change the color of w, we say u *forces* w and write $u \to w$. Given a coloring of G, the *derived set* is the set of black vertices obtained by applying the color-change rule until no more changes are possible. A *zero forcing set* for G is a subset of vertices Z such that if initially the vertices in Z are colored black and the remaining vertices are colored white, then the derived set is all the vertices of G. The *zero forcing number* $\operatorname{Z}(G)$ is the minimum of $|Z|$ over all zero forcing sets $Z \subseteq V(G)$.

Theorem 1.1 [AIM 2008, Proposition 2.4]. *For any graph G, $\operatorname{M}(G) \le \operatorname{Z}(G)$.*

Let $G = (V_G, E_G)$ be a graph and $W \subseteq V_G$. The *induced subgraph* $G[W]$ is the graph with vertex set W and edge set $\{\{v, w\} \in E_G : v, w \in W\}$. The subgraph induced by $V_G \setminus W$ is also denoted by $G - W$, or in the case W is a single vertex $\{v\}$, by $G - v$. Minimum rank is monotone on induced subgraphs, that is, for any $W \subseteq V_G$, $\operatorname{mr}(G[W]) \le \operatorname{mr}(G)$. If e is an edge of $G = (V_G, E_G)$, the subgraph $(V_G, E_G \setminus \{e\})$ is denoted by $G - e$. We denote the complete graph on n vertices by K_n, the cycle on n vertices by C_n and the path on n vertices by P_n. The *union* of $G_i = (V_i, E_i)$, for $i = 1, \ldots, h$, is $\bigcup_{i=1}^{h} G_i = \left(\bigcup_{i=1}^{h} V_i, \bigcup_{i=1}^{h} E_i\right)$. An (edge) *covering* of a graph G is a set of subgraphs $\{G_i, i = 1, \ldots, h\}$ such that $G = \bigcup_{i=1}^{h} G_i$. The following observation is useful when bounding minimum rank from above by using a covering to exhibit a low rank matrix.

Observation 1.2 [Fallat and Hogben 2007]. If $G = \bigcup_{i=1}^{h} G_i$, then

$$\operatorname{mr}(G) \le \sum_{i=1}^{h} \operatorname{mr}(G_i).$$

The *path cover number* $\mathrm{P}(G)$ of G is the smallest positive integer m such that there are m vertex-disjoint induced paths in G such that every vertex of G is a vertex of one of the paths. The path cover number was first used in the study of minimum rank and maximum eigenvalue multiplicity in [Johnson and Leal Duarte 1999] (since the matrices in $\mathcal{S}(G)$ are symmetric, algebraic and geometric multiplicities of eigenvalues are the same, and since the diagonal is free, maximum eigenvalue multiplicity is the same as maximum nullity). Johnson and Duarte [1999] showed that for a tree T, $\mathrm{P}(T) = \mathrm{M}(T)$; however, Barioli et al. [2004] showed that $\mathrm{P}(G)$ and $\mathrm{M}(G)$ are not comparable for graphs unless some restriction is imposed on the type of graph. A graph is *planar* if it can be drawn in the plane with no edge crossings. A graph is *outerplanar* if it has a drawing in the plane without crossing edges such that one face contains all vertices. Recently Sinkovic established the following relationship between $\mathrm{P}(G)$ and $\mathrm{M}(G)$ for outerplanar graphs.

Theorem 1.3 [Sinkovic 2010]. *If G is an outerplanar graph, then* $\mathrm{P}(G) \geq \mathrm{M}(G)$.

A connected graph G is *k-connected* if for any set of vertices S such that $G - S$ is disconnected, $|S| \geq k$. The *dual* G^d of a 3-connected planar graph G is the graph obtained by putting a dual vertex in each region of a plane drawing of G and a dual edge between two dual vertices whenever the original regions share an original edge (we assume the graph is 3-connected to ensure that the dual is determined by the graph rather than a particular plane embedding). At a research meeting devoted to minimum rank at the American Institute of Mathematics, the following questions were asked:

Question 1.4. If G is a 3-connected planar graph, is it true that $\mathrm{M}(G^d) = \mathrm{M}(G)$?

Question 1.5. If G is a 3-connected planar graph, is it true that $\mathrm{Z}(G^d) = \mathrm{Z}(G)$?

In Section 3 we give examples of graphs G such that $\mathrm{M}(G^d) \neq \mathrm{M}(G)$ and $\mathrm{Z}(G^d) \neq \mathrm{Z}(G)$. The examples are taken from the family of estrellas. This family and the related family of ciclos are defined in Section 2, and the minimum ranks, maximum nullities, and zero forcing numbers of some members of these families are established. In Section 4 we determine the vertex spreads and edge spreads of select members of the ciclo and estrella families, thereby computing the minimum ranks, maximum nullities, and zero forcing numbers of additional families of graphs (spreads are defined in Section 4).

2. Ciclo and estrella graph families

Definition 2.1. Let G be a graph and let e be an edge of G. A *t-ciclo of G with e*, denoted by $C_t(G, e)$, is constructed from a t-cycle C_t and t copies of G by identifying each edge of C_t with the edge e in one copy of G. If a symbol for the graph identifies a specific edge, or if G is edge-transitive (so it is not necessary to

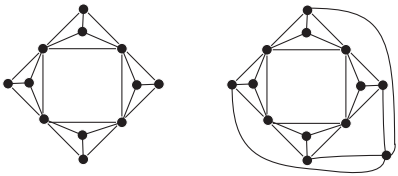


Figure 1. The complete ciclo $C_4(K_4)$ and the complete estrella $S_4(K_4)$.

specify edge e), then the notation $C_t(G)$ is used. A vertex on C_t is called a *cycle* vertex.

The ciclo $C_4(K_4)$ is shown in Figure 1. Ciclos of complete graphs are discussed in Section 2A. The order of $C_t(G)$ is $(|G|-1)t$. Note that although $C_t(G, e)$ is defined as a union of a t-cycle C_t and t copies of G to explain the construction, in fact $C_t(G, e)$ is a union of just the t copies of G.

Definition 2.2. Let G be a graph, let e be an edge of G, and let v be a vertex of G that is not an endpoint of e. A *t-estrella of G with e and v*, denoted by $S_t(G, e, v)$, is the union of a t-ciclo $C_t(G, e)$ and the complete bipartite graph $K_{1,t}$ with each degree one vertex of $K_{1,t}$ identified with one copy of v. If a symbol for the graph identifies a specific edge and vertex, or if G is vertex- and edge-transitive (so it is not necessary to specify e and v), then the notation $S_t(G)$ is used. The degree t vertex of the $K_{1,t}$ used to construct the estrella is called the *star* vertex of the estrella, and every neighbor of the star vertex is called a *starneighbor* vertex. A cycle vertex in the ciclo that is used to construct the estrella is also called a *cycle* vertex in the estrella.

The estrella $S_4(K_4)$ is shown in Figure 1. The order of $S_t(G)$ is $(|G|-1)t+1$. Estrellas of complete graphs are discussed in Section 2B. The families of ciclos and estrellas formed from house graphs (see Sections 2C and 2D) are introduced because of their importance as examples answering the duality questions (see Questions 1.4 and 1.5 above). Related families of ciclos are studied in Sections 2E and 2F. Another natural family of ciclos are the cycle ciclos, discussed in Section 2G.

2A. *The complete ciclo $C_t(K_r)$.*

Definition 2.3. The *complete ciclo*, denoted by $C_t(K_r)$, is the ciclo of the complete graph K_r, with $t, r \geq 3$. (Note that K_r is edge-transitive.) A vertex not on C_t is called a *noncycle* vertex.

The order of $C_t(K_r)$ is $(r-1)t$.

Theorem 2.4. *For $t \geq 3$ and $r \geq 3$,*

$$\mathrm{M}(C_t(K_r)) = \mathrm{Z}(C_t(K_r)) = (r-2)t, \quad \mathrm{mr}(C_t(K_r)) = t.$$

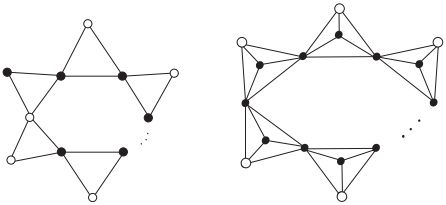


Figure 2. The zero forcing sets for the complete ciclos $C_t(K_3)$ and $C_t(K_r)$.

Proof. First, we will derive a lower bound for the maximum nullity. We know from Observation 1.2 that the minimum rank of a graph will be less than or equal to the sum of the minimum ranks of the subgraphs in a covering of it. Since every $C_t(K_r)$ can be covered by t copies of K_r graphs, each of minimum rank 1, $\mathrm{mr}(C_t(K_r)) \le t$ and $(r-2)t \le \mathrm{M}(C_t(K_r))$.

The zero forcing number can be used to bound the maximum nullity from above. There are many possible zero forcing sets of minimum cardinality, but it suffices to exhibit one for each of the two cases $r = 3$ and $r \ge 4$ (see Figure 2).

Case $r = 3$. A set Z consisting of $t-1$ cycle vertices and one noncycle vertex adjacent to the cycle vertex that is not in Z is a zero forcing set of t vertices.

Case $r \ge 4$. Let Z consist of all the cycle vertices and for each K_r, all but one of the noncycle vertices. Then Z is a zero forcing set because there will always be at least one black noncycle vertex in each K_r that will force the one white noncycle vertex, coloring the entire graph. Note that $|Z| = (r-2)t$.

In either case, $\mathrm{M}(C_t(K_r)) \le \mathrm{Z}(C_t(K_r)) \le (r-2)t$. □

2B. *The complete estrella $S_t(K_r)$.*

Definition 2.5. The *complete estrella*, denoted by $S_t(K_r)$, is the estrella of the complete graph K_r, with $t, r \ge 3$. (Note that K_r is vertex- and edge-transitive.) A vertex in $S_t(K_r)$ that is not the star vertex, not a starneighbor vertex, and not a cycle vertex is called a *standard* vertex.

The order of $S_t(K_r)$ is $(r-1)t+1$, and $S_t(K_4)$ is planar and 3-connected.

Theorem 2.6. *For $t \ge 3$ and $r \ge 4$,*

$$\mathrm{mr}(S_t(K_r)) = t+2 \quad \textit{and} \quad \mathrm{M}(S_t(K_r)) = \mathrm{Z}(S_t(K_r)) = (r-2)t-1.$$

Proof. Note that $|S_t(K_r)| = (r-1)t+1$. Since $S_t(K_r)$ can be covered by t copies of K_r (each of minimum rank 1) and one $K_{1,t}$ (of minimum rank 2), $\mathrm{mr}(S_t(K_r)) \le t+2$ and $(r-2)t-1 \le \mathrm{M}(S_t(K_r))$.

Define a set Z consisting of all cycle vertices and all but one standard vertices; note $|Z| = (r-2)t-1$. We claim Z is a zero forcing set. In each of the complete

graphs that has all its standard vertices in Z, any black standard vertex can force the one white starneighbor vertex. Then any one of the (now) black starneighbor vertices can force the star vertex. Then the star vertex forces the one remaining white starneighbor vertex, and any black neighbor forces the last white vertex. So the entire graph is black, establishing the claim. Thus,

$$\mathrm{M}(S_t(K_r)) \leq \mathrm{Z}(S_t(K_r)) \leq (r-2)t - 1. \qquad \square$$

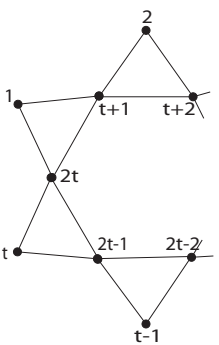
Theorem 2.7. *For $t \geq 3$,*

$$\mathrm{mr}(S_t(K_3)) = t \quad \textit{and} \quad \mathrm{M}(S_t(K_3)) = \mathrm{Z}(S_t(K_3)) = t + 1.$$

Proof. By Theorem 2.4, $\mathrm{mr}(C_t(K_3)) = t$, and since $C_t(K_3)$ is an induced subgraph of $S_t(K_3)$,

$$t = \mathrm{mr}(C_t(K_3)) \leq \mathrm{mr}(S_t(K_3)).$$

To show that $\mathrm{mr}(S_t(K_3)) \leq t$, we construct a matrix of rank t in $\mathcal{S}(C_t(K_3))$ and extend it to a matrix in $\mathcal{S}(S_t(K_3))$ without changing the rank of the matrix. Number the vertices of $C_t(K_3)$ as follows:



Define the $t \times t$ matrix B to be

$$B = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 & -1 \\ -1 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -1 & 1 \end{bmatrix}.$$

Note that the sum of each of the rows and the sum of each of the columns equal zero. Then the $2t \times 2t$ matrix

$$A = \begin{bmatrix} I & B \\ B^T & B^T B \end{bmatrix} \in \mathcal{S}(C_t(K_3))$$

has rank $A = t$. Extend the matrix A to the $(2t+1) \times (2t+1)$ matrix

$$A' = \begin{bmatrix} I_{t\times t} & B & \mathbf{1}_t \\ B^T & B^TB & 0_t \\ \mathbf{1}_t^T & 0_t^T & t \end{bmatrix} \in \mathcal{S}(S_t(K_3)).$$

Note that B^TB shares properties with B in that for each row and column, the sum is zero as well. Thus the entries of the new column $2t+1$ of A' is the sum of the columns of A, and, similarly for the rows. Thus rank $A' = t$, and $\mathrm{mr}(S_t(K_3)) \le t$. □

2C. *The house ciclo $C_t(H_0)$.*

Definition 2.8. A *house* H_0 (also called an *empty house*) is the union of a 3-cycle and a 4-cycle with one edge in common, shown on the left in Figure 3. The symbol H_0 also designates the specific edge e and vertex v shown in the figure (this figure also includes numbering that will be used later). A *house ciclo* is $C_t(H_0) = C_t(H_0, e)$.

The house ciclo $C_4(H_0)$ is shown on the right in Figure 3. Note that the order of $C_t(H_0)$ is $4t$ and $C_t(H_0)$ is outerplanar.

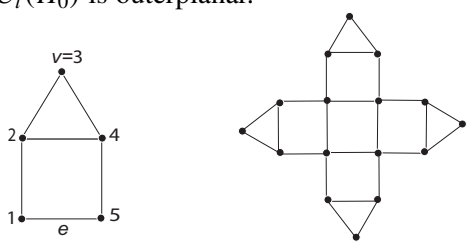
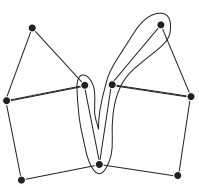


Figure 3. The house H_0 and the house ciclo $C_4(H_0)$.

Observation 2.9. For $t \ge 3$, $\mathrm{P}(C_t(H_0)) \le t$, because we can create a covering with t paths:



Theorem 2.10. *For $t \ge 3$, $\mathrm{M}(C_t(H_0)) = t$ and $\mathrm{mr}(C_t(H_0)) = 3t$.*

Proof. Because house ciclos are outerplanar, Theorem 1.3 and Observation 2.9 give the upper bound $\mathrm{M}(C_t(H_0)) \le \mathrm{P}(C_t(H_0)) \le t$ for the maximum nullity of $C_t(H_0)$. Using the obvious covering of the house ciclo $C_t(H_0)$ by the set of t houses H_0, and the fact that $\mathrm{mr}(H_0) = 3$, we have the same lower bound on maximum nullity: $\mathrm{M}(C_t(H_0)) = |C_t(H_0)| - \mathrm{mr}(C_t(H_0)) \ge |C_t(H_0)| - 3t \ge t$. Therefore, $\mathrm{M}(C_t(H_0)) = t$ and $\mathrm{mr}(C_t(H_0)) = 3t$. □

Theorem 2.11. *For even $t \geq 4$, $\mathrm{Z}(C_t(H_0)) = t$.*

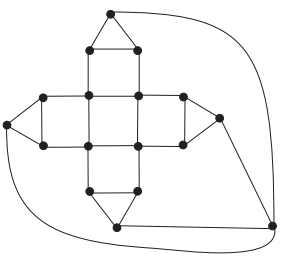
Proof. Since $t = \mathrm{M}(C_t(H_0)) \leq \mathrm{Z}(C_t(H_0))$, it suffices to exhibit a zero forcing set Z with $|Z| = t$. Let Z consist of pairs chosen in alternate houses of $C_t(H_0)$ going around the cycle (2 vertices in the first house, skip the second house, 2 vertices in the third house, skip the fourth house, etc.), where each pair of vertices consists of the peak vertex $v = 3$ and its neighbor 2, labeled as in Figure 3. Because t is even, $|Z| = t$. Within each house that contains two black vertices, the remaining three vertices are forced to turn black. Then, the remaining three white vertices in a house in between two houses having all vertices black will be forced. So Z is a zero forcing set. □

In the case t is odd, the method used in the proof of Theorem 2.11 will produce a zero forcing set of order $t+1$, so for t odd, $\mathrm{Z}(C_t(H_0)) \leq t+1$. For odd $t \leq 9$, it has been established by use of software [DeLoss et al. 2008] that $\mathrm{Z}(C_t(H_0)) = t+1$.

2D. *The house estrella $S_t(H_0)$.*

Definition 2.12. A *house estrella* is $S_t(H_0) = S_t(H_0, e, v)$, where v and e are as shown in Figure 3.

Here is the house estrella $S_4(H_0)$. Note that the order of $S_t(H_0)$ is $4t+1$ and $S_t(H_0)$ is planar and 3-connected.



We adopt the following convention for numbering the vertices of $S_t(H_0)$. We start the numbering on one of the houses from the lower left corner, starting with 1, and complete the numbering clockwise around the house, as in Figure 3. When that house is done, continue to the clockwise-adjacent house. The star vertex is numbered $4t+1$.

Theorem 2.13. *For $t \geq 3$,*

$$\mathrm{mr}(S_t(H_0)) = 3t \quad \textit{and} \quad \mathrm{M}(S_t(H_0)) = t+1.$$

Proof. In Theorem 2.10, it was shown that $\mathrm{mr}(C_t(H_0)) = 3t$, and since $C_t(H_0)$ is an induced subgraph of $S_t(H_0)$, we have $3t = \mathrm{mr}(C_t(H_0)) \leq \mathrm{mr}(S_t(H_0))$.

Next, we will construct a specific matrix $A \in \mathcal{S}(C_t(H_0))$ having rank $A = 3t$ that we can extend to a matrix A' such that $\mathcal{G}(A') = S_t(H_0)$ and rank $A' = 3t$, thus showing that the minimum rank of $S_t(H_0)$ is also $3t$.

Define the submatrices

$$U = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad W = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

The sum of the adjacency matrix of $C_t(H_0)$ and the $4t \times 4t$ identity matrix is the $4t \times 4t$ matrix

$$A = \begin{bmatrix} V & W & 0 & 0 & \dots & 0 & 0 & 0 & U \\ U & V & W & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & U & V & W & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & U & V & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & V & W & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & U & V & W & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & U & V & W \\ W & 0 & 0 & 0 & \dots & 0 & 0 & U & V \end{bmatrix}. \tag{1}$$

Note that V is the submatrix corresponding to the adjacencies between the vertices numbered $4s+1$, $4s+2$, $4s+3$, $4s+4$, and V lies on the diagonal.

Let $\boldsymbol{b}$ be the 0,1-vector describing the adjacencies of the star vertex. If $\boldsymbol{b} \in$ range A, then there exists a vector $\boldsymbol{x}$ such that $A\boldsymbol{x} = \boldsymbol{b}$ and for

$$A' = \begin{bmatrix} A & A\boldsymbol{x} \\ \boldsymbol{x}^T A & \boldsymbol{x}^T A \boldsymbol{x} \end{bmatrix},$$

we have

$$\operatorname{rank} A' = \operatorname{rank} A \quad \text{and} \quad \mathcal{G}(A') = S_t(H_0).$$

Thus it suffices to show that $\boldsymbol{b}$ is in the range of A.

To prove $\boldsymbol{b} \in$ range A, we show that $\boldsymbol{b} \in (\ker A)^\perp$ and apply the fact that for any real symmetric matrix A, $(\ker A)^\perp =$ range A [Han and Neumann 2007, Fact 5.2.15]. Establishing $\boldsymbol{b} \in (\ker A)^\perp$ can be done by finding a basis for the kernel of A and showing that $\boldsymbol{b}$ is orthogonal to the vectors in the basis of the kernel. To construct the basis, we construct t linearly independent null vectors (and note that null $A \le \mathrm{M}(C_t(H_0)) = t$ by Theorem 2.10).

Let

$$\alpha = [0, 0, -1, 1], \quad \omega = [0, -1, 1, 0], \quad \beta = [0, -1, 0, 1], \quad 0 = [0, 0, 0, 0].$$

Then construct the vectors in the following manner:

$$\boldsymbol{v}_1 = [\underbrace{\beta, \beta, \dots, \beta, \beta, \beta}_{t}]^T,$$

$$\begin{aligned}
\boldsymbol{v}_2 &= [\alpha, \underbrace{\beta, \ldots, \beta, \beta}_{t-2}, \omega]^T, \\
\boldsymbol{v}_3 &= [\alpha, \underbrace{\beta, \ldots, \beta}_{t-3}, \omega, 0]^T, \\
&\vdots \\
\boldsymbol{v}_r &= [\alpha, \underbrace{\beta, \ldots, \beta}_{t-r}, \omega, \underbrace{0, \ldots, 0}_{r-2}]^T, \\
&\vdots \\
\boldsymbol{v}_t &= [\alpha, \omega, \underbrace{0, \ldots, 0}_{t-2}]^T.
\end{aligned}$$

To show that the vectors $\boldsymbol{v}_i, i = 1, \ldots, t$ are null vectors of A it is sufficient to observe that

$$[U \;\; V \;\; W]_{4\times 12} \begin{bmatrix}
\beta^T & \omega^T & 0^T & 0^T & \alpha^T & \alpha^T & \alpha^T & \beta^T & \beta^T & \omega^T & \beta^T & \omega^T & 0^T \\
\beta^T & \alpha^T & \alpha^T & \alpha^T & \beta^T & \beta^T & \omega^T & \beta^T & \omega^T & 0^T & \omega^T & 0^T & 0^T \\
\beta^T & \beta^T & \beta^T & \omega^T & \beta^T & \omega^T & 0^T & \omega^T & 0^T & 0^T & \alpha^T & \alpha^T & \alpha^T
\end{bmatrix}_{12\times 13} = 0_{4\times 13}.$$

Next, we show that the vectors $\boldsymbol{v}_i, i = 1, \ldots, t$ are linearly independent, viewing these vectors as block vectors (as constructed). Suppose $\sum_{i=1}^{t} \gamma_i \boldsymbol{v}_i = 0$. The vector $\boldsymbol{v}_1$ has $\beta^T = [0, -1, 0, 1]^T$ as the last block of the vector, so the last coordinate is 1. The vector $\boldsymbol{v}_2$ has $\omega^T = [0, -1, 1, 0]^T$ as the last block of the vector, so the last coordinate is 0, and the last coordinate of $\boldsymbol{v}_i, i \geq 3$ is also 0. Thus $\gamma_1 = 0$. Assuming $\gamma_k = 0$, by examining block $t - k + 1$ of $\sum_{i=k+1}^{t} \gamma_i \boldsymbol{v}_i = 0$, we see that $\gamma_{k+1} = 0$. Thus the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t$ are linearly independent.

To complete the proof it suffices to show that 0,1-vector $\boldsymbol{b}$ describing the adjacencies of the star vertex is orthogonal to $\ker A$. Let $\varphi = [0, 0, 1, 0]$; then $\boldsymbol{b} = [\varphi, \ldots, \varphi]^T$. Note that

$$\varphi \cdot \alpha = -1\,, \quad \varphi \cdot \beta = 0\,, \quad \varphi \cdot \omega = 1.$$

Then

$$\boldsymbol{b} \cdot \boldsymbol{v}_1 = [\varphi, \ldots, \varphi]^T \cdot [\beta, \ldots, \beta]^T = \sum_{i=1}^{t} \varphi \cdot \beta = 0,$$

and for $2 \leq r \leq t$,

$$\begin{aligned}
\boldsymbol{b} \cdot \boldsymbol{v}_r &= [\varphi, \ldots, \varphi]^T \cdot [\alpha, \underbrace{\beta, \ldots, \beta}_{t-r}, \omega, \underbrace{0, \ldots, 0}_{r-2}]^T \\
&= \varphi \cdot \alpha + \sum_{i=1}^{t-r} \varphi \cdot \beta + \varphi \cdot \omega + \sum_{i=1}^{r-2} \varphi \cdot 0 \\
&= -1 + 0 + 1 + 0 = 0.
\end{aligned}$$

Therefore, $\boldsymbol{b} \in (\ker A)^{\perp}$. □

Corollary 2.14. *For even* $t \geq 4$, $\mathrm{Z}(S_t(H_0)) = t+1$.

Proof. The zero forcing set Z of Theorem 2.11 together with the star vertex is a zero forcing set of order $t+1$ and the result then follows from Theorem 2.13. □

For t odd, there is a zero forcing set of order $t+2$, so for t odd, $\mathrm{Z}(S_t(H_0)) \leq t+2$. For odd $t \leq 9$, it has been established by use of software [DeLoss et al. 2008] that $\mathrm{Z}(S_t(H_0)) = t+2$.

2E. *The half-house ciclo* $C_t(H_1)$.

Definition 2.15. A *half-full house* or *half-house* H_1 is a house with one diagonal in the square, as shown on the left in Figure 4. The symbol H_1 also designates the specific edge e and vertex v, as shown in this figure. A *half-house ciclo* is a ciclo of half-houses $C_t(H_1) = C_t(H_1, e)$.

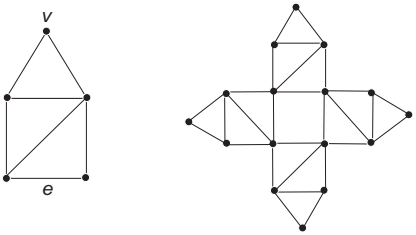


Figure 4. The half-house H_1 and half-house ciclo $C_4(H_1)$.

The half-house ciclo $C_4(H_1)$ is also shown in Figure 4. Note that $\mathrm{mr}(H_1) = 3 = |H_1| - 2$. The order of $C_t(H_1)$ is $4t$ and that $C_t(H_1)$ is outerplanar. Half-full house ciclos have many properties in common with house ciclos. The proofs of the results below are analogous to the proofs of the corresponding results for house ciclos, and are omitted.

Observation 2.16. For $t \geq 3$, $\mathrm{P}(C_t(H_1)) \leq t$.

Theorem 2.17. *For* $t \geq 3$,

$$\mathrm{M}(C_t(H_1)) = t \quad \textit{and} \quad \mathrm{mr}(C_t(H_1)) = 3t.$$

Theorem 2.18. *For even* t, $\mathrm{Z}(C_t(H_1)) = t$.

In the case t is odd, $\mathrm{Z}(C_t(H_1)) \leq t+1$.

2F. *The full-house ciclo* $C_t(H_2)$.

Definition 2.19. A *full house* H_2 is the union of K_4 and K_3 with one edge in common, or equivalently, a house with both diagonals in the square, as shown on the left in Figure 5. The symbol H_2 also designates the specific edge e and vertex v, as shown in this figure (this figure also includes numbering that will be used later). A *full house ciclo* is a ciclo of full houses $C_t(H_2) = C_t(H_2, e)$.

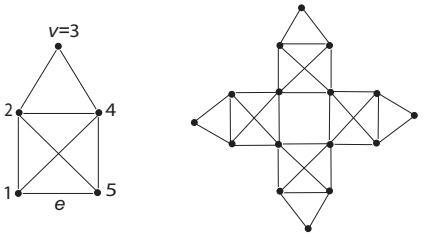


Figure 5. The full house H_2 and full house ciclo $C_4(H_2)$.

The full house ciclo $C_4(H_2)$ is also shown in Figure 5. Note that the order of $C_t(H_2)$ is $4t$ and $\mathrm{mr}(H_2) = 2$. We adopt the following convention for numbering the vertices of $C_t(H_2)$. We start the numbering on one of the houses from the lower left corner, starting with 1, and complete the numbering clockwise around the house, as in Figure 5. When that house is done, continue with the clockwise-adjacent house.

Theorem 2.20. *For* $t \geq 3$, $\mathrm{M}(C_t(H_2)) = \mathrm{Z}(C_t(H_2)) = 2t$ *and* $\mathrm{mr}(C_t(H_2)) = 2t$.

Proof. We can bound the maximum nullity from below by bounding the minimum rank from above using a covering of $C_t(H_2)$ with t copies of the full house. Since a full house has minimum rank 2 and $|C_t(H_2)| = 4t$, $2t \leq \mathrm{M}(C_t(H_2))$.

Next, we can derive an upper bound for the maximum nullity by showing that

$$Z = \{1, 2, 3, 6, 7, 10, 11, \ldots, 4k{+}2, 4k{+}3, \ldots, 4(t{-}2){+}2, 4(t{-}2){+}3, 4(t{-}1){+}2\}$$

is a zero forcing set. There are three black vertices of the four vertices in the first house, one in the last, and two in every other house (where the first four of the five vertices actually in a house are associated with that house to avoid duplication). To see that Z is a zero forcing set, examine the first full house. Since vertices 1, 2, and 3 are black, the other two vertices in house 1 are forced, which means the next house already has its first vertex $5 = 4(2-1)+1$ black, in addition to 6 and 7. This process will continue around the ciclo until we reach the last full house, house t, which now has vertices $4(t-1)+1$, $4(t-1)+2$, and 1 colored, so the remaining two vertices in this house can be forced. Since $|Z| = 2t$,

$$2t \leq \mathrm{M}(C_t(H_2)) \leq \mathrm{Z}(C_t(H_2)) \leq |Z| = 2t,$$

and we have equality throughout. □

2G. *The cycle ciclo* $C_t(C_r)$.

Definition 2.21. A *cycle ciclo* is a ciclo of cycles $C_t(C_r)$, $r \geq 4$.

The cycle ciclo $C_4(C_6)$ is shown in Figure 6. The order of $C_t(C_r)$ is $(r-1)t$ and $C_t(C_r)$ is outerplanar. Cycle ciclos have many properties in common with house

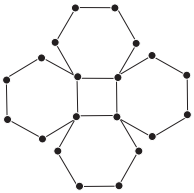


Figure 6. The cycle ciclo $C_4(C_6)$.

ciclos. Note that $\mathrm{mr}(C_r) = r-2 = |C_r|-2$ and $\mathrm{mr}(H_0) = 3 = |H_0|-2$, and $\mathrm{Z}(C_r) = 2 = \mathrm{Z}(H_0)$. The proofs of the results below are analogous to the proofs of the corresponding results for house ciclos, and are omitted.

Observation 2.22. For $t \geq 3$, $\mathrm{P}(C_t(C_r)) \leq t$.

Theorem 2.23. *For* $t \geq 3$, $\mathrm{M}(C_t(C_r)) = t$ *and* $\mathrm{mr}(C_t(C_r)) = (r-2)t$.

Theorem 2.24. *For even* $t \geq 4$, $\mathrm{Z}(C_t(C_r)) = t$.

In the case t is odd, $\mathrm{Z}(C_t(C_r)) \leq t+1$.

2H. ***Summary.*** The results established in this section for certain families of ciclos and estrellas can be summarized as follows:

Graph G	$\lvert G\rvert$	$\mathrm{mr}(G)$	$\mathrm{M}(G)$	$\mathrm{Z}(G)$
$C_t(K_r)$	$(r-1)t$	t	$(r-2)t$	$(r-2)t$
$C_t(H_0)$	$4t$	$3t$	t	t if t even $\leq t+1$ if t odd
$C_t(H_1)$	$4t$	$3t$	t	t if t even $\leq t+1$ if t odd
$C_t(H_2)$	$4t$	$2t$	$2t$	$2t$
$C_t(C_r)$ $(r \geq 4)$	$(r-1)t$	$(r-2)t$	t	t if t even $\leq t+1$ if t odd
$S_t(K_r)(r \geq 4)$	$(r-1)t+1$	$t+2$	$(r-2)t-1$	$(r-2)t-1$
$S_t(K_3)$	$2t+1$	t	$t+1$	$t+1$
$S_t(H_0)$	$4t+1$	$3t$	$t+1$	$t+1$ if t even $\leq t+2$ if t odd

3. Complete estrellas and house estrellas as duals

The next theorem and our previous results show that complete estrellas and house estrellas provide a negative answer to Questions 1.4 and 1.5.

Theorem 3.1. *The dual of the complete estrella* $S_t(K_4)$ *is the house estrella* $S_t(H_0)$.

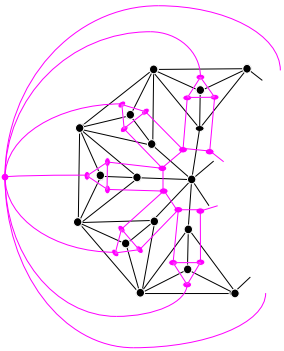


Figure 7. The house estrella $S_t(H_0)$, in lighter color, as the dual of the complete estrella $S_t(K_4)$, in black.

Proof. Since $S_t(K_4)$ is a 3-connected graph, its dual is independent of how it is drawn in the plane, so we draw $S_t(K_4)$ with the star vertex in the center, as shown by the black lines in Figure 7. Each K_4 together with the star vertex produces a house as its dual, so ignoring the infinite region we obtain the house ciclo $C_t(H_0)$ as the dual, shown in a lighter color in the figure. The last step to creating the dual is to add a dual point that represents the infinite region outside the $S_t(K_4)$, and it connects to the vertex numbered 3 of each house (with numbering as in Figure 3), creating the house estrella $S_t(H_0)$ (Figure 7). □

Corollary 3.2. *The example in Theorem 3.1 answers Questions 1.4 and 1.5 in the negative, since* $\mathrm{M}(S_4(K_4)) = \mathrm{Z}(S_4(K_4)) = 7$ *and* $\mathrm{M}(S_4(H_0)) = \mathrm{Z}(S_4(H_0)) = 5$.

4. Rank spread, null spread, and zero spread

If the minimum rank, maximum nullity, and/or zero forcing number are known for a graph G, it is sometimes possible to use this information to determine the same parameter for the graph obtained from G by deleting a vertex or edge. In this section we determine the minimum rank/maximum nullity and zero forcing number of any complete ciclo or complete estrella from which one vertex or one edge has been deleted. Note that a complete ciclo has two types of vertex, a cycle vertex and a noncycle vertex. For a complete estrella there can be four types of vertex: the star vertex, a starneighbor vertex, a cycle vertex, and a standard vertex; note that $S_t(K_3)$ does not have any standard vertices.

4A. ***Vertex spreads of complete ciclos and estrellas.*** Let G be a graph and v be a vertex in G. The *rank spread* of v, defined in [Barioli et al. 2004], is

$$\mathrm{r}_v(G) = \mathrm{mr}(G) - \mathrm{mr}(G - v),$$

and it is known [Nylen 1996] that

$$0 \le \mathrm{r}_v(G) \le 2.$$

By analogy with the rank spread, the null spread and the zero spread were defined in [Edholm et al. 2010]. The *null spread* of v is $\mathrm{n}_v(G) = \mathrm{M}(G) - \mathrm{M}(G-v)$. The *zero spread* of v is $\mathrm{z}_v(G) = \mathrm{Z}(G) - \mathrm{Z}(G-v)$. Clearly, for any graph G and vertex v of G,

$$\mathrm{r}_v(G) + \mathrm{n}_v(G) = 1,$$

and thus

$$-1 \le \mathrm{n}_v(G) \le 1.$$

Theorem 4.1 [Huang et al. 2010; Edholm et al. 2010]. *For every graph G and vertex v of G,*

$$-1 \le \mathrm{z}_v(G) \le 1.$$

As might be expected from the loose relationship between zero forcing number and maximum nullity, the parameters $\mathrm{n}_v(G)$ and $\mathrm{z}_v(G)$ are not comparable, and examples of this are given in [Edholm et al. 2010]. However, under certain circumstances we can use one spread to determine the other.

Observation 4.2. [Barrett et al. 2008] Let G be a graph such that $\mathrm{M}(G) = \mathrm{Z}(G)$ and let v be a vertex of G. Then $\mathrm{n}_v(G) \ge \mathrm{z}_v(G)$, and so if $\mathrm{z}_v(G) = 1$, then $\mathrm{n}_v(G) = 1$ (equivalently, $\mathrm{r}_v(G) = 0$).

Theorem 4.3. *For any vertex v, $\mathrm{M}(C_t(K_r) - v) = \mathrm{Z}(C_t(K_r) - v) = (r-2)t - 1$, or equivalently, $\mathrm{n}_v(C_t(K_r)) = \mathrm{z}_v(C_t(K_r)) = 1$.*

Proof. We exhibit a zero forcing set Z for $C_t(K_r) - v$ such that $|Z| = (r-2)t - 1$ (here $r \ge 3$). Since $\mathrm{Z}(C_t(K_r)) = (r-2)t$ and $\mathrm{z}_v(C_t(K_r)) \le 1$, $\mathrm{z}_v(C_t(K_r)) = 1$ and $\mathrm{Z}(C_t(K_r) - v) = (r-2)t - 1$. Since $\mathrm{M}(C_t(K_r)) = \mathrm{Z}(C_t(K_r))$, by Observation 4.2 $\mathrm{n}_v(C_t(K_r)) = 1$, and thus $\mathrm{M}(C_t(K_r) - v) = (r-2)t - 1$. When exhibiting a zero forcing set, we separate $C_t(K_3)$ from $C_t(K_r)$ with $r \ge 4$. For each of these two cases, there are two types of vertex v, a cycle vertex and a noncycle vertex. The zero forcing sets Z are illustrated as black vertices in Figure 8.

Case $C_t(K_3)$. For a cycle vertex v, let the two noncycle neighbors of v in $C_t(K_3)$ be denoted by u and w. Then Z consists of every noncycle vertex except w. For a noncycle vertex v, let the two neighbors of v (both of which are cycle vertices) be denoted by u and w. Then Z consists of u and every noncycle vertex except for the one adjacent to w.

Case $C_t(K_r)$. Note that each of the one or two copies of K_r in which v was a vertex has now become K_{r-1}. For a cycle vertex v, Z consists of all the remaining cycle vertices and all but one noncycle vertex in each K_r or K_{r-1}. For a noncycle

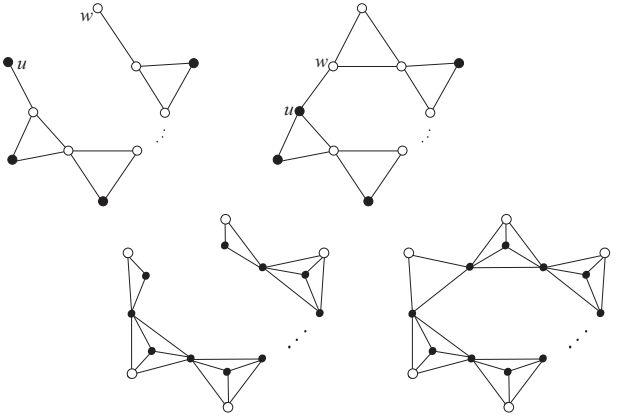


Figure 8. The zero forcing sets for $C_t(K_3)$ (v a cycle vertex and v a noncycle vertex) and $C_t(K_r)$ with $r \geq 4$ (v a cycle vertex and v a noncycle vertex).

vertex v, Z consists of every cycle vertex and all but one noncycle vertex in each K_r or K_{r-1}. □

Theorem 4.4. *For every vertex v,*

$$\mathrm{M}(S_t(K_3) - v) = \mathrm{Z}(S_t(K_3) - v) = t,$$

or equivalently, $\mathrm{n}_v(S_t(K_3)) = \mathrm{z}_v(S_t(K_3)) = 1$.

Proof. First let v be the star vertex of $S_t(K_3)$. Then $S_t(K_3) - v = C_t(K_3)$, so by Theorems 2.4 and 2.7, $\mathrm{n}_v(S_t(K_3)) = \mathrm{z}_v(S_t(K_3)) = 1$. For any vertex v that is not the star vertex, we exhibit a zero forcing set Z for $S_t(K_3) - v$ such that $|Z| = t$, and as in Theorem 4.3 this establishes the theorem. In addition to the star vertex, there are two types of vertex in $S_t(K_3)$, a cycle vertex and a starneighbor vertex. The zero forcing sets Z are illustrated as black vertices in Figure 9.

For a starneighbor vertex v, Z consists of every cycle vertex. For a cycle vertex v, let the two starneighbor vertices adjacent to v in $S_t(K_3)$ be denoted by u and w. Then Z consists of u and every remaining cycle vertex in $S_t(K_3) - v$. □

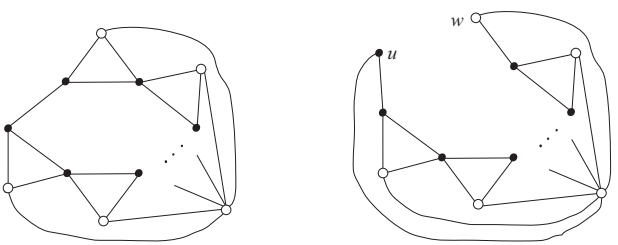


Figure 9. The zero forcing sets for $S_t(K_3) - v$ for v a starneighbor vertex and v a cycle vertex.

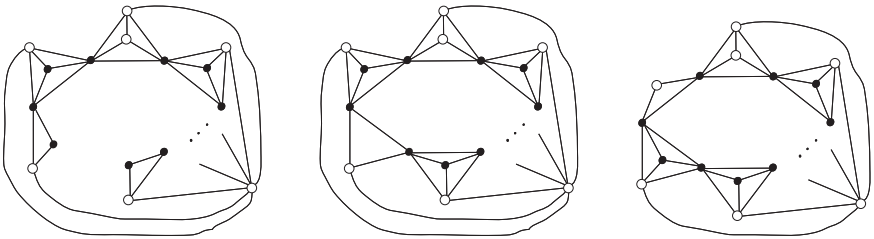


Figure 10. The zero forcing sets for $S_t(K_r)$ for v a cycle vertex, v a standard vertex, and v a starneighbor vertex (with $r \geq 4$).

Theorem 4.5. *Let $r \geq 4$. For every vertex v except the star vertex,*

$$\mathrm{M}(S_t(K_r) - v) = \mathrm{Z}(S_t(K_r) - v) = (r-2)t - 2,$$

or equivalently,

$$\mathrm{n}_v(S_t(K_r)) = \mathrm{z}_v(S_t(K_r)) = 1.$$

If x is the star vertex, then $\mathrm{M}(S_t(K_r) - x) = \mathrm{Z}(S_t(K_r) - x) = (r-2)t$, or equivalently, $\mathrm{n}_x(S_t(K_r)) = \mathrm{z}_x(S_t(K_r)) = -1$.

Proof. First let x be the star vertex of $S_t(K_r)$ with $r \geq 4$. Then $S_t(K_r) - x = C_t(K_r)$, so by Theorems 2.4 and 2.6, $\mathrm{n}_x(S_t(K_r)) = \mathrm{z}_x(S_t(K_r)) = -1$. For any vertex v that is not the star vertex, we exhibit a zero forcing set Z for $S_t(K_r) - v$ of order $(r-2)t - 2$, and as in Theorem 4.3 this establishes the result. The zero forcing sets Z are illustrated as black vertices in Figure 10.

Let v be a cycle vertex, a standard vertex, or a starneighbor vertex, and in $S_t(K_r)$ choose one K_r that does not contain v. Note that each of the one or two copies of K_r in which v was a vertex has now become K_{r-1}. If v is a cycle vertex or a standard vertex, then Z consists of all remaining cycle vertices, all remaining standard vertices in every K_{r-1} or K_r except the chosen K_r, and all but one standard vertices in the chosen K_r. If v is a starneighbor vertex, then Z consists of all cycle vertices, all standard vertices in every K_r except the chosen K_r, and all but one standard vertices in the chosen K_r and in the K_{r-1}. □

4B. ***Edge spreads of complete ciclos and estrellas.*** In analogy with the rank, null, and zero spreads for vertex deletion, spreads for edge deletion were defined in [Edholm et al. 2010]. Let G be a graph and e be an edge in G. The *rank edge spread* of e is $\mathrm{r}_e(G) = \mathrm{mr}(G) - \mathrm{mr}(G-e)$. The *null edge spread* of e is $\mathrm{n}_e(G) = \mathrm{M}(G) - \mathrm{M}(G-e)$. The *zero edge spread* of e is $\mathrm{z}_e(G) = \mathrm{Z}(G) - \mathrm{Z}(G-e)$. Clearly, for any graph G and edge e of G, $r_e(G) + n_e(G) = 0$ [Edholm et al. 2010].

Observation 4.6 [Nylen 1996]. For any graph G and edge e of G, $-1 \leq \mathrm{r}_e(G) \leq 1$ and thus $-1 \leq \mathrm{n}_e(G) \leq 1$.

Theorem 4.7 [Edholm et al. 2010]. *For every graph G and every edge e of G,*

$$-1 \leq \mathrm{z}_e(G) \leq 1.$$

In the same reference it is shown that, although the bounds on the zero edge spread are the same as the bounds on the null edge spread, they are not comparable. As with vertex spread, under certain circumstances we can use one spread to determine the other.

Observation 4.8 [Edholm et al. 2010]. Let G be a graph such that $\mathrm{M}(G) = \mathrm{Z}(G)$ and let e be an edge of G. Then $\mathrm{n}_e(G) \geq \mathrm{z}_e(G)$, and so if $\mathrm{z}_e(G) = 1$, then $\mathrm{n}_e(G) = 1$ (equivalently, $\mathrm{r}_e(G) = 0$).

An edge is classified based on its vertices. For a complete ciclo, there can be three types of edge: cycle-cycle, noncycle-cycle, and noncycle-noncycle (if $r \geq 4$). For a complete estrella there can be six types of edge: cycle-cycle, standard-cycle (if $r \geq 4$), cycle-starneighbor, standard-standard (if $r \geq 5$), standard-starneighbor (if $r \geq 4$), and star-starneighbor.

Theorem 4.9. *For any edge e, $\mathrm{M}(C_t(K_r) - e) = \mathrm{Z}(C_t(K_r) - e) = (r-2)t - 1$, or equivalently, $\mathrm{n}_e(C_t(K_r)) = \mathrm{z}_e(C_t(K_r)) = 1$.*

Proof. We exhibit a zero forcing set Z for $C_t(K_r) - e$ such that $|Z| = (r-2)t - 1$ (here $r \geq 3$). Since $\mathrm{Z}(C_t(K_r)) = (r-2)t$ and $\mathrm{z}_e(C_t(K_r)) \leq 1$, $\mathrm{z}_e(C_t(K_r)) = 1$ and $\mathrm{Z}(C_t(K_r) - e) = (r-2)t - 1$. Since $\mathrm{M}(C_t(K_r)) = \mathrm{Z}(C_t(K_r))$, by Observation 4.8 $\mathrm{n}_e(C_t(K_r)) = 1$, and thus $\mathrm{M}(C_t(K_r) - e) = (r-2)t - 1$. When exhibiting a zero forcing set, we separate $C_t(K_3)$ from $C_t(K_r)$ with $r \geq 4$. The zero forcing sets Z are illustrated as black vertices in Figure 11.

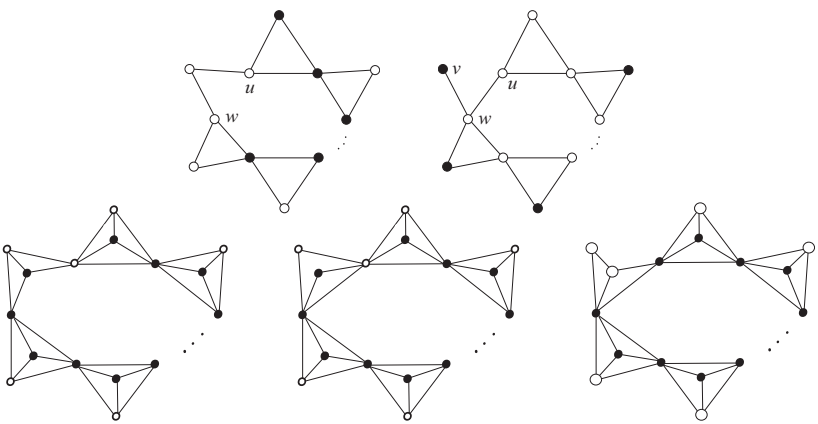


Figure 11. The zero forcing sets for $C_t(K_3) - e$ (e a cycle-cycle edge and e a noncycle-cycle edge) and $C_t(K_r) - e$ with $r \geq 4$ (e a cycle-cycle edge, e a noncycle-noncycle edge, and e a noncycle-cycle edge).

Case $C_t(K_3)$. There are two types of edges e, a cycle-cycle edge and a noncycle-cycle edge. For a cycle-cycle edge $e = \{u, w\}$, Z consists of the noncycle vertex in a K_3 that contains u but not w, and every cycle vertex except u and w. For a noncycle-cycle edge $e = \{v, u\}$, let v be the noncycle vertex of e and let u be the cycle vertex of e. Then Z consists of every noncycle vertex except for the one adjacent to u.

Case $C_t(K_r)$. There are three types of edges: cycle-cycle, noncycle-noncycle, and noncycle-cycle. For e a cycle-cycle edge or noncycle-noncycle edge, Z consists of all cycle vertices except for one of the two cycle vertices in $K_r - e$ and all but one noncycle vertex in each K_r or $K_r - e$; in the case that e is a noncycle-noncycle edge, the noncycle vertex in $K_r - e$ that is not in Z must be an endpoint of e (this is relevant when $r \geq 5$). For a noncycle-cycle edge, Z consists of all the cycle vertices, all but one noncycle vertex in each K_r, and all but two noncycle vertices in the $K_r - e$; one of the two noncycle vertices in $K_r - e$ that is not in Z must be an endpoint of e (this is relevant when $r \geq 5$). □

Theorem 4.10. *For every edge* e, $\mathrm{M}(S_t(K_3) - e) = \mathrm{Z}(S_t(K_3) - e) = t$, *or equivalently*, $\mathrm{n}_e(S_t(K_3)) = \mathrm{z}_e(S_t(K_3)) = 1$.

Proof. We exhibit a zero forcing set Z for $S_t(K_3) - e$ such that $|Z| = t$, and as in Theorem 4.9 this establishes the theorem. The zero forcing sets Z are illustrated as black vertices in Figure 12. There are three types of edges: cycle-cycle, star-starneighbor, and cycle-starneighbor. For a cycle-cycle edge or star-starneighbor edge e, let the two cycle vertices of the K_3 that contains at least one endpoint of e be denoted by u and w. Then Z consists of the starneighbor vertex in the K_3 that contains u but not w, and all cycle vertices except for w. For a cycle-starneighbor edge, Z consists of all cycle vertices. □

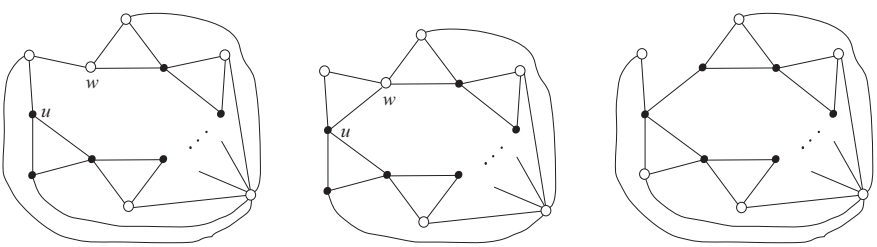


Figure 12. The zero forcing sets for $S_t(K_3) - e$ where e is a cycle-cycle edge, a star-starneighbor edge, and a cycle-starneighbor edge.

Theorem 4.11. *Let* $r \geq 4$. *For every edge* e *except a star-starneighbor edge*,

$$\mathrm{M}(S_t(K_r) - e) = \mathrm{Z}(S_t(K_r) - e) = (r-2)t - 2,$$

or, *equivalently*,

$$\mathrm{n}_v(S_t(K_r)) = \mathrm{z}_v(S_t(K_r)) = 1.$$

If d is a star-starneighbor edge, *then*

$$\mathrm{M}(S_t(K_r) - d) = \mathrm{Z}(S_t(K_r) - d) = (r-2)t - 1,$$

or equivalently,

$$\mathrm{n}_d(S_t(K_r)) = \mathrm{z}_d(S_t(K_r)) = 0.$$

Proof. There can be 6 types of edges: cycle-cycle, standard-cycle, cycle-starneighbor, standard-standard (if $r \geq 5$), standard-starneighbor, and star-starneighbor. For any edge e that is not a star-starneighbor edge, we exhibit a zero forcing set Z for $S_t(K_r) - e$ of order $(r-2)t - 2$, and as in Theorem 4.9 this establishes the result. The zero forcing sets Z are illustrated as black vertices in Figure 13.

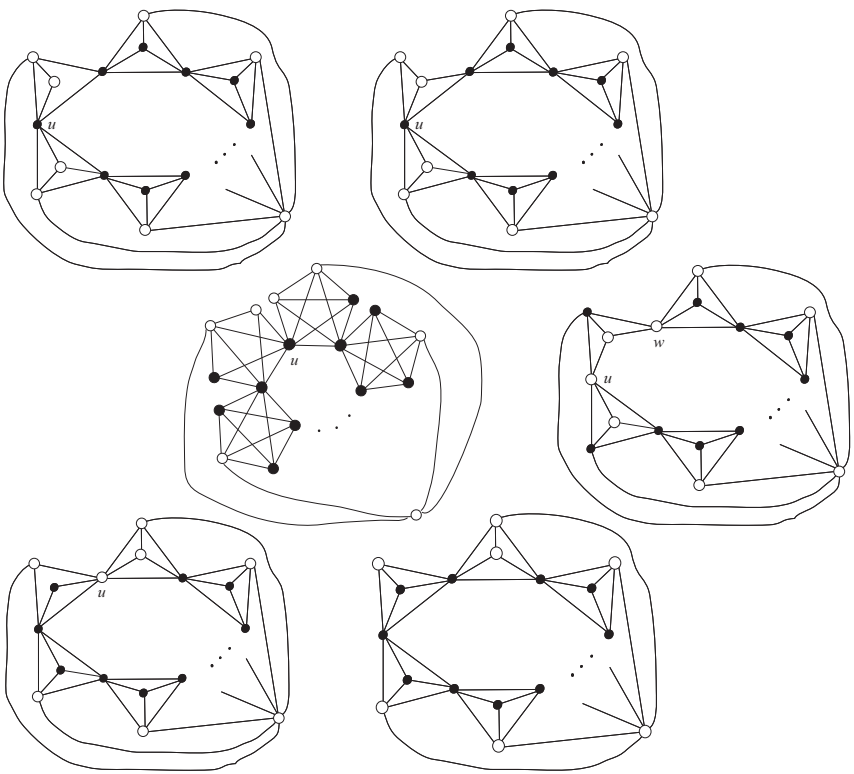


Figure 13. The zero forcing sets for $S_t(K_r) - e$ for e a standard-cycle edge, a cycle-starneighbor edge, a standard-standard edge, a cycle-cycle edge, a standard-starneighbor edge, and a star-starneighbor edge (with $r \geq 4$).

Let e be a standard-cycle edge, a cycle-starneighbor edge, or a standard-standard edge. Let u be a cycle vertex that is not an endpoint of e and is in the $K_r - e$. Then Z consists of all cycle vertices, all standard vertices in each K_r (or $K_r - e$) except those that contain u, and all but one of the standard vertices in the K_r and $K_r - e$ that contain u. In the case that e is a standard-standard edge, the standard vertex in $K_r - e$ that is not in Z must be an endpoint of e (this is relevant when $r \geq 6$).

For a cycle-cycle edge $e = \{w, u\}$, Z consists of all cycle vertices except w and u, all standard vertices in each K_r (or $K_r - e$) except those that contain u, all but one of the standard vertices in the K_r and $K_r - e$ that contain u, and the starneighbor vertex in the K_r and $K_r - e$ that contain u.

For a standard-starneighbor edge, choose one cycle vertex u in the $K_r - e$. Then Z consists of all cycle vertices except for u, all standard vertices in each K_r except the K_r that contains u, all standard vertices in $K_r - e$, and all but one of the standard vertices in the one K_r that contains u.

For a star-starneighbor edge d, let Z consist of all cycle vertices and all standard vertices except one standard vertex in a K_r that does not contain an endpoint of d. Then Z is a zero forcing set for $S_t(K_r) - d$. Since $S_t(K_r) - d$ can be covered by t copies of K_r and one $K_{1,t-1}$, we have $\mathrm{mr}(S_t(K_r) - d) \leq t + 2$. Thus

$$\begin{aligned}(r-2)t - 1 = |S_t(K_r) - d| - (t+2) &\leq \mathrm{M}(S_t(K_r) - d) \\ &\leq \mathrm{Z}(S_t(K_r) - d) \leq (r-2)t - 1,\end{aligned}$$

and we have equality throughout. □

References

[AIM 2008] AIM Minimum Rank – Special Graphs Work Group (F. Barioli, W. Barrett, S. Butler, S. M. Cioabă, D. Cvetković, S. M. Fallat, C. Godsil, W. Haemers, L. Hogben, R. Mikkelson, S. Narayan, O. Pryporova, I. Sciriha, W. So, D. Stevanović, H. van der Holst, K. Vander Meulen, A. W. Wehe), "Zero forcing sets and the minimum rank of graphs", *Linear Algebra Appl.* **428**:7 (2008), 1628–1648. MR 2008m:05166 Zbl 1135.05035

[Barioli et al. 2004] F. Barioli, S. Fallat, and L. Hogben, "Computation of minimal rank and path cover number for certain graphs", *Linear Algebra Appl.* **392** (2004), 289–303. MR 2005i:05115 Zbl 1052.05045

[Barioli et al. 2009] F. Barioli, S. M. Fallat, H. T. Hall, D. Hershkowitz, L. Hogben, H. van der Holst, and B. Shader, "On the minimum rank of not necessarily symmetric matrices: a preliminary study", *Electron. J. Linear Algebra* **18** (2009), 126–145. MR 2010e:05176 Zbl 1169.05345

[Barioli et al. 2010] F. Barioli, W. Barrett, S. M. Fallat, H. T. Hall, L. Hogben, B. Shader, P. van den Driessche, and H. van der Holst, "Zero forcing parameters and minimum rank problems", *Linear Algebra Appl.* **433**:2 (2010), 401–411. MR 2645093 Zbl 05719624

[Barrett et al. 2008] W. Barrett, H. T. Hall, H. van der Holst, and J. Sinkovic, "The minimum rank problem for rectangular grids", lecture by Barrett at Rocky Mountain Discrete Mathematics Days (Laramie, WY), 2008.

[DeLoss et al. 2008] L. DeLoss, J. Grout, T. McKay, J. Smith, and G. Tims, "Program for calculating bounds on the minimum rank of a graph using Sage", 2008. arXiv:0812.1616. A faster zero forcing number program, also in Sage, is included in the Minimum Rank Library by S. Butler, L. DeLoss, J. Grout, H, T, Hall, J, LaGrange, T. McKay, J. Smith, and G. Tims (2010), which is available at http://sage.cs.drake.edu/home/pub/67. For more information contact Jason Grout at jason.grout@drake.edu.

[Edholm et al. 2010] C. J. Edholm, L. Hogben, M. Hyunh, J. LaGrange, and D. D. Row, "Vertex and edge spread of zero forcing number, maximum nullity, and minimum rank of a graph", *Linear Algebra Appl.* (2010).

[Fallat and Hogben 2007] S. M. Fallat and L. Hogben, "The minimum rank of symmetric matrices described by a graph: a survey", *Linear Algebra Appl.* **426**:2-3 (2007), 558–582. MR 2008f:05114 Zbl 1122.05057

[Han and Neumann 2007] L. Han and M. Neumann, "Inner product spaces, orthogonal projection, least squares, and singular value decomposition", pp. 5–1 to 5–16 in *Handbook of Linear Algebra*, edited by L. Hogben, CRC Press, Boca Raton, FL, 2007.

[Hogben 2010] L. Hogben, "Minimum rank problems", *Linear Algebra Appl.* **432**:8 (2010), 1961–1974. MR 2599835 Zbl 05677360

[Huang et al. 2010] L.-H. Huang, G. J. Chang, and H.-G. Yeh, "On minimum rank and zero forcing sets of a graph", *Linear Algebra Appl.* **432**:11 (2010), 2961–2973. MR 2639259 Zbl 1195.05043

[Johnson and Leal Duarte 1999] C. R. Johnson and A. Leal Duarte, "The maximum multiplicity of an eigenvalue in a matrix whose graph is a tree: invariant factors", *Linear and Multilinear Algebra* **46**:1-2 (1999), 139–144. MR 2000e:05114 Zbl 0929.15005

[Nylen 1996] P. M. Nylen, "Minimum-rank matrices with prescribed graph", *Linear Algebra Appl.* **248** (1996), 303–316. MR 97k:15053 Zbl 0864.05069

[Sinkovic 2010] J. Sinkovic, "Maximum nullity of outerplanar graphs and the path cover number", *Linear Algebra Appl.* **432**:8 (2010), 2052–2060. MR 2599841 Zbl 05677366

Received: 2010-05-28 Revised: 2010-10-09 Accepted: 2010-10-10

edgardalmodovar@gmail.com *Department of Mathematics, University of Puerto Rico, Río Piedras Campus, San Juan, PR 00931, United States*

ldeloss@gmail.com *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

LHogben@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

American Institute of Mathematics, 360 Portage Ave, Palo Alto, CA 94306

kahogenson@gmail.com *Department of Mathematics, University of North Dakota, Grand Forks, ND 58202, United States*

kaitlynemurphy@gmail.com *Montclair State University, College of Science and Mathematics, Montclair, NJ 07043, United States*

tpeters@iastate.edu *Department of Mathematics, Iowa State University, Ames, IA 50011, United States*

camila.ale.ramirez@gmail.com *Department of Mathematics, University of Puerto Rico, Río Piedras Campus, San Juan, PR 00931, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.393

A numerical investigation on the asymptotic behavior of discrete Volterra equations with two delays

Immacolata Garzilli, Eleonora Messina and Antonia Vecchio

(Communicated by Kenneth S. Berenhaut)

We describe a numerical approach to the solution of two-delay Volterra integral equations, and we carry out a nonlinear stability analysis on an interesting test equation by means of a parallel investigation both on the continuous and the discrete problem.

1. Introduction

Messina et al. [2008a] present a comparison between the analytical and the numerical solution of the following Volterra integral equation (VIE) with two constant delays:

$$y(t) = \int_{t-\tau_2}^{t-\tau_1} k(t-\tau)g(y(\tau))d\tau \quad t \in [\tau_2, T], \tag{1}$$

with $y(t) = \varphi(t)$, $t \in [0, \tau_2]$, where $\varphi(t)$ is a known function such that

$$\varphi(\tau_2) = \int_0^{\tau_2-\tau_1} k(\tau_2-\tau)g(\varphi(\tau))d\tau. \tag{2}$$

The interest of (1) in the applications is mainly in the modeling of age-structured population dynamics, as described in [Messina et al. 2008a] and the references therein. Here, we continue those investigations with the aim of providing a more complete analysis of the dynamics of the solutions. In particular, we add some new results on the global asymptotic behavior of solutions and simplify some already known proofs. In Section 2, the properties of the continuous solution are summarized and a new result on global asymptotic stability of the nontrivial equilibrium is proved. In Section 3, we consider a numerical method of direct quadrature type and look for conditions on the step size h of a direct quadrature method that lead to a numerical solution which mimics the behavior of the continuous one. The

MSC2000: 45M05, 45M10, 65R20.
Keywords: Volterra integral equations, direct quadrature methods, stability, double delays.

main novelty of this paper with respect to [Messina et al. 2008a] is the compact form that we use to represent the method: this new form allows us to obtain some new results in the discrete case equivalent to those valid for the continuous case, and so to complete the parallelism between the behaviors of the analytical and the numerical solution. Finally, in Section 4 we report some numerical examples that show the nature of these behaviors.

2. The continuous equation

In this section we provide a summary of the theory related to the stability of equilibria of (1) already developed in [Messina et al. 2008a] and we prove a new result on the global asymptotic stability of the nontrivial equilibrium (Theorem 2.6).

As in that paper, we make certain assumptions on the functions φ, g and k of problem (1):

(a) $\varphi(t) \geq 0$, for all $t \in [0, \tau_2]$;

(b) $k(t)$ not identically zero and $k(t) \geq 0$, for all $t \in [\tau_1, \tau_2]$;

(c) $g \in C^1([0, +\infty))$, $g(x) \geq 0$, for all $x \geq 0$ and $g(0) = 0$, $g'(0) > 0$;

(d) $g(x) - xg'(x) \geq 0$, for all $x \geq 0$;

(e) $1/g'(0) \leq x/g(x)$, for all $x > 0$.

These assumptions include some that are significant from a biological point of view (see [Messina et al. 2008a] and the bibliography therein) and guarantee that the solution $y(t)$ is nonnegative for all $t \geq \tau_2$. Define the positive function

$$a(x) = \begin{cases} x/g(x) & \text{if } x > 0, \\ 1/g'(0) & \text{if } x = 0. \end{cases}$$

By hypotheses (d) and (e), $a(x)$ is an increasing function for all $x \geq 0$. In particular, it is strictly increasing for all $x \geq 0$, if $g(x)$ is a nonlinear function, while it is constant otherwise. From now on, we assume that $g(x)$ is nonlinear, hence (d) and (e) are meant as strict inequalities and, in analogy with [Messina et al. 2008a], we consider the following alternative formulation of (1):

$$y(t) = \rho g(y(t - \xi(t))), \quad \xi(t) \in [\tau_1, \tau_2], \tag{3}$$

where

$$\rho = \int_{\tau_1}^{\tau_2} k(x)\, dx, \tag{4}$$

which is more appropriate for our analysis. Obviously, (1) has at least the trivial solution $y^* = 0$. The following theorem shows that this equilibrium is unique for $\rho_0 < 1/g'(0)$, then the value $\rho_0 = 1/g'(0)$ represents a bifurcation point for the

variable ρ; as a matter of fact, when $\rho > \rho_0$, the trivial solution is no longer unique, and another nontrivial equilibrium $y^* = a^{-1}(\rho)$ appears. Let $a^* = \lim_{x\to+\infty} a(x)$.

Theorem 2.1 [Iannelli 1994; Messina et al. 2008a]. *Let ρ be defined as in (4).*

(i) *Equation (1) has one and only one nontrivial equilibrium $y^* = a^{-1}(\rho)$ if and only if $1/g'(0) < \rho < a^*$.*

(ii) *Equation (1) has only the trivial equilibrium if $\rho \le 1/g'(0)$.*

To analyze the nature of these equilibria we recall the following definitions.

Definition 2.1. Let y^* be an equilibrium point for (1). Then y^* is said to be:

- stable if, for all $\epsilon > 0$, there exists $\delta > 0$ such that
$$|\varphi(t) - y^*| < \delta,\ \forall t \ge \tau_2 \implies |y(t) - y^*| < \epsilon,\ \forall t \in [\tau_2, T];$$
- locally attractive if there exists $\delta > 0$ such that
$$|\varphi(t) - y^*| < \delta,\ \forall t \in [0, \tau_2] \implies \lim_{t\to+\infty} |y(t) - y^*| = 0;$$
- globally attractive if, for all $\varphi(t) > 0$,
$$\lim_{t\to+\infty} |y(t) - y^*| = 0;$$
- locally asymptotically stable if it is stable and locally attractive;
- globally asymptotically stable if it is stable and globally attractive.

We now quote some propositions proved in earlier papers, and we prove Theorem 2.6, which assures the global asymptotic stability of the solution $y(t)$ of (1).

Theorem 2.2 [Iannelli 1994; Messina et al. 2010]. *Let y^* be an equilibrium point for (1).*

(i) *If $\rho|g'(y^*)| < 1$, then y^* is locally asymptotically stable;*

(ii) *If $\rho|g'(y^*)| > 1$, then y^* is unstable.*

Theorem 2.3 [Messina et al. 2008a]. *If $g(x)$ is nondecreasing, then the nontrivial equilibrium y^* is locally asymptotically stable.*

Theorem 2.4 [Iannelli 1994]. *If $\rho g'(0) < 1$, then the trivial equilibrium is globally asymptotically stable.*

We recall that, from the biological point of view, the threshold value $\rho g'(0)$ plays the role of the *basic reproduction number.*[1] Furthermore, while it is known

[1] In population dynamics, the basic reproduction number represents the average number of offspring produced over the lifetime of an individual under ideal conditions. In epidemiological models, it represents the mean number of secondary cases that a single infected case causes in a population with no immunity.

in [Messina et al. 2008a] that the global attractivity of $y^* = 0$ implies $\rho \le 1/g'(0)$, a result on the behavior of $y^* = 0$, when $\rho = 1/g'(0)$ is still missing.

Since in many examples of applications the form of the nonlinearity in (1) is of unimodal type (e.g., $g(x) = xe^{-x}$; see for instance [Breda et al. 2007, Section 6; Iannelli 1994, page 81 (5.19)], where, as we explain in the introduction of [Messina et al. 2008a], $\Phi(x) = (g(x))/x)$, we assume, from now on, that $g(x)$ is an unimodal function with mode $\bar{y}$.

Theorem 2.5 [Messina et al. 2008a]. *Let $g(x)$ in (1) be unimodal and let*

$$\frac{1}{g'(0)} \le \rho \le \frac{\bar{y}}{g(\bar{y})}.$$

Then

$$\lim_{t\to+\infty} y(t) = a^{-1}(\rho), \quad \text{for all } \varphi(t) \ge 0.$$

Thanks to these results we can prove the following theorem on the global asymptotic stability of the nontrivial equilibrium.

Theorem 2.6. *Let $g(x)$ in (1) be an unimodal function with mode $\bar{y}$. If*

$$\frac{1}{g'(0)} \le \rho \le \frac{\bar{y}}{g(\bar{y})},$$

then y^* is globally asymptotically stable.

Proof. If $y(t)$ is a solution of (1), then

$$y(t) = \rho g(y(t - \xi(t))) \le \rho g(\bar{y}) \le \bar{y}.$$

This means that each $y(t)$ which is a solution of (1) falls in the interval $[0, \bar{y}]$ where $g(y)$ is increasing; in particular $g'(y^*) > 0$. Since also ρ is positive, then $\rho|g'(y^*)| = \rho g'(y^*)$. What is more, thanks to hypothesis (d), $g(y^*) - y^* g'(y^*) > 0$ and thus $\rho g'(y^*) < 1$ (this last inequality holds since $\rho = y^*/g(y^*)$). Hence, we are in the hypotheses of Theorem 2.2 and so y^* is locally asymptotically stable. Since $g(x)$ is an unimodal function, we are in the hypotheses of Theorem 2.5. Hence, $y^* = a^{-1}(\rho)$ is a globally attractive equilibrium. □

The hypothesis $\rho \le (\bar{y})/g(\bar{y})$ plays a crucial role in the proof because it implies that each $y(t)$ which is a solution of (1) falls in the interval where $g(x)$ in increasing. As a consequence, the previous results on unimodal functions can be extended to increasing functions $g(x)$. In particular, the following theorem holds.

Theorem 2.7. *Let $g(x)$ in (1) be an increasing function. If $\rho \ge 1/g'(0)$, then y^* is globally asymptotically stable.*

Theorem 2.8 [Messina et al. 2008a]. *Let $g(x)$ in (1) be unimodal with mode $\bar{y}$. Assume $\rho > \bar{y}/g(\bar{y})$ and let $k'(x)$ be constant in sign for all $x \in [\tau_1, \tau_2]$. Then the nonequilibrium solutions of (1) cannot be definitively monotone.*

3. The discrete equation

Let a partition of the interval $[0, T]$ be given by

$$\Pi_N = \{t_n : 0 = t_0 < t_1 < \cdots < t_N = T\},$$

where $t_{n+1} - t_n = h$, $n = 0, \ldots, N$, for some fixed h, called the step size. Assume

$$h = \frac{\tau_1}{r_1} = \frac{\tau_2}{r_2}, \tag{5}$$

with r_1, r_2 positive integers. In [Messina et al. 2009] the following direct quadrature method [Brunner and van der Houwen 1986; Linz 1985], adapted to the form of (1), is proposed:

$$y_n = h \sum_{j=r_1}^{r_2} w_j k(jh) g(y_{n-j}), \quad n > r_2, \tag{6}$$

where $y_n \simeq y(t_n)$ and $y_l = \varphi(lh)$, $l = 0, 1, \ldots, r_2$, for $\varphi(t)$ is a known function satisfying condition (2). In [Messina et al. 2008a; 2008b; 2009] some conditions on the step size h were derived for which the numerical solution mimics the behavior of the continuous one. Now, with the help of a new reformulation of (6) we are able to complete such analysis by deriving the discrete version of Theorems 2.2 and 2.3 (Theorems 3.3 and 3.4 respectively) and a new result on the global asymptotic stability of the nontrivial equilibrium (Theorem 3.8).

In order to write (6) as the discrete analogous of (3), we will make use of the discrete mean value theorem that we report and prove here for the sake of completeness.

Theorem 3.1. *Assume $f \in C([a, b])$, with $-\infty < a < b < \infty$ and let $x_1, \ldots, x_n \in [a, b]$. If $\alpha_1, \ldots, \alpha_n$ are n real numbers, all of the same sign, there exists $\xi \in (a, b)$ such that*

$$\sum_{i=1}^{n} \alpha_i f(x_i) = f(\xi) \sum_{i=1}^{n} \alpha_i.$$

Proof. Let $m = \min_{x\in[a,b]} f(x)$ and $M = \max_{x\in[a,b]} f(x)$ and assume $\alpha_j \geq 0$, for all $j = 1, \ldots, n$. Then,

$$m \sum_{j=1}^{n} \alpha_j \leq \sum_{j=1}^{n} \alpha_j f(x_j) \leq M \sum_{j=1}^{n} \alpha_j$$

and hence,

$$m \leq \frac{\sum_{j=1}^{n} \alpha_j f(x_j)}{\sum_{j=1}^{n} \alpha_j} \leq M.$$

Since $f(x)$ takes on all values between m and M (*intermediate value theorem*), there exists a point $\xi \in (a, b)$ such that

$$f(\xi) = \frac{\sum_{j=1}^{n} \alpha_j f(x_j)}{\sum_{j=1}^{n} \alpha_j}. \qquad \square$$

Now, define the quantity

$$\rho_h = h \sum_{j=r_1}^{r_2} w_j k(jh). \tag{7}$$

Observe that $k(jh)w_j$ is constant in sign for all $j = r_1, \ldots, r_2$. By Theorem 3.1, then, there exists $\xi_n \in \left[\min_{n-r_2 \le j \le n-r_1} y_j, \max_{n-r_2 \le j \le n-r_1} y_j \right]$ such that

$$y_n = hg(\xi_n) \sum_{j=r_1}^{r_2} w_j k(jh).$$

Thus, (6) can be formulated, in analogy with the continuous case, in the form

$$y_n = \rho_h g(\xi_n), \text{ with } \xi_n \in \Big[\min_{n-r_2 \le j \le n-r_1} y_j, \max_{n-r_2 \le j \le n-r_1} y_j \Big]. \tag{8}$$

As for the continuous case, hypotheses (a), (b) and (c) and the positiveness of weights w_j guarantee that the discrete solution y_n is nonnegative for all $n \ge 0$. With regard to the existence of equilibrium solutions, we have:

Theorem 3.2 [Messina et al. 2008a]. *Let ρ_h be defined by (7).*

(i) *Equation (6) has one and only one nontrivial equilibrium $y^*(h) = a^{-1}(\rho_h)$ if and only if $1/g'(0) < \rho_h < a^*$.*

(ii) *Equation (6) has only the trivial equilibrium if $\rho_h \le 1/g'(0)$.*

Now we can prove the following results.

Theorem 3.3. *Let $y^*(h)$ be an equilibrium point for (6).*

(i) *If $\rho_h |g'(y^*(h))| < 1$, then $y^*(h)$ is locally asymptotically stable.*

(ii) *If $\rho_h |g'(y^*(h))| > 1$, then $y^*(h)$ is unstable.*

Proof. (1) Suppose $\rho_h |g'(y^*(h))| < 1$. To show that $y^*(h)$ is stable, we fix $\epsilon > 0$ and consider φ such that $|\varphi_j - y^*(h)| < \delta_\epsilon$, $j = 0, \ldots, r_2$, for some $\delta_\epsilon > 0$. Let n take values in $\{r_2, \ldots, r_2 + r_1\}$. From (8) we have $y_n = \rho_h g(\xi_n)$, with

$$\xi_n \in \Big[\min_{j=0\ldots,r_1} \varphi_j, \max_{j=0\ldots,r_1} \varphi_j \Big];$$

hence, $|\xi_n - y^*(h)| < \delta_\epsilon$. For the difference $y_n - y^*(h)$, we have

$$y_n - y^*(h) = \rho_h (g(\xi_n) - g(y^*)) = \rho_h g'(\theta)(\xi_n - y^*(h)),$$

where $\theta \in [\min\{\xi_n, y^*(h)\}, \max\{\xi_n, y^*(h)\}]$; for this reason, $|\theta - y^*(h)| < \delta_\epsilon$. Moreover, since $g'(x)$ is a continuous function such that $|\rho_h g'(y^*(h))| < 1$, there exists $\tilde{y}$ such that

$$|\rho_h g'(y)| < 1 \quad \text{for } y \in [y^*(h) - \tilde{y}, y^*(h) + \tilde{y}].$$

Thus, if we choose $\delta_\epsilon = \min\{\epsilon, \tilde{y}\}$, then $|y_n - y^*(h)| < \epsilon$, $n = r_2, \ldots, r_2 + r_1$. Using this, we easily prove that $|y_n - y^*(h)| < \epsilon$ also for $n = r_2 + r_1, \ldots, r_2 + 2r_1$, and, in general for all $n \geq r_2$. Thus, stability is proved.

Local attractivity follows straightforwardly, by observing that there exists $\delta > 0$ such that $\rho_h |g'(y)| \leq p < 1$, for all $y \in [y^*(h) - \delta, y^*(h) + \delta]$. Thus, by choosing

$$\varphi_1, \ldots, \varphi_{r_2} \in [y^*(h) - \delta, y^*(h) + \delta],$$

and proceeding step by step as n grows, we see that in the k-th interval

$$|y_n - y^*(h)| \leq p^k \delta, \tag{9}$$

where $k \to +\infty$ for $n \to +\infty$. Therefore, $\lim_{n \to +\infty} y_n = y_h^*$.

(2) Consider $\rho_h |g'(y^*(h))| > 1$. To prove the instability of $y^*(h)$ we must find ϵ_0 such that

$$\forall \delta > 0, \ \exists n \in \{0, \ldots, r_2\} \ : \ |y_n - y^*| < \delta,$$

and

$$\exists \bar{n} > r_2 \ : \ |y_{\bar{n}} - y^*(h)| > \epsilon_0.$$

By the continuity of the function g', there exists $d > 0$ such that

$$|g'(y)| \geq r > 1, \ \forall y \in [y^*(h) - d, y^*(h) + d].$$

Take $n \in \{r_2, \ldots, r_2 + r_1\}$. In view of (8) there results

$$|y_n - y^*(h)| = \rho_h |g(\xi_n) - g(y^*(h))| = \rho_h |g'(z)| |\xi_n - y^*(h)|, \tag{10}$$

with $|z - y^*(h)| \leq |\xi_n - y^*(h)|$. Then, for all $\delta > 0$, it is possible to choose the starting values φ_l different from $y^*(h)$, for all $l = 0, \ldots, r_2$ and such that $|\varphi_l - y^*(h)| < \min\{d, \delta\}$, for all $n = 0, \ldots, r_2$. Thus,

$$|\xi_n - y^*(h)| < d, \quad n = r_2, \ldots, r_2 + r_1$$

and, form (10), $|\bar{y} - y^*(h)| < d$. This implies that $\rho_h |g'(z)| > 1$. Hence, choosing $\epsilon_0 = \min_{n \in [0, r_2]} |y_n - y^*(h)|$, we have

$$|y_n - y^*(h)| > |\xi_n - y^*(h)|, \ \forall n \in \{r_2, \ldots, r_2 + r_1\}. \qquad \square$$

Now we prove the discrete counterpart of Theorem 2.3.

Theorem 3.4. *If $g(x)$ is a nondecreasing function, then the nontrivial equilibrium $y^*(h)$ is locally asymptotically stable.*

Proof. Let $y^*(h) \neq 0$, for hypothesis (d), $g(x) - xg'(x) > 0$, for all $x > 0$, then $g(y^*(h)) - y^*(h)g'(y^*(h)) > 0$; since $y^*(h) = \rho_h g(y^*(h))$, we have

$$\rho_h g'(y^*(h)) < 1,$$

that is, $\rho_h |g'(y^*(h))| < 1$, since $g'(x) \geq 0$. The result follows from Theorem 3.3. □

The following theorem was proved in [Messina et al. 2008a], but here the proof has been simplified by the new formulation (8) of (6).

Theorem 3.5. *If $\rho_h g'(0) < 1$, then the trivial equilibrium is globally asymptotically stable.*

Proof. We already know from Theorem 3.3 that $y^*(h) = 0$ is locally asymptotically stable. Now we prove the global attractivity. Let $r_2 \leq n \leq r_2 + r_1$ then, from (8),

$$y_n = \rho_h[g(\xi_n) - g(0)] = \rho_h g'(\xi_{n_0})\xi_n, \tag{11}$$

with $0 \leq \xi_{n_0} \leq \xi_n$ and $\xi_n \in [0, \max_{0\leq j\leq r_2} \varphi_j]$. Thanks to hypotheses (d) and (e),

$$g'(\xi_{n_0}) < \frac{g(\xi_{n_0})}{\xi_{n_0}} < g'(0). \tag{12}$$

From (11) and (12) we obtain $y_n < \rho_h g'(0)\xi_n \leq \rho_h g'(0)\phi$, where $\phi = \max_{0\leq j\leq r_2} \varphi_j$. Let $\alpha = \rho_h g'(0)$. Then $y_n \leq \alpha\phi$, with $\alpha < 1$.

By similar arguments, for $n = r_2 + r_1 \ldots r_2 + 2r_1$, we get $y_n < \alpha^2\phi$, with $\alpha < 1$. The conclusion follows by iterating the same procedure in all the next intervals. □

In analogy with the continuous case we report the following result:

Theorem 3.6 [Messina et al. 2008a]. *If $y^*(h) = 0$ is globally attractive, then*

$$\rho_h \leq \frac{1}{g'(0)}.$$

Next we consider the special case where the function $g(x)$ is unimodal.

Theorem 3.7 [Messina et al. 2008a]. *Assume that $g(x)$ in (1) is unimodal with mode $\bar{y}$. It $1/g'(0) \leq \rho_h \leq \bar{y}/g(\bar{y})$, then*

$$\lim_{n\to+\infty} y_n = a^{-1}(\rho_h).$$

Now, we can prove the discrete version of Theorem 2.6.

Theorem 3.8. *Assume that $g(x)$ in (1) is an unimodal function with mode $\bar{y}$. If $1/g'(0) \leq \rho_h \leq \bar{y}/g(\bar{y})$, then $y^*(h)$ is globally asymptotically stable.*

Proof. If y_n is a solution of (6), then

$$y_n = \rho_h g(\xi_n) \le \rho_h g(\bar{y}) \le \bar{y}.$$

This means that each y_n which is a solution of (6) falls in the interval $[0, \bar{y}]$, where $g(y)$ is increasing; in particular $g'(y^*(h)) > 0$. Since ρ_h is positive as well, we have $\rho_h |g'(y^*(h))| = \rho_h g'(y^*(h))$. What is more, thanks to hypothesis (d), $g(y^*(h)) - y^*(h)g'(y^*(h)) > 0$ and thus $\rho_h g'(y^*(h)) < 1$ (this last inequality holds since $\rho_h = (y^*(h))/(g(y^*(h)))$). Hence, for Theorem 3.3, $y^*(h)$ is locally asymptotically stable. Furthermore, $y^* = a^{-1}(\rho)$ is globally attractive, because, since $g(x)$ is unimodal, we can apply the result in Theorem 3.7. So it is a globally asymptotically stable equilibrium. □

In analogy with the continuous case the following result holds.

Theorem 3.9. *Assume that $g(x)$ is an increasing function. If*

$$\rho_h \ge \frac{1}{g'(0)},$$

then $y^*(h)$ is globally asymptotically stable.

Now, we report a known result that characterize the behavior of the solutions of (6) when the parameter ρ_h is greater than the threshold value $\bar{y}/g(\bar{y})$.

Theorem 3.10 [Messina et al. 2008a]. *Assume $g(x)$ in (6) is unimodal with mode $\bar{y}$ and let $\rho_h > \bar{y}/g(\bar{y})$. Then the nonequilibrium solutions of (6) cannot be definitively monotone.*

4. A case study

All the previous analysis is well illustrated by means of the following problem of the type (1):

$$y(t) = 8R_0 \int_{t-\tau_2}^{t-\tau_1} \left(1 - \frac{1}{\tau_2}(t-\tau)\right) e^{-y(\tau)} y(\tau)\, d\tau, \quad t \in [\tau_2, T]. \tag{13}$$

This equation was considered in [Messina et al. 2008a], where the analytical properties of the solutions were listed and some plots of the numerical solution with respect to time were reported. Here, we summarize the results in that paper and show new ones using a different approach — namely, a comparison between the bifurcation diagrams of the continuous and numerical solutions and plots of the orbits of the numerical solution. Experiments of this kind are quite common for describing the dynamics of population problems.

In (13), $y(t)$ represents the number of adults in the population at time t, while τ_1 is the maturation age, τ_2 the maximum age, and R_0 the basic reproduction number. This equation represents an interesting case study because it includes the major

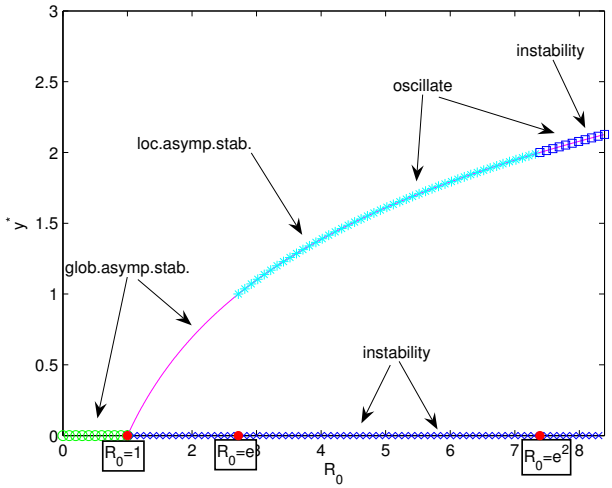


Figure 1. Bifurcation diagram for the parameter R_0 of (13).

features of more complicated models. In particular, $k(t-s)$ is positive and $g(x) = xe^{-x}$ is unimodal with mode $\bar{y} = 1$. If we choose $\tau_1 = 1/2$ and $\tau_2 = 1$, then the parameter $\rho = R_0$ and the nontrivial equilibrium is $y^* = \ln R_0$.

We underline that (13) corresponds to the partial derivative equation described in [Breda et al. 2007, Section 6], modeling a juveniles-adult dynamic.

What makes this equation simple with respect to other problems is that the two classes in which the population is divided (adult $y(t)$ and juveniles $x(t)$) are described by uncoupled equations. More precisely, the number of juveniles $x(t)$ is described by the following integral

$$x(t) = 8R_0 \int_{t-\tau_1}^{t} \Big(1 - \frac{1}{\tau_2}(t-\tau)\Big) e^{-y(\tau)} y(\tau)\, d\tau, \quad t \in [\tau_2, T]. \tag{14}$$

Hence, the complete problem is represented by Equations (13)+(14), where $y(t)$ depends only on itself and $x(t)$ is a function of $y(t)$.

From the diagram in Figure 1, it is clear that, if $R_0 < 1$ only the trivial equilibrium exists and it is globally asymptotically stable, after this threshold value it becomes unstable; as usual we don't know what happens when $R_0 = 1$. At $R_0 = 1$ the solution bifurcates giving rise to a new nontrivial equilibrium which is globally asymptotically stable for all values of $R_0 \le e = \bar{y}/g(\bar{y})$. When $e < R_0 \le e^2 = 1/g'(\rho)$ the solution oscillates and then converges to $y^* = \ln R_0$, while for $R_0 > e^2$ the equilibrium becomes unstable.

In Figure 2 we report the bifurcation diagram related to the numerical solution of the problem described in (13). From the figure it is clear that the dynamics of the continuous and discrete solutions coincide. In particular, the threshold values 1, e, e^2 are the same. What makes the difference is that the dynamic of $y(t)$ is

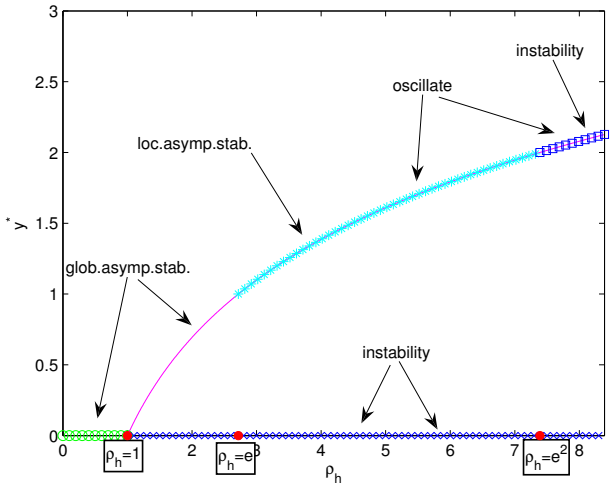


Figure 2. Bifurcation diagram for the parameter ρ_h.

described by the parameter ρ given in (4) (that in our case corresponds to R_0), while the one of y_n by the parameter ρ_h given in (7). However, since $\rho_h \to \rho$ it is always possible to find a sufficiently small step size h such that the two solutions show the same asymptotic behavior.

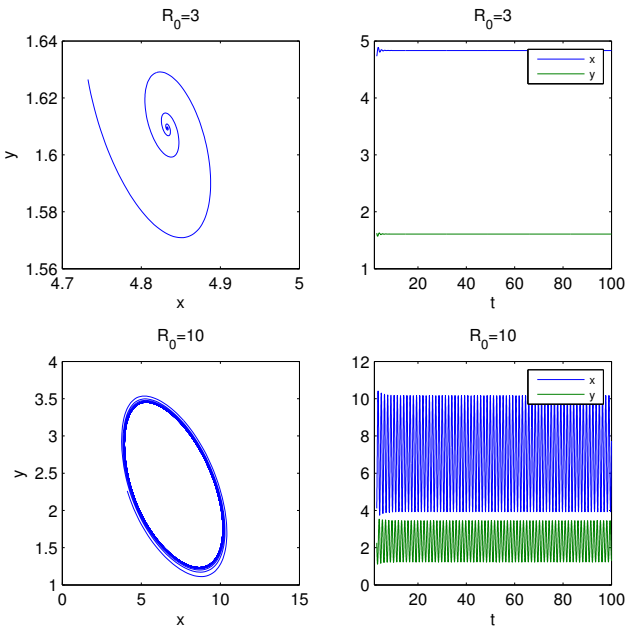


Figure 3. Orbits obtained by numerical computation of the solution of (13)+(14) and corresponding time-dependent plots.

To complete our analysis of problem (13)+(14), we report, in Figure 3, some numerical simulations that show the dynamics of the complete system (13)+(14) for $\rho > 1/g'(0)$ (that is, $R_0 > 1$). In this case, there exists a unique nontrivial equilibrium $P^* = (x^*, y^*) = (\ln R_0, 3 \ln R_0)$. In the first column of the figure, the orbits of the numerical solution clearly show that, in accordance with our investigations, for $1 < R_0 < e^2$ ($R_0 = 5$ in the plot), P^* is a stable equilibrium, while for $R_0 > e^2$, P^* becomes unstable. The two plots reported in the second column show the corresponding time-dependent behaviors, where it is evident that the solution tends to the equilibrium for $R_0 < e^2$ and presents nonstable oscillations after that.

References

[Breda et al. 2007] D. Breda, C. Cusulin, M. Iannelli, S. Maset, and R. Vermiglio, "Stability analysis of age-structured population equations by pseudospectral differencing methods", *J. Math. Biol.* **54**:5 (2007), 701–720. MR 2295748 (2007m:92052)

[Brunner and van der Houwen 1986] H. Brunner and P. J. van der Houwen, *The numerical solution of Volterra equations*, vol. 3, CWI Monographs, North-Holland Publishing Co., Amsterdam, 1986. MR 871871 (88g:65136

[Iannelli 1994] M. Iannelli, *Mathematical theory of age-structured population dynamics*, Applied Mathematics Monographs (C.N.R), Giardini Editori e Stampatori, Pisa, Italy, 1994.

[Linz 1985] P. Linz, *Analytical and numerical methods for Volterra equations*, vol. 7, SIAM Studies in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1985. MR 796318 (86m:65163

[Messina et al. 2008a] E. Messina, E. Russo, and A. Vecchio, "Comparing analytical and numerical solution of a nonlinear two delays integral equations", technical report RT 342/08, Istituto per le Applicazioni del Calcolo "Mauro Picone", 2008. To appear on *Math. Comp. Simul.*

[Messina et al. 2008b] E. Messina, E. Russo, and A. Vecchio, "A stable numerical method for Volterra integral equations with discontinuous kernel", *J. Math. Anal. Appl.* **337**:2 (2008), 1383–1393. MR 2386385 (2008m:45008

[Messina et al. 2009] E. Messina, E. Russo, and A. Vecchio, "A convolution test equation for double delay integral equations", *J. Comput. Appl. Math.* **228**:2 (2009), 589–599. MR 2523175 (2010d:65386

[Messina et al. 2010] E. Messina, Y. Muroya, E. Russo, and A. Vecchio, "Convergence of solutions for two delays Volterra integral equations in the critical case", *Appl. Math. Lett.* **23** (2010), 1162–1165.

Received: 2010-07-09 Revised: 2010-09-15 Accepted: 2010-09-24

immacolata.garzilli@gmail.com *Università degli Studi di Napoli "Federico II", Dipartimento di Matematica e Applicazioni R.Caccioppoli, via Cintia, I-80126 Napoli, Italy*

eleonora.messina@unina.it *Università degli Studi di Napoli "Federico II", Dipartimento di Matematica e Applicazioni R.Caccioppoli, via Cintia, I-80126 Napoli, Italy*

antonia.vecchio@cnr.it *Sede di Napoli – CNR, Istituto per le Applicazioni del Calcolo "Mauro Picone", Via P. Castellino, 111, I-80131 Napoli, Italy*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.405

Visual representation of the Riemann and Ahlfors maps via the Kerzman–Stein equation

Michael Bolt, Sarah Snoeyink and Ethan Van Andel

(Communicated by Michael Dorff)

The Szegő kernel serves as one of the canonical functions associated to a region in the complex plane, and from it one can compute the Riemann (or Ahlfors) map, the essentially unique conformal transformation of the region to the unit disc. We provide an elementary description of the method that Kerzman and Stein used to compute the Szegő kernel, and subsequently, the Riemann and Ahlfors maps. A description, too, is provided for a new tool that generates visual representations of these functions and is included with the open-source computer algebra system Sage.

1. Introduction

In his Ph.D. thesis in 1851, Bernhard Riemann stated a theorem that has come to be regarded as one of the most important results in complex analysis. It says that no matter how pathological the boundary of a simply connected (open) region, one can map the region to the unit disc in such a way that angles are preserved.

Theorem 1 (Riemann mapping theorem). *Let Ω be a (nonempty) simply connected region in the complex plane that is not the entire plane. Then, for any $z_0 \in \Omega$, there exists a bianalytic map f from Ω to the unit disc such that $f(z_0)=0$ and $f'(z_0)>0$.*

To illustrate the result, Figure 1 shows a map for a triangular region, using colors and contour lines to identify the corresponding points of the transformation. (The colors are visible in the electronic version of this paper.) In this example the orthocenter is mapped to the origin without rotation at this point. We generated these images using a new tool, `Riemann_Map()`, that the third author developed to accompany Sage, a freely available, open-source computer algebra system. The tool is now included in the core Sage library, and using a web browser, one can generate similar pictures on any computer that has an internet connection.

MSC2000: primary 30C30; secondary 65E05.
Keywords: Riemann map, Ahlfors map, Szegő kernel, Kerzman–Stein.
All three authors were supported by the National Science Foundation under Grant no. DMS-0702939.

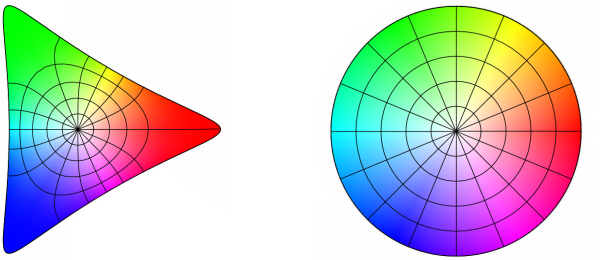


Figure 1. Left: color plot and overlay for the Riemann map of a triangular region. Right: the color scheme for the unit disc.

Although Riemann stated his result in 1851, the first rigorous proof came much later and is due to Carathéodory in 1912. Other proofs have appeared since then, but not all of them provide an easy way to compute the Riemann map. For a proof that is related to the methods used in this paper, see [Garabedian 1991].

The purpose of this paper is threefold. First, we provide a simple description of the method that Kerzman and Stein [1978] used to compute the Riemann map which is based on the Szegő kernel. Next, we provide adaptations of the theory in order to accommodate the case of a multiply connected region and to permit more accurate calculations near the corners of a simply connected region whose boundary is piecewise differentiable. Finally, we describe the numerical implementation of the method and the key features of the tool `Riemann_Map()`, including an adaptation for generating images of the Ahlfors map for a general multiply connected region.

Our implementation of the Nyström method for solving the Kerzman–Stein integral equation is like that used in [Kerzman and Trummer 1986]. Subsequently, that method was modified in [Trummer 1986; O'Donnell and Rokhlin 1989; Murid et al. 1998]. To visualize the Riemann map and Ahlfors map, we use a method devised by Frank Farris [1998] that he calls domain coloring and which uses a color's brightness and hue to indicate the value of a complex function. We mention

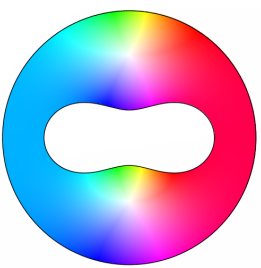


Figure 2. The Ahlfors map for a 2-connected region without a contour overlay. (Color scheme is the same as that of Figure 1.)

that `Riemann_Map()` accepts as data the boundary of a region, given either as a parametrized curve or as a set of boundary points to be interpolated. Using a personal machine, the tool can generate accurate pictures in just seconds.

For ease of presentation, we limit the discussion to regions whose boundary is infinitely differentiable. This means that the boundary curves have a curvature function that is infinitely differentiable with respect to the arc length parameter. The ideas are essentially the same for a twice differentiable region, and many of the results apply even in a still more general context. In particular, `Riemann_Map()` works for domains with piecewise smooth boundary. For a justification of this point, see [Thomas 1996].

The reader who wishes to know more about complex variables than is presented here is encouraged to refer to [Bell 1992; Boas 2010; D'Angelo 2010]. (Bell [1992] offers a completely rigorous treatment of complex analysis in the manner of Kerzman and Stein.) The reader, though, who already has a good grasp of the subject can skip to the last section for an abbreviated users manual for `Riemann_Map()`. We encourage other faculty and student researchers to consider disseminating their work via a platform like Sage. Indeed, we found the review process to be a supportive one, and we were able to get started with relatively little experience working with the programming language Python.

2. Analytic functions and the Cauchy integral formula

The Riemann map and Ahlfors map are examples of analytic functions. For a region $\Omega \subset \mathbb{C}$, there are three equivalent formulations for what this means.

The simplest formulation says that a function $f : \Omega \to \mathbb{C}$ is analytic provided that near any point $z_0 \in \Omega$, it can be expressed as the sum of a power series

$$f(z) = \sum_{j=0}^{\infty} a_j (z - z_0)^j.$$

In fact, when this is the case, the coefficients a_j are the Taylor coefficients for f and are related to the derivatives of f via $a_j = f^{(j)}(z_0)/j!$. The second formulation says that f is analytic provided its real and imaginary parts, $u = \operatorname{Re} f$ and $v = \operatorname{Im} f$, satisfy the partial differential equations,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

which are known as the Cauchy–Riemann equations. This formulation permits an easy demonstration that the real and imaginary parts of an analytic function are harmonic, that is,

$$\triangle u \stackrel{\text{def}}{=} u_{xx} + u_{yy} = 0 \quad \text{and} \quad \triangle v \stackrel{\text{def}}{=} v_{xx} + v_{yy} = 0.$$

The third formulation is familiar from calculus. In this case f is analytic if it is everywhere differentiable, that is,

$$f'(z_0) \stackrel{\text{def}}{=} \lim_{h\to 0} \frac{f(z_0+h)-f(z_0)}{h}$$

exists at each $z_0 \in \Omega$, where it is important to note that h is assumed complex. Of course it is easy to see from this formulation that polynomials with variable z are analytic—one proceeds in the same way as one would compute f' in calculus. We leave for the reader the additional exercise that $f(z) = |z|^2 = z\overline{z}$ is *not* analytic according to this formulation.

Essential to the Kerzman and Stein method is the Cauchy integral formula, one of the most basic results in complex analysis. It says that an analytic function can be expressed as an average of its values along a bounding curve.

Theorem 2 (Cauchy integral formula). *Let f be analytic in a simply connected region $\Omega \subset \mathbb{C}$ and let γ be a simple closed positively oriented curve in Ω. If z_0 is a point that lies interior to γ, then*

$$f(z_0) = \frac{1}{2\pi i} \int_{w\in\gamma} \frac{f(w)\,dw}{w-z_0}.$$

We mention that the equivalence of the second and third formulations of analyticity requires only a small amount of multivariable calculus. To see that a function which is analytic by the first formulation is analytic by the third formulation, one differentiates term-by-term using the standard results about power series. To see that a function which is analytic by the second formulation is analytic by the first formulation, one uses the Cauchy integral formula. The argument needed for this will be apparent after reading the next section.

3. The Cauchy projector

The Kerzman and Stein method begins with the observation that for a general function f defined on the boundary of a region, the Cauchy integral defines an analytic function $\mathcal{C}f$ inside the region. In particular, if one defines

$$\mathcal{C}f(z) = \frac{1}{2\pi i} \int_{w\in\partial\Omega} \frac{f(w)\,dw}{w-z} \qquad \text{for } z \in \Omega,$$

then $\mathcal{C}f$ is analytic inside Ω. To see this, one expands $(w-z)^{-1}$ in a small disc centered at any $z_0 \in \Omega$ using the geometric series,

$$\frac{1}{w-z} = \frac{1}{w-z_0}\left[1 + \frac{z-z_0}{w-z_0} + \left(\frac{z-z_0}{w-z_0}\right)^2 + \left(\frac{z-z_0}{w-z_0}\right)^3 + \cdots\right].$$

The coefficients of the power series for $\mathcal{C}f$, centered at z_0, are then obtained by integration,

$$a_j = \frac{1}{2\pi i} \int_{w \in \partial\Omega} \frac{f(w)\, dw}{(w - z_0)^{j+1}}.$$

For a function f that is integrable on $\partial\Omega$, the series is sure to converge in any disc small enough to fit inside Ω, i.e., small enough for the geometric series to converge for every $w \in \partial\Omega$.

It follows, too, from the Cauchy integral formula, that if f begins as the boundary values of a function that is analytic inside Ω, then the Cauchy integral reproduces the values of that analytic function.

By finally calling on some approximation theory, we are then able to identify a context in which the Cauchy integral behaves as a projection operator. The theory shows that if f begins as an infinitely differentiable function on the boundary, then the function $\mathcal{C}f$, at first defined inside the region, extends continuously and infinitely differentiably on the closure of the region. For the proof of this fact we refer to the first chapter of [Bell 1992].

To help summarize our observations, let $C^\infty(\partial\Omega)$ denote the space of functions that are infinitely differentiable on the boundary and let $A^\infty(\partial\Omega)$ denote the subspace of functions that extend continuously and analytically inside the region. These are vector spaces over $\mathbb{C}$ and $A^\infty(\partial\Omega)$ is a subspace of $C^\infty(\partial\Omega)$. We have then established that the Cauchy integral maps $C^\infty(\partial\Omega)$ to $A^\infty(\partial\Omega)$, and it acts identically on $A^\infty(\partial\Omega)$. Although one might argue that there is an abuse of notation, we will write $\mathcal{C}\colon C^\infty(\partial\Omega) \to A^\infty(\partial\Omega)$ for this projection operator.

To illustrate the construction, we give a direct calculation for the unit disc. We begin with a general function f, defined on the unit circle, that can be expressed as a Fourier series

$$f(e^{it}) = \sum_{j=-\infty}^{\infty} a_j e^{ijt} = \sum_{j \geq 0} a_j e^{ijt} + \sum_{j<0} a_j e^{ijt}.$$

Using the boundary parametrization, $w(t) = e^{it}$ for $0 \leq t < 2\pi$, and expressing in polar form, $z = re^{is}$ for $0 \leq r < 1$ and $0 \leq s < 2\pi$, we evaluate the Cauchy integral by expanding the kernel in a geometric series,

$$\begin{aligned}\mathcal{C}f(re^{is}) &= \frac{1}{2\pi i} \int_0^{2\pi} \frac{f(e^{it})\, ie^{it}\, dt}{e^{it} - re^{is}} = \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=0}^{\infty} r^k e^{ik(s-t)} \sum_{j=-\infty}^{\infty} a_j e^{ijt}\, dt \\ &= \sum_{j \geq 0} a_j\, r^j e^{ijs}.\end{aligned}$$

(The last step uses the fact that $\int_0^{2\pi} e^{int} dt = 2\pi$ if $n = 0$; the integral is zero

otherwise.) We conclude that $\mathcal{C}f(z) = \sum_{j\geq 0} a_j z^j$. Then letting $r \uparrow 1$ gives

$$\mathcal{C}f(e^{it}) = \sum_{j\geq 0} a_j e^{ijt}.$$

For reference, we mention that the situation should be reminiscent of linear algebra, where projection operators map finite dimensional spaces onto lower dimensional subspaces. To illustrate, identify points with position vectors and consider the operator that is represented using the standard basis by the matrix $\left(\begin{smallmatrix} 0 & 2 \\ 0 & 1 \end{smallmatrix}\right)$. This operator maps points in $\mathbb{R}^2$ to points on the line $x - 2y = 0$, and it does so in such a way that points on the line are preserved. We leave these easy facts for the reader to check, and we return to the example in the next section.

Like the example from linear algebra, we mention that the Cauchy projector is linear, since integration is a linear process. In particular, $\mathcal{C}(f + \lambda g) = \mathcal{C}f + \lambda\,\mathcal{C}g$ for $\lambda \in \mathbb{C}$. A fundamental difference, though, is the fact that the Cauchy projector acts between infinite dimensional spaces, as is evident in the example of the unit disc. As will be seen in the next section, however, its skew-hermitian part behaves like its finite dimensional counterpart.

4. The Szegő projector and the Kerzman–Stein equation

We saw in Section 3 that the Cauchy integral provides a projection from $C^\infty(\partial\Omega)$ to $A^\infty(\partial\Omega)$. But the projection is not generally *orthogonal*. To illustrate, Figure 3 shows two projections from $\mathbb{R}^2$ onto the line $x - 2y = 0$. The first is the projection described at the end of the last section, and the second is the orthogonal (shortest distance) projection of $\mathbb{R}^2$ onto the same line.

To make sense of this one needs a notion of distance. In the linear algebra example, distance is measured in the Euclidean way, and this distance arises from the standard dot product. The analogous inner product for functions is given by

$$(f, g) = \int_{\partial\Omega} f\overline{g}\, ds,$$

where ds indicates that integration is done with respect to arc length. The norm of a function is then given by $\|f\| = \sqrt{(f, f)}$ and the distance between functions

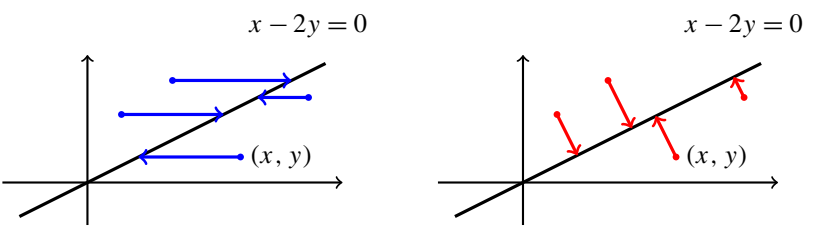


Figure 3. Nonorthogonal and orthogonal projections to $x - 2y = 0$.

is the norm of their difference. Finally, the Szegő projector can be defined as the *orthogonal* projection,

$$\mathcal{S} : C^\infty(\partial\Omega) \xrightarrow{\perp} A^\infty(\partial\Omega).$$

At first it may not be obvious that $A^\infty(\partial\Omega) \subset C^\infty(\partial\Omega)$ is a *closed* subspace — a nontrivial fact since both spaces are infinite dimensional. In particular, it may not be obvious that the Szegő projector maps an infinitely differentiable function to an infinitely differentiable function. These properties, however, can be shown to follow as consequences of the Kerzman and Stein theory. We again refer to [Bell 1992] for a treatment of these delicate facts.

The key insight behind the Kerzman and Stein theory can be described now as follows. The Cauchy integral provides an explicitly computable, though generally nonorthogonal, projection $\mathcal{C}: C^\infty(\partial\Omega) \to A^\infty(\partial\Omega)$. Meanwhile, the Szegő projector represents the uniquely orthogonal, though initially noncomputable, projection $\mathcal{S} : C^\infty(\partial\Omega) \to A^\infty(\partial\Omega)$. The projections are related by the equation,

$$\mathcal{S}(\mathcal{I} + \mathcal{A}) = \mathcal{C}, \tag{1}$$

where $\mathcal{I}$ is the identity operator and $\mathcal{A} = \mathcal{C} - \mathcal{C}^*$ is the Kerzman–Stein operator. In the next section we will see how this leads to a simple way for computing the Riemann map and Ahlfors map.

The effectiveness of the Kerzman–Stein equation balances on the fact that the Kerzman–Stein operator is an integral operator with a well behaved kernel

$$A(z, w) = \frac{1}{2\pi i}\left(\frac{T(w)}{w - z} - \frac{\overline{T(z)}}{\overline{w} - \overline{z}}\right)$$

for $w, z \in \partial\Omega$. Here, $T(w)$ is the unit tangent vector at $w \in \partial\Omega$, so $dw = T(w)\, ds_w$. In particular, the singularities at $w = z$ cancel each other and the kernel is infinitely differentiable on $\partial\Omega \times \partial\Omega$. The significance of this fact is that the Kerzman–Stein operator for a region with finite boundary is compact; that is, it can be approximated in norm by finite rank operators. For details on this point, we direct the reader to any functional analysis text. See, for instance, [Zimmer 1990, Chapter 3].

We leave for the reader to check the claim that the singularities in $A(z, w)$ in fact do cancel. This can be done using a Taylor expansion involving an arc length parametrization of the boundary. (One then makes replacements $w = z(s)$ and $z = z(t)$, so that also $T(w) = z'(s)$ and $T(z) = z'(t)$.) By carrying the expansions a few steps further, one can see additionally that the kernel vanishes precisely when the boundary has constant curvature. It follows that precisely when the region is a disc or half plane, the Cauchy and Szegő projectors are the same.

We also leave for the reader to check the analogue of the Kerzman–Stein equation for the example from linear algebra. In this case, the orthogonal projector is

represented by the matrix

$$\begin{pmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & \frac{1}{5} \end{pmatrix}$$

and the adjoint operator, needed for the computation of $\mathcal{A} = \mathcal{C} - \mathcal{C}^*$, is gotten by taking the transpose of the matrix.

5. Relationship to the Riemann map and Ahlfors map

As described in the introduction, the Riemann map is the essentially unique conformal map from a simply connected region to the unit disc; the Ahlfors map is the essentially unique such conformal map for a multiply connected region. These maps can be expressed as analytic functions with nonvanishing derivatives. In particular, by manipulating the Cauchy–Riemann equations, one can show that at an arbitrary point $z_0 \in \Omega$, an analytic function f accomplishes a rotation by angle $\arg f'(z_0)$. This rotation is paired with a uniform dilation at z_0 by factor $|f'(z_0)|$.

A fundamental relationship between the Riemann map or Ahlfors map and the Szegő projector can be derived using a transformation law for the Szegő kernel. Indeed, the Szegő projector is an integral operator whose kernel can be represented in terms of an orthonormal basis for $A^\infty(\partial\Omega)$. If $\{\phi_j\}_{j\in\mathbb{N}}$ is such a basis, then

$$S(z, \overline{w}) = \sum_{j\in\mathbb{N}} \phi_j(z)\overline{\phi_j(w)},$$

and for a general function f,

$$\mathcal{S}f(z) = \int_{w\in\partial\Omega} S(z, \overline{w})\, f(w)\, ds_w.$$

From the work of Kerzman and Stein, it follows that the Szegő kernel is the solution to the integral equation

$$S(z, \overline{z}_0) - \int_{w\in\partial\Omega} A(z, w)\, S(w, \overline{z}_0)\, ds_w = \frac{1}{2\pi i}\frac{\overline{T(z)}}{\overline{z}_0 - \overline{z}} \quad \text{for } z \in \partial\Omega,\ z_0 \in \Omega. \quad (2)$$

This equation can be seen to follow from (1). In particular, by taking adjoints of (1) one obtains $(\mathcal{I} - \mathcal{A})\mathcal{S} = \mathcal{C}^*$, and following this, one utilizes an approximate identity to obtain (2).

As will be needed, for the case of the unit disc, Δ, one can use a basis $\{\phi_j\}_{j\in\mathbb{N}}$ where $\phi_j(z) = z^{j-1}/\sqrt{2\pi}$ in order to see that

$$S(z, \overline{w}) = \frac{1}{2\pi}\sum_{j\in\mathbb{N}} z^{j-1}\overline{w}^{j-1} = \frac{1}{2\pi}\frac{1}{1 - z\overline{w}}.$$

This is consistent with the earlier observation for the disc that the Szegő projector and Cauchy projector are the same. Indeed, the Cauchy projector for the disc has

kernel

$$C(z, w) = \frac{1}{2\pi i}\frac{T(w)}{w - z} = \frac{1}{2\pi i}\frac{iw}{w - z} = \frac{1}{2\pi}\frac{1}{1 - z\overline{w}}.$$

In the remainder of this section we derive the relationship between the Szegő kernel and Riemann map for a simply connected region. The derivation for the Ahlfors map is handled differently, although the implementation will be the same. This is discussed in the next section. In the next section we also provide details for the numerical solution of (2).

Assuming then that $f : \Omega_1 \to \Omega_2$ is bianalytic, that is, analytic with an analytic inverse, one can show that the operator $\Lambda : C^\infty(\partial\Omega_2) \to C^\infty(\partial\Omega_1)$ defined according to $\Lambda\phi = (\phi \circ f) \cdot \sqrt{f'}$ sends an orthonormal basis for $A^\infty(\partial\Omega_2)$ to an orthonormal basis for $A^\infty(\partial\Omega_1)$. It follows that

$$S_1(z, \overline{w}) = S_2(f(z), \overline{f(w)})\, f'(z)^{1/2}\, \overline{f'(w)^{1/2}}$$

where S_1, S_2 are the Szegő kernels for Ω_1, Ω_2, respectively. Applying this result to the case of the Riemann map $f : \Omega \to \Delta$ normalized for $z_0 \in \Omega$ so that $f(z_0) = 0$ and $f'(z_0) > 0$, one finds that

$$S(z, \overline{z}_0) = \frac{f'(z)^{1/2} f'(z_0)^{1/2}}{2\pi}.$$

From this it follows that

$$f'(z) = \frac{2\pi\, S(z, \overline{z}_0)^2}{S(z_0, \overline{z}_0)}, \tag{3}$$

where the relationship holds first for $z \in \Omega$, and then by continuity, it holds for $z \in \overline{\Omega}$. There is a simple equation for relating the boundary values of f to those of f'. So provided with an efficient algorithm for computing the Szegő kernel, we will have an efficient method for computing the Riemann map.

6. Numerical implementation

We now begin with a region Ω that can be described as the interior of n finite curves that are parametrized by smooth functions $z = z_j(t)$ defined for $0 \le t \le \ell_j$ such that $z_j(\ell_j) = z_j(0)$ and $z_j'(t)$ is nonvanishing, $1 \le j \le n$. Of course, if the region is simply connected, then $n = 1$ and one can drop the subscripts. It is not necessary that the parametrizations be given according to arc length, but it is necessary that the curves are oriented positively with respect to Ω. We further specify a point $z_0 \in \Omega$ that we anticipate having mapped to the origin without rotation.

We first adapt the method used in [Kerzman and Trummer 1986] to compute the Szegő kernel. In particular, after making the replacements

$$\begin{aligned}\phi_j(t) &= |z_j'(t)|^{1/2}\,(2\pi i)^{-1}\,(z_j'(t)/|z_j'(t)|)\,(\overline{z}_0 - \overline{z}_j(t))^{-1},\\ a_{j,k}(t,s) &= |z_j'(t)|^{1/2}\,|z_k'(s)|^{1/2}\,A(z_j(t), z_k(s)),\\ \psi_j(t) &= |z_j'(t)|^{1/2}\,S(z_j(t), \overline{z}_0),\end{aligned}$$

Equation (2) becomes

$$\psi_j(t) - \sum_{k=1}^{n} \int_0^{\ell_k} a_{j,k}(t,s)\psi_k(s)\,ds = \phi_j(t) \tag{4}$$

for $0 \le t \le \ell_j$ and $1 \le j \le n$. With parametrizations provided, the functions ϕ_j and $a_{j,k}$ are explicitly computable. We wish to solve these equations for the functions $\psi_j(t)$ in order to have the Szegő kernel.

Perhaps the easiest way to solve (4) is via the Nyström method. For $m > 0$ one partitions the intervals $[0, \ell_j]$ using $0 = s_0^j < s_1^j < s_2^j < \cdots < s_m^j = \ell_j$ and replaces (4) by its approximation

$$\psi_j(t) - \sum_{k=1}^{n} \sum_{l=1}^{m} a_{j,k}(t, s_l^k)\psi_k(s_l^k)\,\Delta s_l^k = \phi_j(t). \tag{5}$$

Here, $\Delta s_l^k \stackrel{\text{def}}{=} s_l^k - s_{l-1}^k = \ell_k/m$. This equation is next solved explicitly for $\psi_j(t)$ when $t = s_i^j$. Indeed, after replacing $t = s_i^j$, this is tantamount to solving a system of nm linear equations in nm complex unknowns,

$$(I - B)x = y,$$

where the skew-hermitian matrix B has entry $a_{j,k}(s_i^j, s_l^k)$ in its $(n(j-1)+i)$-th row and $(n(k-1)+l)$-th column. With this setup, the $(n(k-1)+l)$-th entry in column vector x is $\psi_k(s_l^k)$ and the $(n(j-1)+i)$-th entry in column vector y is $\phi_j(s_i^j)$.

With values for $\psi_j(s_i^j)$ now determined, the general values for $\psi_j(t)$ can be recovered from (5). In particular, we set

$$\psi_j(t) = \phi_j(t) + \sum_{k=1}^{n} \sum_{l=1}^{m} a_{j,k}(t, s_l^k)\psi_k(s_l^k)\,\Delta s_l^k, \tag{6}$$

where on the right side we use the values $\psi_k(s_l^k)$ already determined. It may appear that with this formula we are redefining $\psi_j(t)$ for $t = s_i^j$, when in fact we are not, since (6) is identical with (5).

Alternatively, we have found it to be effective to replace this last step with a simple linear interpolation of the values $\psi_j(s_i^j)$ for $1 \le i \le m$. In our implementation this helps to prevent the repeated evaluations of ϕ_j and $a_{j,k}$ that otherwise would be needed. (The evaluations may have complicated expressions embedded in the

parametrizations.) To be sure, `Riemann_Map()` seems most effective without the additional evaluations—with linear interpolation one can rather increase m for a finer partition, and as a result, obtain better image resolution.

With values of the Szegő kernel $S(z, \overline{z}_0)$ now known for a designated $z_0 \in \Omega$ and for $z \in \partial\Omega$, we look to recover values for the Riemann map and Ahlfors map. We introduce boundary correspondence functions $\theta_j = \theta_j(t)$ defined for $0 \le t \le \ell_j$ by

$$f(z_j(t)) = e^{i\theta_j(t)},$$

where $f : \Omega \to \Delta$ is the Riemann map or Ahlfors map. Of course, the θ_j are real and unique only to a multiple of 2π. Differentiating this equation gives

$$f'(z_j(t))\, z_j'(t) = i\theta_j'(t)\, e^{i\theta_j(t)} = i\theta_j'(t)\, f(z_j(t)),$$

so that in the simply connected case, we obtain from (3) that

$$\begin{aligned} f(z_j(t)) &= \frac{f'(z_j(t))z_j'(t)}{i\theta_j'(t)} = \frac{z_j'(t)}{i\theta_j'(t)} \frac{2\pi\, S(z_j(t), \overline{z}_0)^2}{S(z_0, \overline{z}_0)} \\ &= \frac{2\pi}{i\theta_j'(t)} \frac{z_j'(t)}{|z_j'(t)|} \frac{\psi_j(t)^2}{S(z_0, \overline{z}_0)}. \end{aligned} \tag{7}$$

Many factors in this equation are positive, so taking arguments yields simply

$$\theta_j(t) = \arg(-i z_j'(t) \psi_j(t)^2). \tag{8}$$

With the boundary values of the Riemann map now determined, one can obtain the interior values via the Cauchy integral formula, where the integrals can be evaluated using the same Riemann sum approximations used earlier in this section.

We mention that there is another way to recover the boundary correspondence function that is better suited for regions with corners. In particular, by taking the modulus of both sides of (7), we obtain

$$\theta_j'(t) = \frac{2\pi\, |\psi_j(t)|^2}{S(z_0, \overline{z}_0)}, \quad \text{where } S(z_0, \overline{z}_0) = \sum_{k=1}^{n} \int_0^{\ell_k} |\psi_k(t)|^2 dt. \tag{9}$$

The correspondence functions can be obtained via integration, using initial conditions derived using (8). This method results in correspondence functions that are continuous at the corners, as they should be, avoiding errors caused by taking the argument of the tangent vector.

It takes us outside the scope of our discussion to establish these formulas for multiply connected regions, but we mention that (8) remains valid for $n > 1$. This follows from an argument based on the identity relating the Szegő and Garabedian kernels [Bell 1992, page 24]. It is not clear to us if (9) remains valid for $n > 1$.

7. The Sage package `Riemann_Map()`

Led by examples, we now give a brief description of the package, `Riemann_Map()`, which was constructed using the methods of the previous sections. We encourage those who are new to Sage to try the examples online by first setting up a free account at http://www.sagemath.org. (We have published there a worksheet 'Riemann_Map() illustrated' that contains the examples.) For each case, we show both the text to be entered in a cell of a worksheet and the Sage output that follows the cell's evaluation. As typical in Sage, one can find documentation for the package by evaluating a cell that contains only the question `Riemann_Map?`.

Example 1: The Riemann map for an ellipse. To use `Riemann_Map()` for a parametrized ellipse (Figure 4), one provides three variables: a function parametrizing the boundary of the ellipse and whose domain is the interval $[0, 2\pi]$, the derivative of this function, and a complex number identifying the point inside the ellipse that is to be mapped to the origin without rotation. To generate a representation for the Riemann map, one subsequently applies the methods `plot_colored()` and `plot_spiderweb()` in order to have a combined representation showing the coloring of the region and the contour overlay. One obtains sharper resolution by increasing the value of `plot_points` at the expense of increased processing time.

```
z(t) = exp(I*t) +.5*exp(-I*t)   # Riemann map for an ellipse
zp(t) = I*exp(I*t) -.5*I*exp(-I*t)
m = Riemann_Map([z],[zp],0)
p = m.plot_colored(plot_points=500) +m.plot_spiderweb()
show(p,axes=false)
```

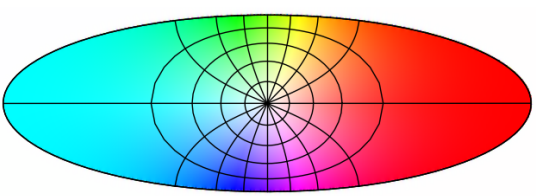


Figure 4. Color representation of the Riemann map for an ellipse.

Example 2: The Riemann map for a square. For a square (Figure 5), one can proceed as for the ellipse using piecewise-defined functions for the parametrization and its derivatives. Alternatively, one can utilize the `polygon_spline()` package along with its methods `value()` and `derivative()` to get these functions more quickly. It is worth noting that command syntax in Sage is shared with the programming language Python, so there is carry-over to learning either of the two languages. We also mention that in the contour overlay, we have increased the

```
ps = polygon_spline([(-1,-1),(1,-1),(1,1),(-1,1)])
z = lambda t: ps.value(t)            # Riemann map for a square
zp = lambda t: ps.derivative(t)
m = Riemann_Map([z],[zp],.3+.3*I)
p = m.plot_colored(plot_points=1000) +m.plot_spiderweb(pts=150)
show(p,axes=false)
```

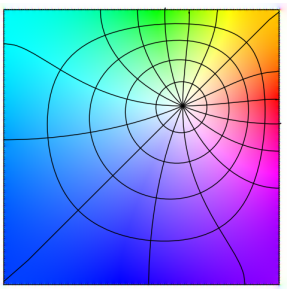


Figure 5. Color representation of the Riemann map for a square.

number of points used to draw concentric rings and radial lines from the default 32 to 150. The extra precision was needed so that the radial lines appear perpendicular to the boundary as they should be.

Example 3: The Ahlfors map for an annulus. For an annulus (Figure 6), one proceeds as in the case of an ellipse, using a parametrization for each of the boundary components. We mention that when `Riemann_Map()` is called, it is necessary that

```
z1(t) = 2*exp(I*t)                   # Ahlfors map for an annulus
z1p(t) = 2*I*exp(I*t)
z2(t) = exp(-I*t); z2p(t) = -I*exp(-I*t)
m = Riemann_Map([z1,z2],[z1p,z2p],sqrt(2)*I)
p = m.plot_colored(plot_points=1000) +m.plot_boundaries()
show(p,axes=false)
```

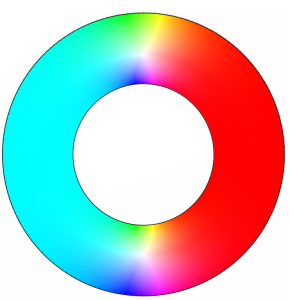


Figure 6. Color representation of the Ahlfors map for an annulus.

the outside curve is provided first in each list of functions. One also must be careful to give parametrizations that are oriented positively with respect to the interior region. Finally, we mention that the methods are not yet sufficiently developed to provide contour overlays for regions with more than one boundary component.

Example 4: The Ahlfors map for a triply connected region. As one more example that combines the elements of the previous examples, we draw the Ahlfors map for a triply connected region composed of a rectangle with two discs removed; see Figure 7. It should be apparent for this example that the Ahlfors map is 3-to-1. As for the previous example, we overlaid the color plot with the region's boundary — this makes the boundary more pronounced and conceals the nearby graininess.

Example 5: The Riemann map for a general region. For our final example, illustrated in Figure 8, we use the companion package `complex_cubic_spline()` to show how to map regions whose boundary is provided by a set of points to be interpolated. For this example, we generated three lists of points which sample two line segments and a circular arc. The combined list of 600 points is suggestive of the boundary of a region that is well-approximated by splines. The subsequent methods `value()` and `derivative()` provide functions giving a parametrization and its derivative for a cubic spline interpolant of the given list. It is worth noting that it is not necessary for the points in the list to be equally spaced, just as it is not necessary for parametrizations to have constant speed.

```
ps = polygon_spline([(-4,-2),(4,-2),(4,2),(-4,2)])
z1 = lambda t: ps.value(t); z1p = lambda t: ps.derivative(t)
z2(t) = -2+exp(-I*t); z2p(t) = -I*exp(-I*t)
z3(t) = 2+exp(-I*t); z3p(t) = -I*exp(-I*t)
m = Riemann_Map([z1,z2,z3],[z1p,z2p,z3p],0)
p = m.plot_colored(plot_points=1000) +m.plot_boundaries()
show(p,axes=false)
```

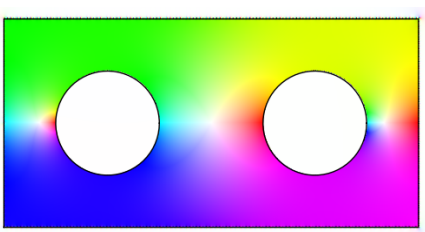


Figure 7. Representation of the Ahlfors map for a triply connected region.

```
li1 = [(sqrt(3)-I)*(t/200)-(3+I)*(1-t/200) for t in range(200)]
li2 = [2*exp(pi*I*(t-100)/600) for t in range(200)]
li3 = [(sqrt(3)+I)*(1-t/200)-(3+I)*(t/200) for t in range(200)]
cs = complex_cubic_spline(li1+li2+li3)
m = Riemann_Map([lambda x: cs.value(x)], \
                    [lambda x: cs.derivative(x)],1-.25*I)
p = m.plot_colored(plot_points=1000) +m.plot_spiderweb(pts=64)
show(p,axes=false)
```

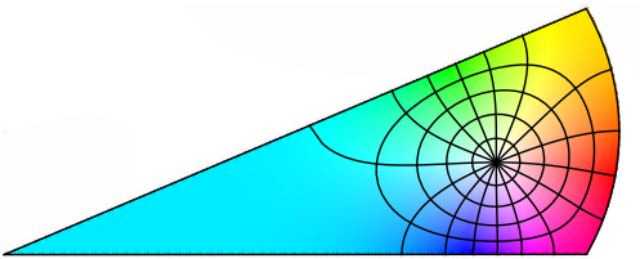


Figure 8. Color representation of the Riemann map for a general region.

References

[Bell 1992] S. R. Bell, *The Cauchy transform, potential theory, and conformal mapping*, CRC Press, Boca Raton, FL, 1992. MR 94k:30013

[Boas 2010] R. P. Boas, *Invitation to complex analysis*, 2nd ed., Mathematical Association of America, Washington, DC, 2010. MR 2674618 Zbl 05769949

[D'Angelo 2010] J. P. D'Angelo, *An Introduction to Complex Analysis and Geometry*, American Mathematical Society, 2010.

[Farris 1998] F. A. Farris, "Reviews: Visual Complex Analysis", *Amer. Math. Monthly* **105**:6 (1998), 570–576. MR 1543280

[Garabedian 1991] P. R. Garabedian, "A simple proof of a simple version of the Riemann mapping theorem by simple functional analysis", *Amer. Math. Monthly* **98**:9 (1991), 824–826. MR 92j:30009 Zbl 0741.30007

[Kerzman and Stein 1978] N. Kerzman and E. M. Stein, "The Cauchy kernel, the Szegö kernel, and the Riemann mapping function", *Math. Ann.* **236**:1 (1978), 85–93. MR 58 #6199

[Kerzman and Trummer 1986] N. Kerzman and M. R. Trummer, "Numerical conformal mapping via the Szegő kernel", *J. Comput. Appl. Math.* **14**:1-2 (1986), 111–123. MR 87f:30017

[Murid et al. 1998] A. H. M. Murid, M. Z. Nashed, and M. R. M. Razali, "Numerical conformal mapping for exterior regions via the Kerzman–Stein kernel", *J. Integral Equations Appl.* **10**:4 (1998), 517–532. MR 99k:30008 Zbl 0919.30008

[O'Donnell and Rokhlin 1989] S. T. O'Donnell and V. Rokhlin, "A fast algorithm for the numerical evaluation of conformal mappings", *SIAM J. Sci. Statist. Comput.* **10**:3 (1989), 475–487. MR 90h:65034 Zbl 0672.30006

[Thomas 1996] A. D. Thomas, "Conformal mapping of nonsmooth domains via the Kerzman–Stein integral equation", *J. Math. Anal. Appl.* **200**:1 (1996), 162–181. MR 97h:46038 Zbl 0846.30005

[Trummer 1986] M. R. Trummer, "An efficient implementation of a conformal mapping method based on the Szegő kernel", *SIAM J. Numer. Anal.* **23**:4 (1986), 853–872. MR 87k:30013

[Zimmer 1990] R. J. Zimmer, *Essential results of functional analysis*, University of Chicago Press, Chicago, IL, 1990. MR 91h:46002 Zbl 0708.46001

Received: 2010-07-23 Revised: 2010-11-16 Accepted: 2010-11-20

mbolt@calvin.edu *Department of Mathematics and Statistics, Calvin College, 3201 Burton St., S.E., Grand Rapids, MI 49546, United States* http://www.calvin.edu/~mdb7/

sarah.snoeyink@colorado.edu *Department of Mathematics, University of Colorado at Boulder, Campus Box 395, Boulder, CO 80309, United States*

esv5@students.calvin.edu *Department of Mathematics and Statistics, Calvin College, 3201 Burton St., S.E., Grand Rapids, MI 49546, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.421

A topological generalization of partition regularity

Liam Solus

(Communicated by Chi-Kwong Li)

In 1939, Richard Rado showed that any complex matrix is partition regular over $\mathbb{C}$ if and only if it satisfies the columns condition. Recently, Hogben and McLeod explored the linear algebraic properties of matrices satisfying partition regularity. We further the discourse by generalizing the notion of partition regularity beyond systems of linear equations to topological surfaces and graphs. We begin by defining, for an arbitrary matrix Φ, the metric space (M_Φ, δ). Here, M_Φ is the set of all matrices equivalent to Φ that are (not) partition regular if Φ is (not) partition regular; and for elementary matrices, E_i and F_j, we let $\delta(A, B) = \min\{m = l+k : B = E_1 \dots E_l A F_1 \dots F_k\}$. Subsequently, we illustrate that partition regularity is in fact a local property in the topological sense, and uncover some of the properties of partition regularity from this perspective. We then use these properties to establish that all compact topological surfaces are partition regular.

1. Introduction

Let $\mathbb{C}$ be the set of complex numbers, and let $\mathbb{M}_{u,v}(\mathbb{C})$ be the set of all $u \times v$ matrices with complex entries. Let $A = [a_{i,j}] \in \mathbb{M}_{u,v}(\mathbb{C})$ be given, and let $\vec{a}_j$ denote the column j of A. Then A satisfies the *columns condition* if and only if there exists an $m \in \{1, \dots, v\}$ and a partition $\{I_1, \dots, I_m\}$ of $\{1, \dots, v\}$ into nonempty sets such that

(i) $\sum_{j \in I_1} \vec{a}_j = \vec{\mathbf{0}}$, and

(ii) for each $t \in \{2, 3, \dots, m\}$ (if any), $\sum_{i \in I_t} \vec{a}_i$ is in the span of $\{\vec{a}_i : i \in \bigcup_{j=1}^{t-1} I_j\}$.

A is said to be *partition regular* if it satisfies the *columns condition* [Hindman 2007; Rado 1943]. The study of partition regularity has long been a combinatorial endeavor, which mostly uses the columns condition to check if a given matrix is partition regular. However, Hogben and McLeod [2010] recently showed that the columns condition is interesting in its own right, and provided a more linear

MSC2000: primary 05C99, 05E99, 15A06, 54H10, 57N05; secondary 15A99, 54E35.
Keywords: partition regularity, columns condition, graphs, metric space, discrete topology, topological surface, triangulation.

algebraic perspective on partition regularity. We employ this new perspective to extend the notion of partition regularity into geometrical and topological settings.

For an arbitrary complex matrix Φ, we construct a metric space characterized by the partition regularity of Φ (Section 2). We then use this metric space about Φ to generate a topological space that recasts partition regularity as a local property. We show a few topological properties of these spaces, and then demonstrate how their systems of neighborhoods can describe the "degree" of partition regularity as applied to a given matrix (Section 2). Finally, using some well known connections between graph theory and linear algebra, we construct topological spaces that allow us to define partition regularity as a property of topological surfaces and graphs. In Section 3, we show that all compact topological surfaces are partition regular. We then demonstrate that not all graphs are partition regular.

We take a *topological surface* to be a two-dimensional real manifold that is Hausdorff. A *graph* $G = (V, E)$ is a nonempty set V of vertices, along with a set E of edges, where an edge is a two-element subset of vertices. A *walk* is an alternating sequence $(v_0, e_1, v_1, e_2, \ldots, e_m, v_m)$ of vertices and edges. A graph G is *connected* if there exists a walk between any two distinct vertices of G. A *component* is a connected subgraph of G, and a set S of edges of G is a *disconnecting set* if $G \setminus S$ has more than one component. The *edge connectivity* of G is the minimum size of a disconnecting set of G. An *orientation* Γ of G is obtained by assigning a direction to each edge of G, and thus replacing the edge $\{i, j\}$ with the arc (i, j). An orientation Γ of G is *strongly connected* if there exists an alternating sequence $(v_0, e_1, v_1, e_2, \ldots, e_m, v_m)$ of vertices and arcs between any two vertices of Γ. The *oriented incidence matrix* of Γ is the rational matrix denoted $D_\Gamma = [d_{i,e}]$, where if $e = (i, j)$, then $d_{i,e} = -1$, $d_{j,e} = 1$, and $d_{k,e} = 0$ for $k \neq i$ and $k \neq j$.

For any matrix A in $\mathbb{M}_{m,n}(\mathbb{C})$, we let a *type-1* elementary operation be a row (column) permutation, a *type-2* elementary operation be multiplication of a given row (column) of A by a scalar $\beta \in \mathbb{C}$, and a *type-3* elementary operation be the addition of a scalar multiple of one row (column) of A to another. We call the associated matrices of each elementary operation T1, T2, and T3 matrices, respectively.

2. Topologically rich spaces associated with partition regularity

Let $A, B \in \mathbb{M}_{m,n}(\mathbb{C})$. We say that B is *equivalent* to A if there exist invertible matrices P and Q for which $B = PAQ$. This is an equivalence relation on $\mathbb{M}_{m,n}(\mathbb{C})$, and we let $[A]$ denote the equivalence class of A. Since P and Q are each the product of a finite number of elementary matrices we can identify P and Q, with the sequence of nonidentity elementary matrices $\langle x \rangle_{i=1}^{l}$ that when applied to matrix A produces matrix B. Since A and B are both in $[A]$, there must exist a minimal sequence of elementary operations. Let $l_{A,B}$ be the nonnegative integer denoting

the length of this minimal sequence. Then we can define the function

$$\delta : [A] \times [A] \longrightarrow \mathbb{R}$$

such that

$$\delta(A, B) = l_{A,B}.$$

Theorem 2.1. *Let A be in $\mathbb{M}_{m,n}(\mathbb{C})$. Then $([A], \delta)$ is a metric space.*

Proof. The nonnegativity of δ follows trivially from the definition of $l_{A,B}$ for any pair of matrices A, B in $[A]$.

To see that δ is symmetric, notice that if $\delta(A, B) = l_{A,B}$, then there exist invertible matrices P and Q associated with the minimal sequence $\langle x\rangle_{i=1}^{l_{A,B}}$ such that $B = PAQ$. It follows that $A = P^{-1}BQ^{-1}$, and thus P^{-1} and Q^{-1} may be associated with the sequence $\langle x\rangle_{i=1}^{m}$, where $m = l_{A,B}$ is equal to the number of elementary matrices in the product $P^{-1}Q^{-1}$. Assume that $\langle x\rangle_{i=1}^{m}$ is not minimal, then there must exist a sequence $\langle x\rangle_{i=1}^{l_{B,A}}$ such that $l_{B,A} < m$. Find its associated invertible matrices U and V such that $A = UBV$. So, $B = U^{-1}AV^{-1}$ and there is an associated sequence $\langle x\rangle_{i=1}^{n}$. But $n = l_{B,A}$, and since $m = l_{A,B}$, this contradicts the minimality of $\langle x\rangle_{i=1}^{l_{A,B}}$. Thus,

$$\delta(A, B) = l_{A,B} = m = l_{B,A} = \delta(B, A),$$

and δ is a symmetric function.

To see that δ satisfies the triangle inequality, let A, B, and C be in $[A]$, and pick G, P, U, V, N, and Q such that $B = GAP$, $C = UAV$, and $B = NCQ$. Then

$$\delta(A, B) = l_{A,B}, \quad \delta(A, C) = l_{A,C}, \quad \delta(C, B) = l_{C,B}.$$

Now let $m = \delta(A, C) + \delta(C, B)$. Then there exists a sequence $\langle x\rangle_{i=1}^{m}$ associated with the invertible matrices NU and VQ such that

$$B = (NU)A(VQ).$$

Thus, since A can be changed to B using $\delta(A, C)+\delta(C, B)$ elementary operations, it follows from the minimality of $\delta(A, B)$ that

$$\delta(A, B) \leq \delta(A, C) + \delta(C, B).$$

So, δ satisfies the triangle inequality, and $([A], \delta)$ is indeed a metric space. □

Let $\Phi \in \mathbb{M}_{m,n}(\mathbb{C})$ be a partition regular matrix, and let M_Φ denote the set of all matrices that are partition regular and equivalent to Φ. Notice that (M_Φ, δ) is a metric space. Furthermore, the range of our metric δ is a subset of the nonnegative integers.

Theorem 2.2. *Let $\mathcal{T}$ be the metric topology induced by δ on M_Φ. Then $(M_\Phi, \mathcal{T})$ is a discrete topological space.*

Proof. Let $A, B \in \mathbb{M}_{m,n}(\mathbb{C})$. We need only prove that $\delta(A, B) = 0$ if and only if $A = B$. To see this, notice that for any A, B in M_Φ, $\delta(A, B) = l_{A,B}$ is a nonnegative integer. So, if $\delta(A, B) = 0$, then the minimal number of nonidentity elementary matrices that must be applied to A to produce B is zero. So it must be that $A = B$. Conversely, if $A = B$, then the minimal number of elementary operations that must be applied to A to reach B is equal to 0. Thus, $\delta(A, B) = 0$.

For A_0 in M_Φ consider the open ball of radius $\frac{1}{2}$ about A_0:

$$\mathcal{B}_{A_0,1/2} = \left\{A \in M_\Phi : \delta(A_0, A) < \tfrac{1}{2}\right\} = \{A \in M_\Phi : \delta(A_0, A) = 0\} = \{A_0\}.$$

Therefore, the singletons of M_Φ are all open sets, and so $\mathcal{T}$ is equal to the power set of M_Φ. Thus, the pair $(M_\Phi, \mathcal{T})$ is a discrete topological space. □

Notice that if Φ is not partition regular, then there is a corresponding discrete space consisting of all matrices that are not partition regular and equivalent to Φ. This allows us to establish a "degree" of partition regularity for any arbitrary matrix.

Definition 2.3. Let $A \in \mathbb{M}_{u,v}(\mathbb{C})$ be given.

(a) The *progress* of A is the minimum number, l, of elementary operations that must be performed on A to produce a partition regular matrix. We say that A *has progress* l, and write $\text{pr}(A) = l$, moreover, we write $pr(A) = \infty$ if A cannot be changed into a partition regular matrix via elementary operations.

(b) The *antiprogress* of A is the minimum number, l, of elementary operations that must be performed on A to produce a matrix that is not partition regular. We say that A *has antiprogress* l, and write $\text{apr}(A) = l$, moreover, we write $apr(A) = \infty$ if A cannot be changed into a matrix that is not partition regular via elementary operations.

Any $A \in \mathbb{M}_{u,v}(\mathbb{C})$ has both a progress and an antiprogress. Moreover, A has progress 0 if and only if A is partition regular, and A has antiprogress 0 if and only if A is not partition regular.

We are interested in the collection of matrices that proceed from a given matrix A in M_Φ. The following definitions describe such collections.

Definition 2.4. (a) A *filament* is a sequence of equivalent matrices in M_Φ satisfying the following conditions:

(i) The sequence begins with Φ.
(ii) No matrix in the sequence is repeated.
(iii) The sequence is finite if and only if the last matrix in the sequence has antiprogress 1.
(iv) Each matrix of the sequence is obtained by performing a single elementary operation on the preceding matrix in the sequence.

A filament is called *finite* if the sequence is finite. Otherwise, it is called *infinite*.

(b) A *subfilament associated with A* is a sequence of equivalent matrices in M_Φ starting with matrix A, that satisfies (ii), (iii), and (iv). A subfilament is called *finite* if the sequence is finite. Otherwise, it is called *infinite*.

Example 2.5. Let $\Phi = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Then

$$\left(\begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \right) \quad \text{and} \quad \left(\begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \right)$$

are finite filaments in M_Φ since

$$\begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

are both not partition regular. If $A = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$, then

$$\left(\begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \right)$$

is a finite subfilament associated with A.

The following theorem provides a better description of exactly how the size of M_Φ relates to the degree of partition regularity of Φ.

Theorem 2.6. *For any partition regular matrix Φ in $\mathbb{M}_{m,n}(\mathbb{C})$, M_Φ contains an infinite number of infinite filaments.*

Proof. Since row equivalent matrices share the same nullspace, we have, as an immediate consequence of [Hogben and McLeod 2010, Theorem 2.3], that partition regularity is invariant under elementary row operations. It follows from the definition of the columns condition that partition regularity is also invariant under type-1 column operations. Thus, there exist exactly four types of elementary operations that cannot produce a matrix that has antiprogress 0 from a partition regular matrix. For any $A \in M_\Phi$ we may apply a type-1 row operation for every possible pair of rows in A, and similarly for any type-1 column operation. This gives $\binom{m}{2}$ possible type 1 row operations, and $\binom{n}{2}$ possible type-1 column operations that can be applied to A and produce a matrix in M_Φ. type-2 row operations also cannot produce a matrix that has antiprogress 0 from A. Since any scalar in $\mathbb{C} - \{0\}$ may be applied to a single row of A, there exist $m\,|\mathbb{C} - \{0\}|$ possible type-2 row operations that can be applied to A. Finally, there are $|\mathbb{C} - \{0\}|$ ways to scale a given row of A, and $m - 1$ rows to which this scaled row may be added. Since this may be done

for any row of A, there exist $m(m-1)\,|\mathbb{C}-\{0\}|$ type-3 row operations that can be applied to A and can produce a matrix in M_Φ. Thus, for any $A \in M_\Phi$, there are exactly

$$m(m-1)\,|\mathbb{C}-\{0\}| + m\,|\mathbb{C}-\{0\}| + \binom{m}{2} + \binom{n}{2}$$

elementary operations that will not produce a matrix that has antiprogress 0 when applied to A. Thus, for each matrix produced by applying exactly these operations to Φ, we can produce another $m(m-1)\,|\mathbb{C}-\{0\}| + m\,|\mathbb{C}-\{0\}| + \binom{m}{2} + \binom{n}{2}$ that are still partition regular. Continuing in this fashion, we produce

$$\left[m(m-1)\,|\mathbb{C}-\{0\}| + m\,|\mathbb{C}-\{0\}| + \binom{m}{2} + \binom{n}{2}\right]^{\aleph_0}$$

infinite sequences of matrices contained in M_Φ.

Now, if

$$(\Phi, A_1, A_2, A_3, \ldots)$$

is an infinite sequence of matrices in M_Φ, where each A_i is produced by applying exactly one of the infinitely many operations described above to A_{i-1}, then there exist only a finite subset of these operations such that $A_i = A_j$ for some

$$A_j \in \{\Phi, A_1, \ldots, A_{i-1}\}.$$

Notice that only the identity operation may be applied to A_{i-1} to produce A_{i-1}. Then to see the result, assume there exists some elementary operation E_0 such that E_0 applied to A_{i-1} produces $A_i \in M_\Phi$ and

$$A_i \in \{\Phi, A_1, \ldots, A_{i-2}\}.$$

If E_0 is a type-1 row operation, there exists some

$$A_j \in \{\Phi, A_1, \ldots, A_{i-2}\}$$

such that switching exactly two rows of A_j produces A_{i-1}. Since there exist for each

$$A_j \in \{\Phi, A_1, \ldots, A_{i-2}\}$$

exactly two rows that may be switched to produce A_{i-1}, there exist at most $(i-1)$ possibilities for E_0 to be a type-1 row operation such that

$$A_i = E_0 A_{i-1} \in \{\Phi, A_1, \ldots, A_{i-2}\}.$$

Similarly, there exist at most $(i-1)$ possibilities for E_0 to be a type-1 column operation such that

$$A_i = A_{i-1} E_0 \in \{\Phi, A_1, \ldots, A_{i-2}\}.$$

In the case that E_0 is a type-2 row operation, and

$$E_0 A_{i-1} \in \{\Phi, A_1, \ldots, A_{i-2}\},$$

then we may produce some $A_j \in \{\Phi, A_1, \ldots, A_{i-2}\}$ by scaling exactly one row of A_{i-1} by a single $\alpha \in \mathbb{C}$. Since $\mathbb{C}$ is a unique factorization domain, it follows that if $E_0 A_{i-1} = A_j$, E_0 is unique. Therefore, at most, we may find one E_0 such that $E_0 A_{i-1} = A_j$ for each $A_j \in \{\Phi, A_1, \ldots, A_{i-2}\}$. Thus, there are at most $(i-1)$ type-2 column operations such that $A_i \in \{\Phi, A_1, \ldots, A_{i-2}\}$.

Finally, let E_0 be a type-3 column operation such that

$$E_0 A_{i-1} = A_j \in \{\Phi, A_1, \ldots, A_{i-2}\}.$$

Then exactly one row of A_j, call it row $(A_j)_k$, is not equal to row $(A_{i-1})_k$, and

$$(A_j)_k = \alpha (A_{i-1})_{k'} + (A_{i-1})_k,$$

where $\alpha \in \mathbb{C}$, and $(A_{i-1})_{k'}$ is one of the $(m-1)$ rows of A_{i-1} that is not row $(A_{i-1})_k$. It follows that

$$(A_j)_k - (A_{i-1})_k = \alpha (A_{i-1})_{k'},$$

and again $\alpha \in \mathbb{C}$ must be unique. Thus, for a given

$$A_j \in \{\Phi, A_1, \ldots, A_{i-2}\},$$

there are at most $(m-1)$ possibilities for E_0 such that $E_0 A_{i-1} = A_j$. Therefore, there are at most $(m-1)(i-1)$ possibilities for E_0 to be a type-3 column operation such that $A_i \in \{\Phi, A_1, \ldots, A_{i-2}\}$.

We may now conclude from these various cases that there are at most

$$3(i-1) + (m-1)(i-1) + 1$$

elementary operations that when applied to A_{i-1} will yield for A_i an element of $\{\Phi, A_1, \ldots, A_{i-1}\}$. Thus, for each A_i in the sequence there still exist an infinite number of elementary operations such that $A_{i+1} \in M_\Phi$ and $A_{i+1} \notin \{\Phi, A_1, \ldots, A_i\}$. It follows that there exist an infinite number of infinite filaments in M_Φ, for any partition regular matrix Φ. □

For any partition regular matrix, Φ, the space M_Φ is large. Moreover, for any two partition regular matrices, Φ and Ψ, the cardinality of the collection of infinite filaments in M_Φ is the same as the cardinality of the collection of infinite filaments in M_Ψ. This makes it difficult to use the size of these spaces to say that one matrix is "more" partition regular than another.

Lemma 2.7. *Let $A' \in M_\Phi$, and let $\kappa(A')$ be an infinite subfilament associated with A', such that A' is the only matrix on $\kappa(A')$ that may also be on a finite filament. Consider the map*

$$p : M_\Phi \longrightarrow M_\Phi$$

such that

$$p(A) = \begin{cases} A & \text{if } A \text{ is on a finite filament,} \\ \Phi & \text{if } A \text{ is on some } \kappa(A') \text{ for some } A' \in M_\Phi, \text{ and } A \neq A'. \end{cases}$$

Let $P_\Phi = Im(p)$, the image of p. Then the space P_Φ has the quotient topology induced by p.

Proof. It is clear p is a surjection. Furthermore, since M_Φ has the discrete topology, then if U is open in P_Φ, it must be that U is open in M_Φ. □

This new topological space consists of exactly those finite sequences of matrices that will allow Φ to escape the condition of partition regularity. Therefore, the sizes of these spaces offer a better characterization of the degree of partition regularity of a given matrix. Now consider the following corollary to Theorem 2.1.

Corollary 2.8. *Let $\mathscr{D}_\Phi = \{A \in P_\Phi : \text{apr}(A) = 1\}$. Then $(\mathscr{D}_\Phi, \delta)$ is a metric space, and $(\mathscr{D}_\Phi, \mathscr{T})$ is a discrete topological space.*

Proof. By Theorem 2.1 we know that

$$\delta : [\Phi] \text{ X } [\Phi] \longrightarrow \mathbb{R}$$

is a metric on $[\Phi]$. Since $\mathscr{D}_\Phi$ is a subset of $[\Phi]$, we know that

$$\delta : \mathscr{D}_\Phi \text{ X } \mathscr{D}_\Phi \longrightarrow \mathbb{R}$$

is a metric on $\mathscr{D}_\Phi$. Thus, $(\mathscr{D}_\Phi, \delta)$ is a metric space. Since every subset of $\mathscr{D}_\Phi$ is contained in M_Φ, then we know for every A_0 in $\mathscr{D}_\Phi$, there exists an open ball

$$\mathscr{B}_{A_0,1/2} = \{A_0\}.$$

Therefore, every singleton in $\mathscr{D}_\Phi$ is open, and we conclude that $(\mathscr{D}_\Phi, \mathscr{T})$ is a discrete topological space. □

Lemma 2.9. *If P_Φ is compact, then it is finite. Similarly, if $\mathscr{D}_\Phi$ is compact, then it is finite.*

Proof. We will demonstrate the result for P_Φ. The proof works analogously for $\mathscr{D}_\Phi$. Let P_Φ be compact. Since the collection $\mathscr{F}$ consisting of all singletons in P_Φ forms an open cover of P_Φ, there exists a finite subcover $\mathscr{F}'$ contained in $\mathscr{F}$. Assume that $\mathscr{F}'$ is a proper subcollection of $\mathscr{F}$. Then the set $\mathscr{Q} = \{A \in P_\Phi : \{A\} \in \mathscr{F}'\}$ has power set $\mathscr{P}(\mathscr{Q})$ equal to the set of all sets that may be formed by taking the union of elements of $\mathscr{F}'$. Similarly, $\mathscr{R} = \{A \in P_\Phi : \{A\} \in \mathscr{F}\}$ has power set $\mathscr{P}(\mathscr{R})$

equal to the set of all sets that may be formed by taking the union of elements of $\mathcal{F}$. Since $\mathcal{F}'$ is a proper subcollection of $\mathcal{F}$, it must be that $\mathcal{P}(\mathcal{Q})$ is a proper subcollection of $\mathcal{P}(\mathcal{R})$. Thus, we can choose $\mathcal{U} \in \mathcal{P}(\mathcal{R}) \setminus \mathcal{P}(\mathcal{Q})$ that is not equal to the empty set. Then

$$\mathcal{U} = \bigcup_{B \in \mathcal{U}} \{B\},$$

and so any subset $\{B\}$ is not in $\mathcal{P}(\mathcal{Q})$. Thus, the matrix B cannot be in any open set contained in $\mathcal{F}'$, a contradiction. Therefore, $\mathcal{F}' = \mathcal{F}$, and consequently, $\mathcal{F}$ is a finite open cover of P_Φ. Since $\mathcal{F}$ is the collection of all singletons in P_Φ, we conclude that P_Φ must be finite. □

Notice that for any Φ in $\mathbb{M}_{m,n}(\mathbb{C})$, the topological spaces M_Φ, P_Φ, and $\mathcal{D}_\Phi$ are all *Hausdorff*. We can think of $\mathcal{D}_\Phi$ as the boundary set of P_Φ, since no matrix in P_Φ can be "closer" to leaving P_Φ than those matrices with antiprogress 1. This relationship between P_Φ and $\mathcal{D}_\Phi$ grants us the following theorem.

Theorem 2.10. *For any partition regular matrix Φ, the quotient space P_Φ is compact if and only if $\mathcal{D}_\Phi$ is finite.*

Proof. Necessity of the statement follows from Lemma 2.9. To demonstrate sufficiency, recall that P_Φ consists of only finite filaments. Every A in $\mathcal{D}_\Phi$ is the last matrix of a finite filament. For such an A, pick P and Q such that $A = P\Phi Q$, and represent P and Q with the finite sequence of elementary operations $\langle x \rangle_{i=1}^{l}$. Then we can think of the finite filament ending in A as the finite, ordered set of matrices

$$\left(\Phi, \langle x \rangle_{i=1}^{1}(\Phi), \langle x \rangle_{i=1}^{2}(\Phi), \ldots, \langle x \rangle_{i=1}^{l-1}(\Phi), A\right),$$

where $\langle x \rangle_{i=1}^{n}(\Phi)$ represents the matrix produced by applying the first n operations of $\langle x \rangle_{i=1}^{l}$ to Φ. If we let

$$\langle x \rangle_A(\Phi) = \left(\Phi, \langle x \rangle_{i=1}^{1}(\Phi), \langle x \rangle_{i=1}^{2}(\Phi), \ldots, \langle x \rangle_{i=1}^{l-1}(\Phi), A\right),$$

then

$$P_\Phi = \bigcup_{A \in \mathcal{D}_\Phi} \langle x \rangle_A(\Phi).$$

It follows that P_Φ contains a finite number of matrices if and only if $\mathcal{D}_\Phi$ is finite. Now let $\mathbb{X}$ denote the set of all open sets in P_Φ. Since every singleton is open in P_Φ, we know that $\mathbb{X}$ is finite if and only if $\mathcal{D}_\Phi$ is finite.

Assume that $\mathcal{D}_\Phi$ is finite. Let $\mathcal{F}$ be an open cover of P_Φ (without duplicates), and assume that $\mathcal{F}$ is not finite. Then $|\mathbb{X}| < |\mathcal{F}|$. Thus, $\mathcal{F}$ must contain more open sets of P_Φ than are in the set $\mathbb{X}$, a contradiction. We conclude that $\mathcal{F}$ must be finite, and since any open cover $\mathcal{F}$ is a subcover of itself, then it must be that every open cover of P_Φ contains a finite subcover. Thus, P_Φ is compact. □

Corollary 2.11. *$\mathcal{D}_\Phi$ is compact if and only if $\mathcal{D}_\Phi$ is finite.*

Proof. Necessity of the statement again follows from Lemma 2.9. So assume that $\mathcal{D}_\Phi$ is finite. Then, P_Φ is compact. Since $P_\Phi \setminus \mathcal{D}_\Phi$ is open in P_Φ, then $\mathcal{D}_\Phi$ is closed. Therefore, $\mathcal{D}_\Phi$ is a closed subset of a compact space, and we conclude that $\mathcal{D}_\Phi$ is compact. □

Theorem 2.12. *Let Φ be a partition regular matrix, and let the set $\mathcal{D}_\Phi$ be infinite. Then P_Φ is the union of an infinite number of disjoint, compact subspaces.*

Proof. Let $\mathcal{G} = \{G_i : i \in \mathbb{N}\}$ be a partition of $\mathcal{D}_\Phi$ into nonempty, disjoint and finite subsets. Since each subset is finite and $\mathcal{D}_\Phi$ contains an infinite number of elements, there must exist an infinite number of subsets in $\mathcal{G}$. Now let $B_1 \subset P_\Phi$ be the set of all matrices on a filament that terminates with a matrix contained in G_1. Then for $i > 1$, let $B_i \subset P_\Phi$ be the set of all matrices on a filament that terminates with a matrix contained in G_i, but are not contained in $\bigcup_{j=1}^{i-1} B_j$.

If a filament ends in a matrix $A \in G_i$, then there exists a pair of invertible matrices P and Q, such that $A = P\Phi Q$, that can be represented by a finite sequence of elementary operations $\langle x \rangle_{i=1}^{l}$. Thus, the filament terminating with A may be written as the finite, ordered set of matrices

$$\langle x \rangle_A(\Phi) = \left(\Phi, \langle x \rangle_{i=1}^{1}(\Phi), \langle x \rangle_{i=1}^{2}(\Phi), \ldots, \langle x \rangle_{i=1}^{l-1}(\Phi), A\right),$$

where $\langle x \rangle_{i=1}^{n}(\Phi)$ represents the matrix produced by applying the first n operations of $\langle x \rangle_{i=1}^{l}$ to Φ. Then, for all $i \in \{1, 2, 3, \ldots\}$, the set

$$\bigcup_{A \in G_i} \langle x \rangle_A(\Phi)$$

is the union of a finite number of finite sets, and therefore is also finite. Consequently, each B_i is finite for all i since,

$$B_i \subseteq \bigcup_{A \in G_i} \langle x \rangle_A(\Phi).$$

Now let $\mathbb{X}$ be the set of all open sets in B_i. Since B_i is finite and has the discrete topology, $\mathbb{X} = \mathcal{P}(B_i)$, and is also finite. Assume that $\mathcal{F}$ is an open cover of B_i that contains an infinite number of open sets. Since $\mathcal{F}$ is a collection of open sets of B_i, it must be that $\mathcal{F} \subseteq \mathbb{X}$. Thus, $|\mathcal{F}| \leq |\mathbb{X}|$, which contradicts the finite size of $\mathbb{X}$. So $\mathcal{F}$ must be finite, and every open cover of B_i is finite. Thus, every open cover of B_i contains a finite subcover, namely itself. Therefore, B_i is a compact subspace of P_Φ. Since $\{B_i : i \in \mathbb{N}\}$ is an infinite set of disjoint subspaces of P_Φ and $P_\Phi = \bigcup_{i \in \mathbb{N}} B_i$, it follows that P_Φ is the union of an infinite number of disjoint, compact subspaces. □

3. Partition regular topological surfaces and graphs

In this section, we will create a topological space that has geometry describing the degree of partition regularity for an associated topological surface, and then we will show how these spaces may also be created for an arbitrary, finite graph.

Every topological surface $\mathcal{S}$ has a triangulation, and $\mathcal{S}$ is compact if and only if it has a triangulation consisting of a finite number of triangles. So let $\mathcal{S}$ be a topological surface, and let $T(\mathcal{S})$ be a triangulation of $\mathcal{S}$. Then, the set of all vertices and edges in $T(\mathcal{S})$ form a connected graph G. Let $\{\Gamma_i : i \in \mathbb{N}\}$ be the collection of all orientations of G.

Proposition 3.1. *The collection $\{\Gamma_i : i \in \mathbb{N}\}$ is finite if and only if the surface $\mathcal{S}$ is compact.*

Proof. For necessity, let $\{\Gamma_i : i \in \mathbb{N}\}$ be finite and assume $\mathcal{S}$ is not compact. Then $T(\mathcal{S})$ does not consist of a finite number of triangles. Thus, there exists an infinite number of edges in G, each of which may be assigned one of two directions. Let Γ_0 be in $\{\Gamma_i : i \in \mathbb{N}\}$ with $\mathcal{K}$ as the edge index set. Now, let Γ_n be the orientation of G obtained by reversing only the direction of edge n in Γ_0. Then the set $\{\Gamma_n : i \in \mathcal{K}\}$ is infinite. However,

$$\{\Gamma_n : i \in \mathcal{K}\} \subset \{\Gamma_i : i \in \mathbb{N}\},$$

which gives a contradiction.

For sufficiency, notice that if $\mathcal{S}$ is compact, then $T(\mathcal{S})$ contains a finite number of triangles. Therefore, E, the set of all edges in G, is a finite set. Since each edge may have one of two directions, then $|\{\Gamma_i : i \in \mathbb{N}\}| = 2^{|E|}$. □

Let G be a finite graph and Γ_i an orientation of G. We know that D_{Γ_i}, the oriented incidence matrix of Γ_i, is partition regular if and only if Γ_i is strongly connected [Hogben and McLeod 2010, Theorem 2.4]. Here we consider the subcollection $\mathcal{C}$ of $\{\Gamma_i : i \in \mathbb{N}\}$ consisting of all strongly connected orientations of G.

Theorem 3.2. *The collection $\mathcal{C}$, of all strongly connected orientations of G, is a nonempty and proper subcollection of $\{\Gamma_i : i \in \mathbb{N}\}$, for any triangulation of any topological surface $\mathcal{S}$.*

Proof. Let $T(\mathcal{S})$ be a triangulation of some topological surface $\mathcal{S}$. Then the graph G consisting of all vertices and edges in $T(\mathcal{S})$ is a connected graph. It is well known that a graph G has a strongly connected orientation if and only if the edge connectivity of G is greater than or equal to 2. Therefore, G associated with $T(\mathcal{S})$ will have a strongly connected orientation if and only if it does not have edge connectivity equal to 1. We know that an edge $\{i, j\}$ of G must be an edge of at least one triangle. Now let

$$w = \big\{v_1, \{1, 2\}, v_2, \ldots, v_i, \{i, j\}, v_j, \ldots, \{n-1, n\}, v_n\big\}$$

be a walk on G that uses edge $\{i, j\}$. If we remove edge $\{i, j\}$, then we can define the walk

$$w' = \big\{v_1, \{1, 2\}, v_2, \ldots, v_i, \{i, k\}, v_k, \{k, j\}, v_j, \ldots, \{n-1, n\}, v_n\big\},$$

where v_k is the third vertex in some triangle containing edge $\{i, j\}$. Thus, $G\backslash\{\{i, j\}\}$ is still connected, and consequently, G cannot have edge connectivity equal to 1. Therefore, there exists a strongly connected orientation of G, and $\mathscr{C}$ is nonempty.

To see that $\mathscr{C} \neq \{\Gamma_i : i \in \mathbb{N}\}$, notice that for any vertex v_i of G, there exists an orientation Γ_i in $\{\Gamma_i : i \in \mathbb{N}\}$ such that any edge connected to v_i has v_i as its head. Thus, Γ_i cannot be strongly connected, and $\mathscr{C}$ must be a proper subcollection of $\{\Gamma_i : i \in \mathbb{N}\}$. □

Therefore, if $\mathcal{S}$ is compact, then for each Γ_i in $\mathscr{C}$, D_{Γ_i} is partition regular. Furthermore, for each D_{Γ_i} we can create the associated quotient space $P_{D_{\Gamma_i}} = P_{\Gamma_i}$.

Definition 3.3. Given a compact topological surface $\mathcal{S}$, let $\mathfrak{T}$ be the set of all triangulations of $\mathcal{S}$. For $t \in \mathfrak{T}$, let $\mathscr{C}_t$ be the set of all strongly connected orientations of the graph associated with t. Then $\mathcal{S}$ is a *partition regular surface* if the product space

$$\prod_{t \in \mathfrak{T}} \Big(\prod_{\Gamma_i \in \mathscr{C}_t} P_{\Gamma_i} \Big)$$

is nonempty.

Theorem 3.4. *Let $\mathcal{S}$ be a compact topological surface. Then $\mathcal{S}$ is partition regular.*

Proof. Recall that the set $\mathscr{C}_t$ is nonempty for any triangulation t, of any surface $\mathcal{S}$. Thus, there is at least one quotient space for each distinct triangulation of $\mathcal{S}$ in the product topology associated with $\mathcal{S}$. Consequently, the product topology associated with $\mathcal{S}$ is not an empty space, and thus $\mathcal{S}$ is partition regular. □

Definition 3.5. For any finite graph G, let $\mathscr{C}$ be the set of all strongly connected orientations of G. Then G is a *partition regular graph* if the product space

$$\prod_{\Gamma_i \in \mathscr{C}_t} P_{\Gamma_i}$$

is nonempty.

In contrast to Theorem 3.4, the following theorem shows that not all finite graphs are partition regular.

Theorem 3.6. *Let G be a finite tree. Then G is not a partition regular graph.*

Proof. Since no orientation of a tree graph is strongly connected, then every quotient space in the product topology associated with a tree graph is empty. Consequently, any such product topology is empty, and no tree graph is a partition regular graph. □

4. Conclusion

As is captured in the contrasting scenarios presented in Theorems 3.4 and 3.6, partition regularity may be thought of as a property with varying degree that is dependent on the object being studied. For instance, we began with a compact topological surface $\mathcal{S}$, traced the notion of order in the context of topological surfaces, through the graph theoretical context, and finally the matrix theoretical context. Consequently, we were able to construct topological spaces characterizing the degree of order for the surface $\mathcal{S}$. We have also seen that every compact topological surface is a partition regular surface, and thus exhibits, as should be expected, some level of order. We may now explore the concept of partition regularity for compact topological surfaces and finite graphs. Moreover, we now possess structures that allow us to no longer think of an object as simply being partition regular, but instead, as having some degree of partition regularity. Subsequently, we may begin to relate matrices, graphs, and topological surfaces based on their relative degrees of partition regularity.

Acknowledgements

The author thanks the 2010 NSF REU Program at Mount Holyoke College, and Professor Jillian McLeod of Mount Holyoke College for her critical review of the material in this paper.

The author also thanks the faculty and staff of the Oberlin College Department of Mathematics.

References

[Hindman 2007] N. Hindman, "Partition regularity of matrices", pp. 265–298 in *Combinatorial number theory* (Carrollton, GA, 2005), edited by B. Landman et al., de Gruyter, Berlin, 2007. Also available as article #A18 in *Integers Elec. J. Combin. Number Theory* **7**:2 (2007), accessible from http://www.integers-ejcnt.org/vol7-2.html. MR 2008g:05216 Zbl 1125.05105

[Hogben and McLeod 2010] L. Hogben and J. McLeod, "A linear algebraic view of partition regular matrices", *Lin. Alg. Appl.* **433** (2010), 1809–1820. Zbl 05811008

[Rado 1943] R. Rado, "Note on combinatorial analysis", *Proc. London Math. Soc.* **48** (1943), 122–160. MR 5,87a Zbl 0028.33801

Received: 2010-08-02 Revised: 2010-12-21 Accepted: 2010-12-22

lsolus@oberlin.edu *Department of Mathematics, Oberlin College, OCMR 2293, 135 W Lorain Street, Oberlin, OH 44074, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.435

Energy-minimizing unit vector fields

Yan Digilov, William Eggert, Robert Hardt, James Hart,
Michael Jauch, Rob Lewis, Conor Loftis, Aneesh Mehta,
Esther Perez, Leobardo Rosales, Anand Shah and Michael Wolf

(Communicated by Frank Morgan)

Given a surface of revolution with boundary, we study the extrinsic energy of smooth tangent unit-length vector fields. Fixing continuous tangent unit-length vector fields on the boundary of the surface of revolution, we ask if there is a unique smooth tangent unit-length vector field continuously achieving the boundary data and minimizing energy amongst all smooth tangent unit-length vector fields also continuously achieving the boundary data.

1. Introduction

Let $\mathcal{S}$ be a surface of revolution given by the parametrization

$$\Phi(\theta, t) = (r(t)\cos\theta, r(t)\sin\theta, t), \quad \theta \in \mathbb{R}, t \in (0, h),$$

where $r(t) \in C^\infty([0, h])$ is positive in $[0, h]$. Let $\mathfrak{X}^1(\mathcal{S})$ be the set of smooth tangent unit-length vector fields on $\mathcal{S}$. For $\boldsymbol{V} \in \mathfrak{X}^1(S)$ we define the *extrinsic* energy of $\boldsymbol{V}$ to be

$$E(\boldsymbol{V}) = \iint_S |D\boldsymbol{V}|^2 \, d\,\text{Area},$$

where $D\boldsymbol{V}$ is the differential of the map $\boldsymbol{V} : \mathcal{S} \to \mathbb{R}^3$. Using the parametrization Φ, we get

$$E(\boldsymbol{V}) = \int_0^h \int_0^{2\pi} \left(\frac{r(t)}{\sqrt{1 + r'(t)^2}} \right) \left| \frac{\partial \boldsymbol{V}}{\partial t} \right|^2 + \left(\frac{\sqrt{1 + r'(t)^2}}{r(t)} \right) \left| \frac{\partial \boldsymbol{V}}{\partial \theta} \right|^2 d\theta \, dt.$$

Suppose V_0 and V_h are continuous unit-length tangent vector fields, defined respectively on $\{\Phi(\theta, 0) : \theta \in \mathbb{R}\}$ and $\{\Phi(\theta, h) : \theta \in \mathbb{R}\}$. For $\boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S})$, we

MSC2000: primary 53A05; secondary 49Q99.
Keywords: calculus of variations, energy, first variation, vector fields, surfaces of revolution.
This research was supported by National Science Foundation grant DMS-0739420.

write $\boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h$ if $\boldsymbol{V}$ continuously achieves the boundary data V_0, V_h on $\mathcal{S}$. Precisely, $\boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h$ if for every $\vartheta \in \mathbb{R}$ we have

$$\lim_{(\theta,t)\to(\vartheta,0)} \boldsymbol{V}(\Phi(\theta,t)) = V_0(\Phi(\vartheta,0)), \qquad \lim_{(\theta,t)\to(\vartheta,h)} \boldsymbol{V}(\Phi(\theta,t)) = V_h(\Phi(\vartheta,h)).$$

We pose the following question: *Suppose V_0 and V_h are continuous unit-length tangent vector fields defined respectively on $\{\Phi(\theta,0):\theta\in\mathbb{R}\}$ and $\{\Phi(\theta,h):\theta\in\mathbb{R}\}$. Does there exist a unique $\boldsymbol{V}\in\mathfrak{X}^1(\mathcal{S})$ with $\boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h$ so that $E(\boldsymbol{V}) < E(\tilde{\boldsymbol{V}})$ for any other $\tilde{\boldsymbol{V}}\in\mathfrak{X}^1(\mathcal{S})$ with $\tilde{\boldsymbol{V}}|_{\partial\mathcal{S}} = V_0, V_h$?*

We give partial answers to the question of existence and uniqueness. Theorem 3.2 shows the existence of minimizers for a certain class of boundary data, and Theorem 4.1 allows us to conclude uniqueness in a parametric sense in general, and outright for the case of the unit cylinder with horizontal boundary data (see Corollary 5.2). Only first-year graduate analysis is needed for most of the results, although some references to regularity of weak solutions to ordinary differential equations and approximations by smooth functions in $W^{1,2}$ is mentioned in the proofs of Theorem 3.2 and Theorem 5.1.

We describe the effect of the shape of $\mathcal{S}$ on the minimizer. Observe that where $r'(t)$ is large $\partial V/\partial t$ can be large in magnitude without paying much in energy. Hence, we can seek to minimize energy by letting $\boldsymbol{V}$ not vary much from the boundary data near $t = 0, h$, and then where $r'(t)$ is large we let $\boldsymbol{V}$ quickly change to a vector field of low energy. In the case of the unit cylinder, Figure 2 (page 448) shows that it is best to steadily homotopy between the boundary data. However, for the surface given by $r(t) = \sin t + 2$ (Figure 3, right), it is better to homotopy to a vector field with low energy in the regions where $r'(t)$ is large, as suggested by Figure 3, left. This illustrates that the $\partial\boldsymbol{V}/\partial t$ term is important, and so we list t-derivatives first in our calculations.

In case of the cylinder $r(t) = 1$ with height h, replacing $D\boldsymbol{V}$ with the covariant derivative of $\boldsymbol{V}$ leaves us to study 2π-periodic harmonic functions defined over $\mathbb{R}\times(0,h)$. In general, intrinsic energy of unit vector fields is also called *total bending*, and has been studied in the more general setting of Riemannian manifolds of any dimension; see [Wiegmink 1995] for an introduction. In [Borrelli et al. 2003], for example, it is shown that the infimum intrinsic energy in the odd-dimensional sphere S^{2k+1} for $k\geq 2$ is given by the energy of the horizontal tangent unit vector field defined on S^{2k+1} except at two antipodal points $\{P,-P\}$. This value, however, is not attained by any smooth tangent unit vector field over S^{2k+1} as shown in [Brito and Walczak 2000].

Minimizing the extrinsic energy over all smooth vector fields can be studied using similar techniques, as will follow. Although the set of vector fields over which we must minimize is larger, we avoid the difficulties arising in the unit-length

case by the necessity to work with the angle functions φ introduced in Section 2. Instead, denoting by $\boldsymbol{V}$ a tangent vector field on $\mathcal{S}$ and using the parametrization $\boldsymbol{V} = a(\theta, t)\Phi_t + b(\theta, t)\Phi_\theta$, we work directly with the smooth functions a, b in the general case.

2. First variation

We derive a partial differential equation which a minimizing $\boldsymbol{V}$ must solve, using a standard technique from the calculus of variations. First, given $\boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S})$ we can find a function $\varphi(\theta, t)$ so that

$$\boldsymbol{V}(\theta, t) = \begin{pmatrix} -\sin\theta\cos\varphi \\ \cos\theta\cos\varphi \\ 0 \end{pmatrix} + \frac{1}{\sqrt{1+r'(t)^2}} \begin{pmatrix} r'(t)\cos\theta\sin\varphi \\ r'(t)\sin\theta\sin\varphi \\ \sin\varphi \end{pmatrix}.$$

Thus, $\varphi(\theta, t)$ measures the angle between $\boldsymbol{V}(\theta, t)$ and the horizontal tangent vector field $(-\sin\theta, \cos\theta, 0)$. Our choice of angle function φ is not unique, and may be chosen to be discontinuous. This occurs for example in the proof of Theorem 5.1. Choosing φ continuous may require us to make $|\varphi|$ large. However, $\sin\varphi$, $\cos\varphi$, and $\sin 2\varphi$ will be smooth in $\mathbb{R}\times(0, h)$, continuous even at $t = 0, h$, and independent of φ. Using smoothness of $\sin\varphi$, $\cos\varphi$ we can define φ_t, φ_θ smooth in $\mathbb{R} \times (0, h)$ and independent of φ. Whenever $\boldsymbol{V}$ is given by an angle function φ, we shall write $\boldsymbol{V} = \boldsymbol{V}(\varphi)$.

For $\boldsymbol{V} = \boldsymbol{V}(\varphi)$, we can write the energy $E(\boldsymbol{V}) = E(\varphi)$ in terms of φ:

$$E(\varphi) = \int_0^h \int_0^{2\pi} T(t)(\varphi_t)^2 + \Theta(t)(\varphi_\theta)^2 \, d\theta\, dt + \int_0^h \int_0^{2\pi} P_c(t)\cos^2\varphi + P_s(t)\sin^2\varphi + Q(t)\, d\theta\, dt \quad (2\text{-}1)$$

where

$$T(t) = \frac{r(t)}{\sqrt{1+r'(t)^2}}, \qquad \Theta(t) = \frac{1+r'(t)^2(3+3r'(t)^2+r'(t)^4)}{r(t)(1+r'(t)^2)^{5/2}},$$

$$P_c(t) = \frac{1+4r'(t)^2+2r'(t)^4}{r(t)(1+r'(t)^2)^{5/2}}, \qquad P_s(t) = \frac{2r(t)^2 r''(t)^2}{r(t)(1+r'(t)^2)^{5/2}},$$

$$\text{and } Q(t) = \frac{r'(t)^2}{r(t)\sqrt{1+r'(t)^2}}.$$

If $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ minimizes energy on $\mathcal{S}$ with respect to the boundary data V_0, V_h, then let $\eta \in C_c^\infty((0, 2\pi) \times (0, h))$ (that is, a smooth function with compact support in $(0, 2\pi) \times (0, h)$). We then let $\boldsymbol{V}^s \in \mathfrak{X}^1(\mathcal{S})$ be the vector field given by the angle

function $\varphi + s\eta$. Then $E(\boldsymbol{V}^s)$ achieves a minimum at $s = 0$, and so

$$\frac{d}{ds} E(\boldsymbol{V}^s)\Big|_{s=0} = 0.$$

Differentiating (2-1) under the integral with respect to s gives:

$$\int_0^h \int_0^{2\pi} 2T(t)\varphi_t \eta_t + 2\Theta(t)\varphi_\theta \eta_\theta - ((P_c(t) - P_s(t)) \sin 2\varphi)\eta \, d\theta \, dt = 0.$$

Since η has compact support in $(0, 2\pi) \times (0, h)$, we may use integration by parts in the first and second terms to get

$$\int_0^h \int_0^{2\pi} [-2(T(t)\varphi_t)_t - 2(\Theta(t)\varphi_\theta)_\theta - (P_c(t) - P_s(t)) \sin 2\varphi]\eta \, d\theta \, dt = 0.$$

We thus have that φ must satisfy the second-order partial differential equation:

$$(T(t)\varphi_t)_t + (\Theta(t)\varphi_\theta)_\theta + (P_c(t) - P_s(t))\Big(\frac{\sin 2\varphi}{2}\Big) = 0, \tag{2-2}$$

which we call the *Euler–Lagrange equation* associated to the energy $E(\varphi)$.

In case of the cylinder $\mathcal{C}$ with $r(t) = 1$ and height h, the energy (2-1) becomes

$$E(\varphi) = \int_0^h \int_0^{2\pi} (\varphi_t)^2 + (\varphi_\theta)^2 + \cos^2 \varphi \, d\theta \, dt.$$

Equation (2-2) in this case is

$$\varphi_{tt} + \varphi_{\theta\theta} + \frac{\sin 2\varphi}{2} = 0,$$

for which the only constant solutions are $\varphi = k\pi/2$ with $k \in \mathbb{Z}$. Although when k is odd $E(k\pi/2) = 0$, we can show by example that for large h the horizontal vector field $\varphi = \pi$ is not a minimizer. Corollary 5.2 will show that for $h < \sqrt{8}$ the horizontal vector field is a minimizer, and it remains to find the largest h_0 so that this true for all $h < h_0$.

The equation

$$\varphi_{tt} + \varphi_{\theta\theta} + \frac{\sin 2\varphi}{2} = 0$$

is a special case of a form of equations called the *sine-Gordon equations*, which arise in differential geometry and various areas of physics. This particular form arises in the study of ferromagnetics in physics; see [Chen et al. 2004] for example, and in the study of harmonic maps in differential geometry, see [Hu 1982].

3. Existence

In this section we aim to prove the existence of minimizers with boundary data V_0,V_h which make a constant angle with the horizontal vector field $(-\sin\theta, \cos\theta, 0)$.

Lemma 3.1. *Suppose V_0, V_h are continuous tangent unit-length boundary data on $\partial\mathcal{S}$ such that each can be written using a constant angle function. Let $\boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S})$ with $\boldsymbol{V}|_{\partial\mathcal{S}} = V_0$, V_h and $\boldsymbol{V} = \boldsymbol{V}(\varphi)$. If $\varphi_\theta \not\equiv 0$, then there is a vector field $\tilde{\boldsymbol{V}} \in \mathfrak{X}^1(\mathcal{S})$ with $\tilde{\boldsymbol{V}}|_{\partial\mathcal{S}} = V_0$, V_h so that $E(\tilde{\boldsymbol{V}}) < E(\boldsymbol{V})$, and so that we can write $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}(\tilde{\varphi})$ where $\tilde{\varphi} \in C([0,h]) \cap C^\infty((0,h))$ and $\tilde{\varphi}(0) \in [0, 2\pi)$.*

Proof. Suppose $E(\boldsymbol{V}) < \infty$, otherwise we simply take $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}(\varphi_0 + (t/h)(\varphi_h - \varphi_0))$ where $\varphi_0 \in [0, 2\pi)$, φ_h are constants so that the boundary data $V_0 = V_0(\varphi_0)$ and $V_h = V_h(\varphi_h)$. Let $\boldsymbol{V} = \boldsymbol{V}(\varphi)$, we thus have $\int_0^h \int_0^{2\pi} \Theta(t)(\varphi_\theta)^2 \, d\theta \, dt > 0$. Consider the integrable function

$$f(\theta) = \int_0^h T(t)(\varphi_t)^2 + P_c(t)\cos^2\varphi + P_s(t)\sin^2\varphi + Q(t)\, dt.$$

We then have $\inf_{\theta\in[0,2\pi)} f(\theta) < \infty$. Choose $\theta_0 \in [0, 2\pi)$ so that

$$f(\theta_0) < \inf_{\theta\in[0,2\pi)} f(\theta) + \frac{1}{2\pi}\int_0^h \int_0^{2\pi} \Theta(t)(\varphi_\theta)^2 \, d\theta \, dt.$$

Define $\tilde{\varphi}(\theta, t) = \varphi(\theta_0, t)$, and let $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}(\tilde{\varphi}) \in \mathfrak{X}^1(S)$. Evidently $\tilde{\boldsymbol{V}}|_{\partial\mathcal{S}} = V_0$, V_h and

$$E(\boldsymbol{V}) = \int_0^{2\pi} f(\theta)\, d\theta + \int_0^h \int_0^{2\pi} \Theta(t)(\varphi_\theta)^2 \, d\theta \, dt > \int_0^{2\pi} f(\theta_0)\, d\theta = E(\tilde{\varphi}) = E(\tilde{\boldsymbol{V}}).$$

Since $\tilde{\varphi}$ only depends on t, we can redefine $\tilde{\varphi}$ so that

$$\tilde{\varphi} \in C([0,h]) \cap C^\infty((0,h)).$$

We can also translate by some $2\pi k$ with $k \in \mathbb{Z}$, without changing the energy $E(\tilde{\varphi})$, so that $\tilde{\varphi}(0) \in [0, 2\pi)$. □

Theorem 3.2. *Suppose $V_0 = V_0(\varphi_0)$, $V_h = V_h(\varphi_h)$ are continuous tangent unit-length boundary data on $\partial\mathcal{S}$, where φ_0, φ_h are constants. Then there exists*

$$\boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S}), \qquad \text{with } \boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h,$$

minimizing energy. Moreover,

$$\boldsymbol{V} = \boldsymbol{V}(\varphi), \qquad \text{with } \varphi \in C^\infty([0,h]).$$

Proof. The argument follows the proof of the existence of minimizers to the Dirichlet energy using weak compactness [Evans 1998, Section 8.2]. Let

$$E = \inf\{E(\boldsymbol{V}) : \boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S}), \boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h\},$$

note that $E < \infty$. Define C_{E+1} to be the set of $\varphi \in C([0,h]) \cap C^\infty((0,h))$ with energy $E(\varphi) \le E+1$ and $\varphi(0) \in [0, 2\pi)$ so that $\boldsymbol{V}(\varphi)|_{\partial\mathcal{S}} = V_0, V_h$. By Lemma 3.1 it suffices to show $E = \inf_{\varphi\in C_{E+1}} E(\varphi)$ is attained. Let $\overline{C}_{E+1}$ be the closure of C_{E+1} in $L^2([0,h])$.

Lemma 3.3. *Every $\overline{\varphi} \in \overline{C}_{E+1}$ is continuous in $[0,h]$ with a weak derivative in $L^2([0,h])$. Moreover, we can find a sequence $\varphi_k \in C_{E+1}$ converging uniformly to $\overline{\varphi}$.*

Proof. Take a sequence $\varphi_k \in C_{E+1}$. Let $T_{\min} = \min_{t\in[0,h]} T(t)$. It follows that the φ_k are equicontinuous in $[0,1]$, since by Cauchy–Schwartz

$$\begin{aligned}|\varphi_k(x) - \varphi_k(y)| &= \left|\int_x^y (\varphi_k)_t\,dt\right| \le \sqrt{|x-y|}\Big(\int_0^h ((\varphi_k)_t)^2\,dt\Big)^{1/2}\\ &= \sqrt{|x-y|}\Big(\int_0^h \frac{T(t)}{T_{\min}}\cdot((\varphi_k)_t)^2\,dt\Big)^{1/2} \le \sqrt{\frac{E+1}{T_{\min}}}\cdot\sqrt{|x-y|}.\end{aligned}$$

Since $0 \le \varphi_k(0) < 2\pi$, there is by Arzelà–Ascoli a subsequence of the φ_k having a uniformly convergent subsequence. Therefore $\overline{C}_{E+1} \subseteq C([0,h])$.

Let $\eta \in C_c^\infty((0,1))$ and $\overline{\varphi} \in \overline{C}_{E+1}$ with $\varphi_k \in C_{E+1}$ converging uniformly to $\overline{\varphi}$. Then

$$\int_0^h \overline{\varphi}\eta_t\,dt = \lim_{k\to\infty}\int_0^h \varphi_k\eta_t\,dt = -\lim_{k\to\infty}\int_0^h (\varphi_k)_t\eta\,dt.$$

However, note that the sequence $(\varphi_k)_t$ is a bounded sequence in $L^2([0,h])$. By Alaoglu's theorem, a subsequence of the $(\varphi_k)_t$ converges weakly to some

$$\overline{\varphi}_t \in L^2([0,h]).$$

We therefore have

$$\int_0^h \overline{\varphi}\eta_t\,dt = -\int_0^h \overline{\varphi}_t\eta\,dt,$$

and so $\overline{\varphi}$ has weak derivative $\overline{\varphi}_t$ in $L^2([0,h])$. □

Returning to the proof of Theorem 3.2, given $\overline{\varphi} \in \overline{C}_{E+1}$ we can define the energy $E(\overline{\varphi})$ by

$$E(\overline{\varphi}) = 2\pi\int_0^h T(t)(\overline{\varphi}_t)^2 + P_c(t)\cos^2\overline{\varphi} + P_s(t)\sin^2\overline{\varphi} + Q(t)\,dt,$$

where $\overline{\varphi}_t$ is the weak derivative in $L^2([0,h])$ of $\overline{\varphi}$. Also define

$$E_{\overline{C}_{E+1}} = \inf_{\overline{\varphi}\in\overline{C}_{E+1}} E(\overline{\varphi}),$$

so that $E_{\overline{C}_{E+1}} \le E$.

We show there is a $\overline{\varphi}\in\overline{C}_{E+1}$ with $E(\overline{\varphi}) = E_{\overline{C}_{E+1}}$. Take a sequence $\overline{\varphi}_k \in \overline{C}_{E+1}$ so that $E(\overline{\varphi}_k) \searrow E_{\overline{C}_{E+1}}$. The sequence $\overline{\varphi}_k$ will also be equicontinuous with $\overline{\varphi}_k(0) \in [0, 2\pi)$, and hence a subsequence will converge uniformly to some $\overline{\varphi} \in \overline{C}_{E+1}$. Arguing as in Lemma 3.3, we can show $(\overline{\varphi}_k)_t \to \overline{\varphi}_t$ weakly in $L^2([0,h])$, and since $T(t)$ is bounded in $[0,h]$, we have $T(t)^{\frac{1}{2}}(\overline{\varphi}_k)_t \to T(t)^{\frac{1}{2}}\overline{\varphi}_t$ weakly in $L^2([0,h])$ as well. From this it follows that

$$\int_0^h T(t)(\overline{\varphi}_t)^2\,dt \le \varliminf_{k\to 0} \int_0^h T(t)((\overline{\varphi}_k)_t)^2\,dt,$$

and since $\overline{\varphi}_k \to \overline{\varphi}$ uniformly, we can show

$$\int_0^h P_c(t)\cos^2\overline{\varphi}_k + P_s(t)\sin^2\overline{\varphi}_k\,dt \to \int_0^h P_c(t)\cos^2\overline{\varphi} + P_s(t)\sin^2\overline{\varphi}\,dt.$$

We therefore have $E(\overline{\varphi}) \le \lim_{k\to\infty} E(\overline{\varphi}_k) = E_{\overline{C}_{E+1}}$, and so $E(\overline{\varphi}) = E_{\overline{C}_{E+1}}$.

Now, taking $\overline{\varphi}$, let $\eta \in C_c^\infty((0,h))$ and consider $\overline{\varphi}_s = \overline{\varphi} + s\eta$. Although we may not have $\overline{\varphi}_s \in \overline{C}_{E+1}$, observe that $\overline{\varphi}$ still minimizes the energy over the closure in $L^2([0,h])$ of the set of functions φ as in C_{E+1} except with $E(\varphi) \le E+2$. We can thus conclude $E(\overline{\varphi}) \le E(\overline{\varphi}_s)$ for all sufficiently small s. As in computing the Euler–Lagrange equation (2-2), we have that $\overline{\varphi}$ is a *weak solution* to the second-order ODE in $(0,h)$:

$$(T(t)\overline{\varphi}_t)_t + (P_c(t) - P_s(t))\frac{\sin 2\overline{\varphi}}{2} = 0,$$

meaning that for any $\eta \in C_c^\infty((0,h))$ we have

$$\int_0^h T(t)\overline{\varphi}_t\cdot\eta_t + (P_c(t) - P_s(t))\frac{\sin 2\overline{\varphi}}{2}\cdot\eta\,dt = 0.$$

Using standard regularity theory [Evans 1998, Section 6.3, Theorems 1 and 2], we conclude that $\overline{\varphi}\in C^\infty([0,h])$. □

4. Uniqueness

The following theorem will allow us to conclude uniqueness in certain circumstances. Let

$$T_{\min} = \min_{t\in[0,h]} T(t),\ \Theta_{\min} = \min_{t\in[0,h]} \Theta(t), \qquad P_{c-s} = \sup_{t\in[0,h]} |P_c(t) - P_s(t)|.$$

Theorem 4.1. *Let* $0 < h < \sqrt{\dfrac{8(T_{\min}+\Theta_{\min})}{P_{c-s}}}$, *and suppose that*

$$\varphi \in C^1(\mathbb{R}\times[0,h]) \cap C^2(\mathbb{R}\times(0,h))$$

is 2π*-periodic in* θ *and satisfies the Euler–Lagrange equation* (2-2) *in* $(0,2\pi)\times(0,h)$. *Then* φ *is uniquely determined by its boundary values* $\varphi(\theta,0)$, $\varphi(\theta,h)$.

The requirement that $\varphi(\theta,t)$ is 2π-periodic in θ geometrically means that for each fixed $t\in[0,h]$, as θ increases from 0 to 2π the vector field $\boldsymbol{V}=\boldsymbol{V}(\varphi(\theta,t))$ spins clockwise as many times as it does counterclockwise as measured from the horizontal vector field $(-\sin\theta,\cos\theta,0)$.

To prove the theorem we need first the following Poincaré inequality:

Lemma 4.2. *Suppose* $\varphi\in C^1(\mathbb{R}\times[0,h])$ *satisfies* $\varphi(\theta,0)=\varphi(\theta,h)=0$ *for each* $\theta\in\mathbb{R}$. *Then*

$$\int_0^h\int_0^{2\pi}\varphi^2\,d\theta\,dt \le \frac{h^2}{8}\int_0^h\int_0^{2\pi}(\varphi_t)^2+(\varphi_\theta)^2\,dt\,d\theta.$$

Proof. Writing

$$\varphi(\theta,t)=\int_0^t\frac{\partial}{\partial s}\varphi(\theta,s)\,ds=-\int_t^h\frac{\partial}{\partial s}\varphi(\theta,s)\,ds,$$
$$\int_0^h\varphi^2\,dt=\int_0^{h/2}\varphi^2\,dt+\int_{h/2}^h\varphi^2\,dt,$$

we have

$$\int_0^h\varphi^2\,dt=\int_0^{h/2}\Big(\int_0^t\frac{\partial\varphi}{\partial s}\,ds\Big)^2dt+\int_{h/2}^h\Big(\int_t^h\frac{\partial\varphi}{\partial s}\,ds\Big)^2dt.$$

Using Cauchy–Schwartz,

$$\begin{aligned}\int_0^h\varphi^2dt &\le \int_0^{h/2}t\Big(\int_0^t\Big(\frac{\partial\varphi}{\partial s}\Big)^2ds\Big)dt+\int_{h/2}^h(h-t)\Big(\int_t^h\Big(\frac{\partial\varphi}{\partial s}\Big)^2ds\Big)dt\\ &\le \int_0^{h/2}(\varphi_t)^2+(\varphi_\theta)^2\,dt\int_0^{h/2}t\,dt+\int_{h/2}^h(\varphi_t)^2+(\varphi_\theta)^2\,dt\int_{h/2}^h(h-t)\,dt,\end{aligned}$$

which gives $\int_0^h\varphi^2\,dt\le\frac{1}{8}h^2\int_0^h\varphi_t^2+\varphi_\theta^2\,dt$. Integrating with respect to θ gives the result. □

Proof of Theorem 4.1. Suppose $\varphi_1,\varphi_2\in C^1(\mathbb{R}\times[0,h])\cap C^2(\mathbb{R}\times(0,h))$ are solutions to (2-2), both 2π-periodic in θ and satisfying

$$\varphi_1(\theta,0)=\varphi_2(\theta,0),\quad \varphi_1(\theta,h)=\varphi_2(\theta,h).$$

Multiplying

$$(T(t)(\varphi_1-\varphi_2)_t)_t+(\Theta(t)(\varphi_1-\varphi_2)_\theta)_\theta+(P_c(t)-P_s(t))\Big(\frac{\sin 2\varphi_1}{2}-\frac{\sin 2\varphi_2}{2}\Big)=0$$

by $\varphi_1-\varphi_2$ and integrating gives

$$\int_0^h\int_0^{2\pi}\big[(T(t)(\varphi_1-\varphi_2)_t)_t+(\Theta(t)(\varphi_1-\varphi_2)_\theta)_\theta\big](\varphi_1-\varphi_2)$$
$$+(P_c(t)-P_s(t))\Big(\frac{\sin 2\varphi_1}{2}-\frac{\sin 2\varphi_2}{2}\Big)(\varphi_1-\varphi_2)\,d\theta\,dt=0.$$

Since $(\varphi_1-\varphi_2)(\theta,0)=(\varphi_1-\varphi_2)(\theta,h)=0$ and φ_1,φ_2 are 2π-periodic in θ, then integration by parts gives:

$$\int_0^h\int_0^{2\pi}T(t)((\varphi_1-\varphi_2)_t)^2+\Theta(t)((\varphi_1-\varphi_2)_\theta)^2\,d\theta\,dt$$
$$=\int_0^h\int_0^{2\pi}(P_c(t)-P_s(t))\Big(\frac{\sin 2\varphi_1}{2}-\frac{\sin 2\varphi_2}{2}\Big)(\varphi_1-\varphi_2)\,d\theta\,dt.$$

We now use the inequality $|\sin x-\sin y|\le|x-y|$ to get

$$(T_{\min}+\Theta_{\min})\int_0^h\int_0^{2\pi}((\varphi_1-\varphi_2)_t)^2+((\varphi_1-\varphi_2)_\theta)^2\,d\theta\,dt$$
$$\le P_{c-s}\int_0^h\int_0^{2\pi}(\varphi_1-\varphi_2)^2\,d\theta\,dt.$$

Lemma 4.2 now implies

$$\int_0^h\int_0^{2\pi}((\varphi_1-\varphi_2)_t)^2+((\varphi_1-\varphi_2)_\theta)^2\,d\theta\,dt$$
$$\le\frac{P_{c-s}}{(T_{\min}+\Theta_{\min})}\frac{h^2}{8}\int_0^h\int_0^{2\pi}((\varphi_1-\varphi_2)_t)^2+((\varphi_1-\varphi_2)_\theta)^2\,d\theta\,dt.$$

When

$$h<\sqrt{\frac{8(T_{\min}+\Theta_{\min})}{P_{c-s}}}$$

we see that $\varphi_1=\varphi_2$ must occur. □

Theorem 4.1 together with Lemma 3.1 imply the following corollary:

Corollary 4.3. *Let*

$$h<\sqrt{\frac{8(T_{\min}+\Theta_{\min})}{P_{c-s}}}$$

and take boundary data V_0, V_h each with constant angle function. Suppose $\boldsymbol{V}=\boldsymbol{V}(\varphi)$ and $\tilde{\boldsymbol{V}}=\tilde{\boldsymbol{V}}(\tilde\varphi)$ are minimizers with $\varphi,\tilde\varphi\in C^\infty([0,h])$. If $\varphi(0)=\tilde\varphi(0)$ and $\varphi(h)=\tilde\varphi(h)$, then $\varphi=\tilde\varphi$ and so $\boldsymbol{V}=\tilde{\boldsymbol{V}}$.

For the boundary data $V_0 = V_0(\pi/2)$ and $V_h = V_h(-\pi/2)$, if $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ is a minimizer then so is $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}(\pi - \varphi) \neq \boldsymbol{V}$. In the next section we show that uniqueness holds without reference to the angle functions in certain cases.

5. Twisting in the unit cylinder

Recall that in a cylinder or in a frustum of a cone the vector field $\boldsymbol{V} = \boldsymbol{V}(k\pi/2)$ with k odd minimizes energy over all vector fields in $\mathfrak{X}^1(\mathcal{S})$. This allows us to show that minimizers ought not to "twist" too much, if the boundary data does not. We show the case of the cylinder.

Theorem 5.1. *Let $V_0 = V_0(\varphi_0)$ and $V_h = V_h(\varphi_h)$ be continuous boundary data on $\partial\mathscr{C}$. Suppose for some $n \in \mathbb{Z}$ and all $\theta \in \mathbb{R}$ we have*

$$\frac{n\pi}{2} - \frac{\pi}{2} < \varphi_0(\theta), \quad \varphi_h(\theta) < \frac{n\pi}{2} + \frac{\pi}{2}.$$

Then for any $\epsilon > 0$ and $\boldsymbol{V} \in \mathfrak{X}^1(\mathscr{C})$ with $\boldsymbol{V}|_{\partial\mathscr{C}} = V_0, V_h$ and $E(\boldsymbol{V}) < \infty$, there is $\tilde{\boldsymbol{V}} \in \mathfrak{X}^1(\mathscr{C})$ with $\tilde{\boldsymbol{V}}|_{\partial\mathscr{C}} = V_0, V_h$, $E(\tilde{\boldsymbol{V}}) < E(\boldsymbol{V}) + \epsilon$, and so that we can write $\tilde{\boldsymbol{V}} = \tilde{\boldsymbol{V}}(\tilde{\varphi})$ using an angle function $\tilde{\varphi}$ with $n\pi/2 - \pi/2 < \tilde{\varphi} < n\pi/2 + \pi/2$.

We remark that the calculation $\cos^2(\varphi \pm n\pi) = \cos^2(\varphi)$ is used in the proof; the argument as given cannot be used in case φ_0, φ_h have values in a period of length π centered at an angle not of the form $n\pi/2$ with $n \in \mathbb{Z}$.

Proof. Take a vector field $\boldsymbol{V} \in \mathfrak{X}^1(\mathscr{C})$ with boundary data V_0, V_h, and write $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ using an angle function satisfying $n\pi/2 - \pi \leq \varphi < n\pi/2 + \pi$. We choose φ to be smooth at all points where $\varphi \neq n\pi/2 - \pi$, so that in particular φ is smooth near $t = 0, h$.

Suppose $\{(\theta, t) : |\varphi(\theta, t) - n\pi/2| \geq \pi/2\} = \{(\theta, t) : \cos(\varphi(\theta, t) - n\pi/2) \leq 0\}$ is a nonempty set. Applying Sard's theorem to the smooth function $\cos(\varphi - n\pi/2)$ in $[0, 2\pi] \times [0, h]$, we can choose $\theta_1 < \pi/2$ so that $|\varphi_0(\theta) - n\pi/2|, |\varphi_h(\theta) - n\pi/2| < \theta_1$ for all $\theta \in [0, 2\pi]$ and so that $\{(\theta, t) \in [0, 2\pi] \times [0, h] : |\varphi(\theta, t) - n\pi/2| = \theta_1\}$ is a finite collection of closed Jordan curves together with Jordan arcs with endpoints at $\{0, 2\pi\} \times (0, h)$. See Figure 1 for example.

Let

$$A_{<\theta_1} = \{(\theta, t) \in (0, 2\pi) \times (0, h) : |\varphi - n\pi/2| < \theta_1\}.$$

Necessarily, φ is smooth in $A_{<\theta_1}$. Also let $A_{>\theta_1} = \{(\theta, t) \in (0, 2\pi) \times (0, h) : |\varphi - n\pi/2| > \theta_1\}$. We then have $\overline{A}_{>\theta_1} \subset [0, 2\pi] \times (0, h)$ (see the shaded region in Figure 1 for example).

Define the function

$$R_{\theta_1} : \left[\frac{n\pi}{2} - \pi, \frac{n\pi}{2} + \pi\right) \to \left[\frac{n\pi}{2} - \theta_1, \frac{n\pi}{2} + \theta_1\right]$$

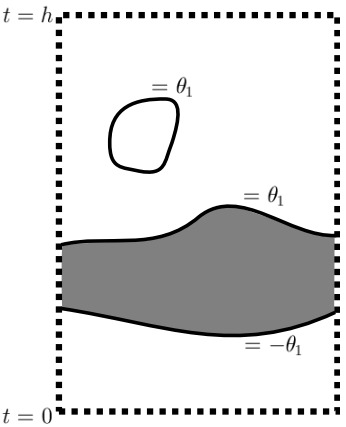


Figure 1. Sard's theorem for $\varphi(\theta, t) - n\pi/2 = \pm\theta_1$. In this case φ is discontinuous in the shaded region, which is $A_{>\theta_1}$.

by

$$R_{\theta_1}(x) = \begin{cases} -\dfrac{\theta_1}{\pi - \theta_1}\left(x - \left(\dfrac{n\pi}{2} - \pi\right)\right) + \dfrac{n\pi}{2} & \text{for } x \in \left[\dfrac{n\pi}{2} - \pi, \dfrac{n\pi}{2} - \theta_1\right], \\ x & \text{for } x \in \left(\dfrac{n\pi}{2} - \theta_1, \dfrac{n\pi}{2} + \theta_1\right], \\ -\dfrac{\theta_1}{\pi - \theta_1}\left(x - \left(\dfrac{n\pi}{2} + \pi\right)\right) + \dfrac{n\pi}{2} & \text{for } x \in \left(\dfrac{n\pi}{2} + \theta_1, \dfrac{n\pi}{2} + \pi\right). \end{cases}$$

Considering $R_{\theta_1}(\varphi(\theta, t))$, we see that $R_{\theta_1} \circ \varphi = \varphi$ for $(\theta, t) \in A_{<\theta_1}$. Furthermore, we can immediately see that $R_{\theta_1} \circ \varphi$ is Lipschitz near every point with $\varphi(\theta, t) \neq n\pi/2 - \pi$. However, note that the function defined by

$$\begin{cases} \varphi(\theta, t) + \pi & \text{if } \varphi(\theta, t) < n\pi/2, \\ \varphi(\theta, t) - \pi & \text{if } \varphi(\theta, t) > n\pi/2, \end{cases}$$

is smooth at points where $\varphi(\theta, t) = n\pi/2 - \pi$. Hence, $R_{\theta_1} \circ \varphi$ is Lipschitz in $(0, 2\pi) \times (0, h)$.

Next, since $R_{\theta_1} \circ \varphi$ is Lipschitz, it is differentiable almost everywhere [Evans 1998, Theorem 6, Section 5.8], and we may still define the energy of $R_{\theta_1} \circ \varphi$ by

$$E(R_{\theta_1} \circ \varphi) = \int_0^h \int_0^{2\pi} ((R_{\theta_1} \circ \varphi)_t)^2 + ((R_{\theta_1} \circ \varphi)_\theta)^2 + \cos^2 R_{\theta_1} \circ \varphi \, d\theta \, dt.$$

Observe then that

$$E(R_{\theta_1} \circ \varphi) = \int_{A_{<\theta_1}} (\varphi_t)^2 + (\varphi_\theta)^2 + \cos^2 \varphi \, d\theta \, dt + \left(\frac{\theta_1}{\pi - \theta_1}\right)^2 \int_{A_{>\theta_1}} (\varphi_t)^2 + (\varphi_\theta)^2 \, d\theta \, dt + \int_{A_{>\theta_1}} \cos^2 R_{\theta_1} \circ \varphi \, d\theta \, dt.$$

Choose a sequence $\theta_k \nearrow \pi/2$ so that we have regions $A_{<\theta_k}, A_{>\theta_k}$ as above (Sard's theorem says this can be done for almost every θ near $\pi/2$). Note that the sets $A_{<\theta_k} \to \{|\varphi - (n\pi/2)| < \pi/2\}$, $A_{>\theta_k} \to \{|\varphi - (n\pi/2)| \geq \pi/2\}$, and the functions $R_{\theta_k} \circ \varphi \to R_{\pi/2} \circ \varphi$ pointwise. Since $E(\boldsymbol{V}) < \infty$ we have, by the dominated convergence theorem

$$\lim_{k\to\infty} E(R_{\theta_k} \circ \varphi) = \int_{\{|\varphi-(n\pi/2)|<\pi/2\}} (\varphi_t)^2 + (\varphi_\theta)^2 + \cos^2\varphi \, d\theta \, dt$$
$$+ \int_{\{|\varphi-(n\pi/2)|\geq\pi/2\}} (\varphi_t)^2 + (\varphi_\theta)^2 \, d\theta \, dt + \int_{\{|\varphi-(n\pi/2)|\geq\pi/2\}} \cos^2 R_{\pi/2} \circ \varphi \, d\theta \, dt.$$

Since

$$\cos^2\left(-\left(\varphi - \left(\frac{n\pi}{2} - \pi\right)\right) + \frac{n\pi}{2}\right) = \cos^2\left(-\left(\varphi - \left(\frac{n\pi}{2} + \pi\right)\right) + \frac{n\pi}{2}\right) = \cos^2\varphi,$$

we have $\cos^2 R_{\pi/2} \circ \varphi = \cos^2\varphi$. Thus $\lim_{k\to\infty} E(R_{\theta_k} \circ \varphi) = E(\boldsymbol{V})$.

Note that $R_{\theta_k} \circ \varphi$ is Lipschitz with derivatives in $L^2_{\text{loc}}(\mathbb{R} \times [0, h])$. In other words,

$$R_{\theta_k} \circ \varphi \in W^{1,2}_{\text{loc}}(\mathbb{R} \times [0, h])$$

(see [Evans 1998, Section 5.2.2]) and we can find a θ-periodic function

$$\varphi^\infty \in C(\mathbb{R} \times [0, h]) \cap C^\infty(\mathbb{R} \times (0, h)),$$

so that $\|(R_{\theta_k} \circ \varphi) - \varphi^\infty\|_{W^{1,2}([0,2\pi]\times[0,h])}$ is as small as we please. For this, see [Evans 1998, Section 5.3]; in particular we can apply Theorem 3 of Section 5.3.3 to a bounded smooth region $U \subset \mathbb{R} \times (0, h)$ containing $(0, 2\pi) \times (0, h)$, and we can ensure our approximating functions are θ-periodic.

However, φ^∞ may not have the correct boundary data, and so we fix φ^∞ as follows. Choose $\sigma > 0$ so that $R_{\theta_k} \circ \varphi = \varphi$ for $t \in [0, 2\sigma) \cup (h - 2\sigma, h]$. Pick a smooth function g with $0 \leq g \leq 1$ so that $g(t) = 1$ for $t \in [0, \sigma) \cup (h - \sigma, h]$, $g(t) = 0$ for $t \in (2\sigma, h - 2\sigma)$, and $|g'(t)| \leq 2/\sigma$. Define

$$\tilde{\varphi}(\theta, t) = g(t) R_{\theta_k}(\varphi(\theta, t)) + (1 - g(t))\varphi^\infty(\theta, t).$$

We now compute $E(\tilde{\varphi})$. First,

$$\tilde{\varphi}_t = (\varphi^\infty)_t + g_t((R_{\theta_k} \circ \varphi) - \varphi^\infty) + g((R_{\theta_k} \circ \varphi) - \varphi^\infty)_t.$$

By the inequality $(x + y)^2 \leq (1 + \epsilon)x^2 + ((\epsilon + 1)/\epsilon)y^2$ we have

$$(\tilde{\varphi}_t)^2 \leq (1 + \epsilon)((\varphi^\infty)_t)^2 + \left(\frac{\epsilon + 1}{\epsilon}\right)\left(g_t((R_{\theta_k} \circ \varphi) - \varphi^\infty) + g((R_{\theta_k} \circ \varphi) - \varphi^\infty)_t\right)^2.$$

Then $(x + y)^2 \leq 2x^2 + 2y^2$ along with $|g_t| \leq 2/\sigma$ imply

$$(\tilde{\varphi}_t)^2 \leq (1+\epsilon)((\varphi^\infty)_t)^2 + \left(\frac{\epsilon + 1}{\epsilon}\right)\left(\frac{8}{\sigma^2}((R_{\theta_k} \circ \varphi) - \varphi^\infty)^2 + 2(((R_{\theta_k} \circ \varphi) - \varphi^\infty)_t)^2\right).$$

Second, we similarly have

$$(\tilde{\varphi}_\theta)^2 \le (1+\epsilon)((\varphi^\infty)_\theta)^2 + \left(\frac{\epsilon+1}{\epsilon}\right)(((R_{\theta_k}\circ\varphi)-\varphi^\infty)_\theta)^2.$$

Third, since $|\cos^2 x - \cos^2 y| \le 2|x-y|$, we have

$$\cos^2\tilde{\varphi} \le \cos^2\varphi^\infty + 2|(R_{\theta_k}\circ\varphi)-\varphi^\infty|.$$

We therefore have by the definition of $\|(R_{\theta_k}\circ\varphi)-\varphi^\infty\|_{W^{1,2}([0,2\pi]\times[0,h])}$

$$E(\tilde{\varphi}) \le (1+\epsilon)E(\varphi^\infty) + \left[\left(\frac{\epsilon+1}{\epsilon}\right)\left(\frac{8}{\sigma^2}+3\right)+2\right]\|(R_{\theta_k}\circ\varphi)-\varphi^\infty\|^2_{W^{1,2}([0,2\pi]\times[0,h])}.$$

Given $\epsilon > 0$, we can choose $R_{\theta_k}\circ\varphi$ so that $E(R_{\theta_k}\circ\varphi) < E(\varphi)+\epsilon$. We can then choose φ^∞ with $\|(R_{\theta_k}\circ\varphi)-\varphi^\infty\|^2_{W^{1,2}([0,2\pi]\times[0,h])}$ sufficiently small so that $E(\varphi^\infty) < E(\varphi)+\epsilon$ as well. Since σ depends only on φ, we can make $E(\tilde{\varphi})$ as close to $E(\varphi)$ as we like. □

We remark that even given Theorem 5.1, we cannot immediately argue as in Theorem 3.2 to show the existence of a minimizer in case the boundary data does not twist too much. To see this, take a sequence φ_k as in Theorem 5.1 with $E(\varphi_k)$ converging to the infimum energy. Unlike in the proof of Theorem 3.2, it is unclear whether the sequence φ_k is equicontinuous. Although we can conclude the φ_k converge to some function φ *weakly* in L^2, it is not clear whether the sequence $\cos\varphi_k$ converges weakly to $\cos\varphi$. Thus, we cannot conclude that $E(\varphi)$ is the infimum energy.

Corollary 5.2. *The minimizer $\boldsymbol{V}$ with horizontal boundary data on $\mathscr{C}$ is unique for $h < \sqrt{8}$.*

Proof. Let $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ be a minimizer, so φ must be θ-independent, and we can write $\varphi(0) = 0$. Now, if $\varphi(h) = 0$, then by Theorem 4.1 we get $\varphi = 0$. If instead, suppose $\varphi(h) = 2\pi$, then let

$$t_0 = \inf\left\{t \in [0,h] : \varphi(t) = \frac{\pi}{2}\right\} \quad \text{and} \quad t_1 = \sup\left\{t \in [0,h] : \varphi(t) = \frac{5\pi}{2}\right\}.$$

Define $\tilde{\varphi}(t) = -\varphi(t)$ for $0 \le t < t_0$, $\tilde{\varphi}(t) = -(\pi/2)$ for $t_0 \le t < t_1$, and $\tilde{\varphi}(t) = \varphi(t) - 2\pi$ for $t_1 \le t \le h$. In this case note $E(\tilde{\varphi}) < E(\varphi)$, and we can smooth $\tilde{\varphi}$ and still conclude the same. This is a contradiction, and so we must have $\varphi = 0$. □

6. Computer approximations

In this section we present two numerical approximations of solutions to (2-2) for two surfaces of revolution. To sidestep the possibility of suffering Runge's phenomenon [Runge 1901], our numerical approximations sample Chebyshev points; these are points which cluster near the boundary of $[0, 2\pi] \times [0, h]$. To handle

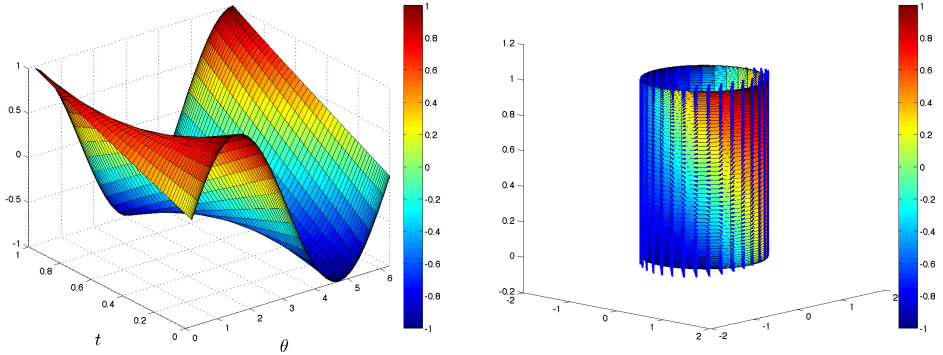


Figure 2. Left: plot of $\varphi(\theta, t)$ for $\varphi(\theta, 0) = \sin\theta$, $\varphi(\theta, 1) = \cos\theta$. Right: plot of $\boldsymbol{V}(\varphi)$ for the same φ.

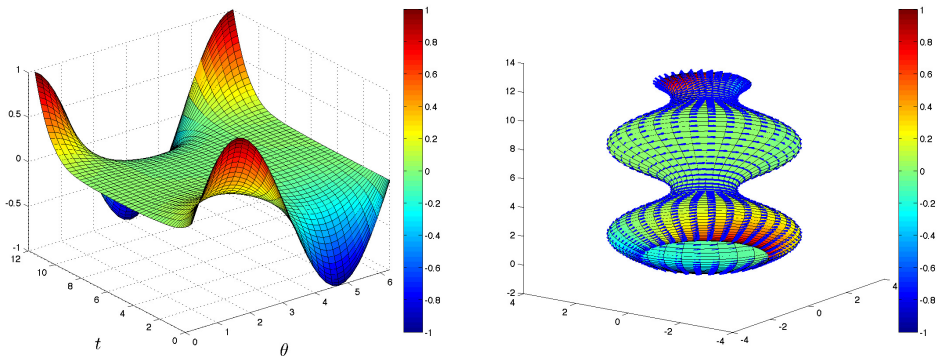


Figure 3. Left: plot of $\varphi(\theta, t)$ for $r(t) = \sin(t) + 2$. Right: plot of $\boldsymbol{V}(\varphi)$ for the same φ.

periodicity in the θ variable, we borrow some theory about Fourier discretization matrices from [Trefethen 2000]. These matrices allow us to solve our differential equation on the interior of the cylinder $(\mathbb{R} \bmod 2\pi) \times [0, h]$ while leaving the boundary conditions fixed.

Our program allows us to input a height h, a radius function $r(t)$, and two functions $\varphi_0(\theta)$ and $\varphi_h(\theta)$ that describe the boundary conditions, and finds a very close approximation of a function $\varphi(\theta, t)$ which satisfies (2-2) with boundary data φ_0, φ_h over $[0, 2\pi] \times [0, h]$.

First, we take the unit cylinder with unit height, and we take boundary data $\varphi_0(\theta) = \sin(\theta)$ and $\varphi_1(\theta) = \cos(\theta)$. We plot the solution $\varphi(\theta, t)$ and also $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ in Figure 2. Second, we take the surface with $r(t) = \sin(t) + 2$, and set $\varphi_0(\theta) = \sin(\theta)$, $\varphi_{12}(\theta) = \cos(\theta)$. We again plot $\varphi(\theta, t)$ and $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ in Figure 3.

7. Future projects

There are a number of projects well-suited for future VIGRE at Rice internships for undergraduates. We mention a few.

The first problem is to extend Theorem 3.2 to the case when the boundary data V_0, V_h cannot be written using constant angle functions. A preliminary challenge is to show the existence of $\boldsymbol{V} \in \mathfrak{X}^1(\mathcal{S})$ with $\boldsymbol{V}|_{\partial\mathcal{S}} = V_0, V_h$ in the case of general continuous boundary data. (When V_0, V_h are smooth, this can be done using an argument similar to the end of the proof of Theorem 5.1.) Theorem 5.1 provides a first step in showing the existence of minimizers, at least if we assume V_0, V_h do not twist too much.

Related to Theorem 5.1 is finding the largest h_0 so that Corollary 5.2 continues to hold with $h < h_0$. This is related to the analogous question for Theorem 4.1 and Lemma 4.2, and similar to the well-studied question of finding the optimal constant in the usual Poincaré inequality [Bebendorf 2003].

Another direction is to consider the following inverse problem: given an angle function $\varphi(\theta, t)$, find the surface of revolution $\mathcal{S}$ such that $\boldsymbol{V} = \boldsymbol{V}(\varphi)$ minimizes energy with respect to the boundary data $\boldsymbol{V}(\varphi(\theta, 0))$, $\boldsymbol{V}(\varphi(\theta, h))$. A different problem with a similar flavor is to find, given an angle function $\varphi(\theta, t)$, which surface of revolution $\mathcal{S}$ is such that $E(\varphi)$ is the least.

The torus of revolution also provides a fountain of projects, by asking which smooth tangent unit-length vector field minimizes energy. Some work has been done by the authors in this direction [Rosales et al. 2010], most notably in computing the relationship between the radii of the tube and the distance to the center of the tube of the torus with the energies of the normalizations of the coordinate vector fields, when the torus is given the usual parametrization.

8. Acknowledgments

This work was conducted in the summers of 2009 and 2010, over the course of two undergraduate internship programs at Rice University under the supervision of Dr. Robert Hardt, Dr. Leobardo Rosales, and Dr. Michael Wolf. In 2009 the participating undergraduate students were Yak Digilov, Bill Eggert, Michael Jauch, Rob Lewis, and Hector Perez; in 2010, they were James Hart, Conor Loftis, Aneesh Mehta, and Anand Shah. The internship, *VIGRE at Rice*, is an initiative sponsored by the National Science Foundation to carry out innovative educational programs in which research and education are integrated and in which undergraduates, graduate students, postdocs, and faculty are mutually supportive. Reports from the internships can be found in [Rosales et al. 2009; 2010].

We thank graduate students Christoper Davis, Evelyn Lamb, Renee Laverdiere, and Ryan Scott for helping supervise the students, Dr. Rolf Ryham for volunteering

advice, and Dr. Mark Embree of the Department of Computational and Applied Mathematics at Rice for being instrumental in achieving the computational work.

References

[Bebendorf 2003] M. Bebendorf, "A note on the Poincaré inequality for convex domains", *Z. Anal. Anwendungen* **22**:4 (2003), 751–756. MR 2004k:26025 Zbl 1057.26011

[Borrelli et al. 2003] V. Borrelli, F. Brito, and O. Gil-Medrano, "The infimum of the energy of unit vector fields on odd-dimensional spheres", *Ann. Global Anal. Geom.* **23**:2 (2003), 129–140. MR 2003m:53046 Zbl 1031.53090

[Brito and Walczak 2000] F. G. B. Brito and P. Walczak, "On the energy of unit vector fields with isolated singularities", *Ann. Polon. Math.* **73**:3 (2000), 269–274. MR 2001k:53049 Zbl 0997.53025

[Chen et al. 2004] G. Chen, Z. Ding, C.-R. Hu, W.-M. Ni, and J. Zhou, "A note on the elliptic sine-Gordon equation", pp. 49–67 in *Variational methods: open problems, recent progress, and numerical algorithms* (Flagstaff, AZ, 2002), edited by J. M. Neuberger, Contemp. Math. **357**, Amer. Math. Soc., Providence, RI, 2004. MR 2005f:35059 Zbl 02144556

[Evans 1998] L. C. Evans, *Partial differential equations*, Graduate Studies in Mathematics **19**, American Mathematical Society, Providence, RI, 1998. MR 99e:35001 Zbl 0902.35002

[Hu 1982] H. S. Hu, "Sine-Laplace equation, sinh-Laplace equation and harmonic maps", *Manuscripta Math.* **40**:2-3 (1982), 205–216. MR 84i:58037 Zbl 0511.35061

[Rosales et al. 2009] L. Rosales et al., "Minimizing the energy of vector fields on surfaces of revolution", online report, 2009, available at http://cnx.org/content/m30944/latest/.

[Rosales et al. 2010] L. Rosales et al., "Minimizing the energy of vector fields on surfaces of revolution, II", online report, 2010, available at http://cnx.org/content/m34976/latest/.

[Runge 1901] C. Runge, "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten", *Z. Math. Phys.* **46** (1901), 224–243. JFM 32.0272.02

[Trefethen 2000] L. N. Trefethen, *Spectral methods in MATLAB*, Software, Environments, and Tools **10**, Soc. Ind. Appl. Math., Philadelphia, 2000. MR 2001c:65001 Zbl 0953.68643

[Wiegmink 1995] G. Wiegmink, "Total bending of vector fields on Riemannian manifolds", *Math. Ann.* **303**:2 (1995), 325–344. MR 97a:53050 Zbl 0834.53034

Received: 2010-09-27 Revised: 2010-10-12 Accepted: 2010-10-14

Yan.M.Digilov@rice.edu william.j.eggert@rice.edu james.e.hart@rice.edu
Michael.A.Jauch@rice.edu Rob.Lewis@rice.edu Conor.T.Loftis@rice.edu
anm5@rice.edu Hector.Perez@rice.edu anand.shah@rice.edu
Rice University, 6100 S. Main Street, MS 136, Houston, TX 77005-1892, United States

hardt@rice.edu *Department of Mathematics, Rice University, 6100 S. Main Street, MS 136, Houston, TX 77005-1892, United States*

lrosales@rice.edu *Department of Mathematics, Rice University, 6100 S. Main Street, MS 136, Houston 77005-1892, United States*

mwolf@rice.edu *Department of Mathematics, Rice University, 6100 S. Main Street, MS 136, Houston 77005-1892, United States*
http://math.rice.edu/~mwolf

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.451

Some conjectures on the maximal height of divisors of $x^n - 1$

Nathan C. Ryan, Bryan C. Ward and Ryan Ward

(Communicated by Kenneth S. Berenhaut)

Define $B(n)$ to be the largest height of a polynomial in $\mathbb{Z}[x]$ dividing $x^n - 1$. We formulate a number of conjectures related to the value of $B(n)$ when n is of a prescribed form. Additionally, we prove a lower bound for $B(n)$.

1. Introduction

The height $H(f)$ of a polynomial f is the largest coefficient of f in absolute value. Let

$$\Phi_n(x) = \prod_{\substack{1 \le a \le n \\ (a,n)=1}} (x - e^{2\pi i a/n})$$

be the n-th cyclotomic polynomial. For example, for a prime p, we have

$$\Phi_p(x) = 1 + x + \cdots + x^{p-1}.$$

Define the function $A(n) := H(\Phi_n(x))$. This function was originally studied by Erdős and has been much investigated since then. The second of the following two facts reduces the study of $A(n)$ to square-free n:

$$\Phi_{np}(x) = \frac{\Phi_n(x^p)}{\Phi_n(x)} \text{ if } p \nmid n \quad \text{and} \quad \Phi_{np}(x) = \Phi_n(x^p) \text{ if } p \mid n. \tag{1-1}$$

The variant we study in the present paper was first defined in [Pomerance and Ryan 2007] and studied further in [Kaplan 2009]. In [Pomerance and Ryan 2007] the function

$$B(n) = \max\{H(f) : f \mid x^n - 1 \text{ and } f \in \mathbb{Z}[x]\}$$

is defined and a fairly good asymptotic bound is found. In the same paper there are two explicit formulas for n of a certain form: it is shown that $B(p^k) = 1$ and $B(pq) = \min\{p, q\}$. In the present paper, for n of a prescribed form, we are interested in finding explicit formulas for $B(n)$, discovering bounds for $B(n)$,

MSC2000: 11C08, 11Y70, 12Y05.
Keywords: cyclotomic polynomials, heights.

determining which divisors of $x^n - 1$ have height $B(n)$ and understanding the image of $B(n)$. One might consider the present paper a continuation of [Kaplan 2009], where it was shown that $B(p^2q) = \min\{p^2, q\}$ and where upper bounds were found for $B(n)$. Kaplan also found a better upper bound as well as a lower bound for $B(pqr)$, where $p < q < r$ are primes.

Our main theoretical result is a lower bound for $B(p^aq^b)$, but most of the content of the paper consists of conjectures about $B(n)$ of the kind described above. The conjectures are verified by extensive data computed in Sage (www.sagemath.org) and tabulated in [Ryan et al. 2010].

The paper is organized as follows. In Section 2 we describe our computations: the method and the scale. Section 3 provides a reasonably good lower bound for $B(n)$ in terms of its prime factorization. The first of the subsequent two sections, Section 4, is about $B(n)$ for n that are divisible by two distinct primes. Section 5 investigates what happens when 3 or more primes divide n. We conclude the paper with three further variants on the arithmetic function $B(n)$. For the first of these three variants, related data have also been tabulated in [Ryan et al. 2010].

2. Computations

Much of what is included in the present paper is the result of a great deal of machine computation. The function $B(n)$ is very difficult to compute. The best way we know to compute $B(n)$ is to do the following: observe that any f that would give a maximal height is a product of cyclotomic polynomials since

$$x^n - 1 = \prod_{d|n} \Phi_d(x). \tag{2-1}$$

So, to compute $B(n)$ we need to compute the set of divisors of n and its power set. We then iterate over the power set, multiplying the corresponding cyclotomic polynomials in each set. The largest height among the polynomials in this very long list is the value of $B(n)$.

We have computed $B(n)$ for almost 300,000 values of n, the largest being 56,796,482. This includes all n with four or fewer prime factors, and in particular every n less than 1000.

These computations were done in Sage, and took 30 processors several months on various systems at Bucknell University: many were run on a cluster node with dual quad core 3.33 GHz Xeons with 64GB of RAM. For example, $B(720)$ took 113 hours to compute and $B(840)$ took 550 hours.

The resulting data can be accessed freely at [Ryan et al. 2010]. We store all data we consider useful for formulating conjectures about $B(n)$. This includes n, $B(n)$, and the set of sets of cyclotomic polynomials which multiply to yield the maximal height.

form of n	conjecture	ranges	# data points
p^2q^2	4.1	$2 \le p < q < 60$	463
$2q^b$	4.2	$2 < q < 300,\ b = 2$	96
		$2 < q < 100,\ b = 3$	24
		$2 < q < 75,\ b = 4$	20
		$2 < q < 10,\ b = 5$	4
		$2 < q < 10,\ b = 6$	4
		$q \in \{3, 5\},\ b = 7$	2
		$q \in \{3, 5\},\ b = 8$	2
pq^b	4.2, 4.4	$2 < p < q < 85,\ b = 3$	301
		$2 < p < q < 35,\ b = 4$	92
		$2 < p < q < 15,\ b = 5$	14
		$2 < p < q < 10,\ b = 6$	13
pqr	5.1	$2 \le p < q < r < 150$	55530
$pqrs$	5.1	$2 \le p < q < r < s < 15$	1045
pqr^b	5.2	$2 \le q < r < 50,\ b = 2$	1490
		$2 \le q < r < 35,\ b = 3$	171
		$2 \le q < r < 35,\ b = 4$	13

Table 1. Summary of data motivating the conjectures in this paper. The data can be accessed at [Ryan et al. 2010].

We note that far less comprehensive computations have been done in [Abbott 2009], and a smaller set of data can be found at [Garcia 2006].

We present in the next section the conjectures we have formulated based on these computational data; the values of n so studied are summarized in Table 1.

3. Lower bound

We start by stating a lower bound for the function $B(n)$. (We thank Pieter Moree and the anonymous referee for independently pointing out this improvement to our earlier result.)

Theorem 3.1. *Suppose $n = uv$, with u and v coprime positive integers. Then $B(n) \ge \min\{u, v\}$.*

Proof. Since u and v are coprime, we note that $x^u - 1$ and $x^v - 1$ have $x - 1$ as greatest common divisor. Consider the divisor

$$(x^u - 1)(x^v - 1)/(x - 1)^2 \quad \text{of } x^{uv} - 1.$$

Let $w = \min\{u, v\}$ and observe that the coefficient of x^{w-1} is w. □

This result can be rephrased as follows:

Corollary 3.2. *We have* $B(p_1^{e_1} \cdots p_s^{e_s}) \geq \min\{p_1^{e_1}, \ldots, p_s^{e_s}\}$.

We observe that this bound is surprisingly good for the data we have computed, at least when n is divisible by two primes. Of the 5396 n in the database of the form $p^a q^b$, $B(n) = \min\{p^a, q^b\}$ a majority of the time (we exclude $(a, b) \in \{(1, 1), (1, 2), (2, 1)\}$ in this total as in those cases it is a theorem that $B(n) = \min\{p^a, q^b\}$).

4. When n is divisible by two primes

Evaluation of the function $A(p^a q^b)$ is straightforward. To see that $A(p^a q^b) = 1$, one can write down an explicit formula for $\Phi_{pq}(x)$ (see, e.g., [Lam and Leung 1996]) and then use (1-1). The situation for $B(p^a q^b)$ is not all like the situation for $A(p^a q^b)$.

By means of a thorough case-by-case analysis, one can find an explicit formula for $B(pq^2)$ [Kaplan 2009, Theorem 6] where p and q are distinct primes. The proof proceeds by computing the height of every possible divisor of $x^{pq^2} - 1$ and identifying which of those is largest. In that spirit we make the note of the following:

Conjecture 4.1. Let $p < q$ be primes. Then $B(p^2 q^2)$ is the larger of

$$H(\Phi_p(x)\Phi_q(x)\Phi_{p^2q}(x)\Phi_{pq^2}(x)) \quad \text{and} \quad H(\Phi_p(x)\Phi_q(x)\Phi_{p^2}(x)\Phi_{q^2}(x)).$$

For example,

$$B(3^2 \cdot 5^2) = H(\Phi_3\Phi_5\Phi_{3^2\cdot 5}\Phi_{3\cdot 5^2}) \neq H(\Phi_3\Phi_5\Phi_{3^2}\Phi_{5^2}),$$
$$B(5^2 \cdot 11^2) = H(\Phi_5\Phi_{11}\Phi_{5^2}\Phi_{11^2}) \neq H(\Phi_5\Phi_{11}\Phi_{5^2\cdot 11}\Phi_{5\cdot 11^2}).$$

In addition to not having a proof for this conjecture, we also lack an explicit formula for the height of the polynomial. The conjecture has been checked for the primes indicated in Table 1.

An even more difficult problem is to deduce a formula for n of a more arbitrary form. For example, our computations suggest the following conjecture.

Conjecture 4.2. Let $p < q$ be odd primes.

(i) For any positive integer b, $B(2q^b) = 2$.

(ii) Suppose $b > 2$. Then $B(pq^b) > p$.

The difficulty here is that a case by case analysis as described above is not feasible.

We have computed data verifying the first part of the conjecture as indicated in Table 1. The cases $b = 1$ and $b = 2$ in the first part are theorems in [Pomerance

and Ryan 2007] and [Kaplan 2009], respectively. We have verified the second half of the conjecture as indicated in Table 1.

The previous conjectures deals with what values of $B(pq^b)$ you get when you have two fixed primes and let one of the exponents vary. A related question is what happens when you have one fixed prime and two fixed exponents.

Theorem 4.3. *Fix a prime p and positive integers a and b. Then $B(p^a q^b)$ takes on only finitely many values as q ranges through the set of primes.*

Proof. This is a rephrasing of a special case of [Kaplan 2009, Theorem 4]. □

As a result of investigating this theorem computationally, we make the following observation:

Conjecture 4.4. For a fixed odd prime p and fixed positive integer b, the finite list of values $B(pq^b)$ as $q > p$ varies are all divisible by p.

We have checked this for the same range as which we have checked the second half of Conjecture 4.2. We observe that $B(7^2 83^2) = 64$, showing that the hypothesis on the factorization of n as pq^b is necessary.

5. When n is divisible by more than two primes

For products of three distinct primes, as noted in [Kaplan 2009, p. 2687], one of the products

$$\Phi_p(x)\Phi_q(x)\Phi_r(x)\Phi_{pqr}(x) \quad \text{or} \quad \Phi_1(x)\Phi_{pq}(x)\Phi_{pr}(x)\Phi_{qr}(x)$$

appears to give the largest height. Most of the time the first product gives the largest height. According to out data, of the 27492 n of the form pqr we have computed, the vast majority of the time the first product does give the maximal height while the second product only gives the maximal height only around half of the time (often they both give the maximal height). In general, one can make the following conjecture.

Conjecture 5.1. Let $n = p_1 \cdots p_t$ be square free. Then $B(n)$ is given by either

$$\prod_{\substack{d \mid n \\ \omega(d) \text{ even}}} \Phi_d(x) \quad \text{or} \quad \prod_{\substack{d \mid n \\ \omega(d) \text{ even}}} \Phi_d(x),$$

where $\omega(d)$ is the number of primes dividing d.

The conjecture is true when $t = 1$ and $t = 2$ [Pomerance and Ryan 2007, Lemma 2.1]. Our data supporting the conjecture for other n is listed in Table 1; in addition, the conjecture has been checked for $n = 2310$, the smallest product of five distinct primes.

For odd n, the analogue to Conjecture 4.4 would be: $B(pqr^b)$ is divisible by p. This statement is false for squarefree n, since $B(3 \cdot 31 \cdot 1009) = 599$, which is not divisible by 3. On the other hand, we can make the following conjecture.

Conjecture 5.2. Let $n = pqr^b$ where $p < q < r$, and $b > 1$. Then $B(n)$ is divisible by p. Moreover, $B(n) > p$.

Once more, our evidence for this is in Table 1. This conjecture is analogous to Conjectures 4.2 and 4.4.

6. Conclusions and future work

Above we have explicitly described several conjectures about the function $B(n)$. Implicitly, we have also suggested that proving explicit formulas for $B(n)$, especially by case-by-case analysis, is extremely difficult. In fact, even conjecturing formulas is difficult. A new method for proving formulas will be required before more progress can be made.

In addition to the obvious task of proving any of the conjectures included here and developing a new approach to proving these formulas, we propose the following related problems:

(1) Define the length of a polynomial $f = \sum_{n=0}^{d} a_n x^n$ to be $L(f) = \sum_{n=0}^{d} |a_n|$ and let

$$C(n) := \max\{L(f) : f \,|\, x^n - 1,\ f \in \mathbf{Z}[x]\}.$$

(2) Let $\mathbb{Q}(\zeta_n)$ be the n-th cyclotomic field and define the function

$$D(n) := \max\{H(f) : f \in \mathbb{Q}(\zeta_n)[x],\ f \,|\, x^n - 1 \text{ and } f \text{ monic}\}.$$

Can any explicit formulas or bounds be found for these functions? The database at [Ryan et al. 2010] has data related to the first of these two problems.

In [Decker and Moree 2010], a number of problems related to $B(n)$ have been described. The authors investigate, among other things, the set of coefficients of divisors of $x^n - 1$ and show that in some cases the coefficients of each divisor are a list of consecutive integers (sometimes excluding zero). In the future, we may return to the questions posed by Decker and Moree and investigate them computationally. This problem was suggested to us by Pieter Moree and the anonymous referee.

References

[Abbott 2009] J. Abbott, "Bounds on factors in $\mathbb{Z}[x]$", preprint, 2009. arXiv 0904.3057

[Decker and Moree 2010] A. Decker and P. Moree, "Coefficient convexity of divisors of $x^n - 1$", preprint, 2010. arXiv 1010.3938

[Garcia 2006] F. Garcia, entry A114536 in *The on-line encyclopedia of integer sequences*, edited by N. J. A. Sloane, 2006.

[Kaplan 2009] N. Kaplan, "Bounds for the maximal height of divisors of $x^n - 1$", *J. Number Theory* **129**:11 (2009), 2673–2688. MR 2010h:11161 Zbl 05603993

[Lam and Leung 1996] T. Y. Lam and K. H. Leung, "On the cyclotomic polynomial $\Phi_{pq}(X)$", *Amer. Math. Monthly* **103**:7 (1996), 562–564. MR 97h:11150 Zbl 0868.11016

[Pomerance and Ryan 2007] C. Pomerance and N. C. Ryan, "Maximal height of divisors of $x^n - 1$", *Illinois J. Math.* **51**:2 (2007), 597–604. MR 2008j:12012 Zbl 05197699

[Ryan et al. 2010] N. C. Ryan, B. C. Ward, and R. E. Ward, Database on cyclotomic polynomials, 2010, available at http://www.eg.bucknell.edu/~theburg/projects/data/wards/cyclo.py/index.

Received: 2010-09-29 Revised: 2010-11-23 Accepted: 2010-12-01

nathan.ryan@bucknell.edu *Department of Mathematics, Bucknell University, Lewisburg, PA 17837, United States*

bryan.ward@bucknell.edu *Department of Mathematics, Bucknell University, Lewisburg, PA 17837, United States*

ryan.ward@bucknell.edu *Department of Mathematics, Bucknell University, Lewisburg, PA 17837, United States*

INVOLVE 3:4 (2010)
https://doi.org/10.2140/involve.2010.3.459

Computing corresponding values of the Neumann and Dirichlet boundary values for incompressible Stokes flow

John Loustau and Bolanle Bob-Egbe

(Communicated by Kenneth S. Berenhaut)

We consider the Stokes equation for a flow through a partially obstructed channel and determine the relationship between Dirichlet boundary values (velocities) and Neumann boundary values (forces) for the FEM discrete form. For the steady state case we find a linear relationship. For the transient case the relationship depends on the time stepping procedure and includes the relationship at prior states. We resolve the issue for trapezoid and Adams–Bashford-2 time stepping. Since Stokes flow may be considered as the startup phase of Navier–Stokes flow, we give particular attention to a flow with a startup function.

1. Introduction

Our interest in boundary value questions for incompressible Stokes flow arises from the following setting. Commonly, finite element methods (FEM) are used to derive approximate solutions for the vector field of an incompressible fluid flow. These techniques involve first rendering a discrete form of the Navier–Stokes equation for the spatial variables via FEM and then employing finite difference techniques (FDM) to realize the flow in time. In this context the nonlinearity of the Navier–Stokes equations requires knowledge of the prior flow state at each time step. In practice the flow is assumed to begin at rest and then pass through a Stokes phase when the Reynolds number is small. The end step of this phase then provides the initial step data for the time step FDM applied to the Navier–Stokes equations. Authors often emphasize the importance of the Stokes phase to success in the resulting calculations with the Navier–Stokes phase [Gresho and Sani 2000].

When setting up the linear system of equations for a flow problem, the boundary values are initially applied in the Stokes phase then carried forward to the Navier–Stokes phase. For the case of a channel flow past an obstruction, authors commonly

MSC2000: 47N40.

Keywords: incompressible Stokes flow, finite element method, boundary values, computational fluid dynamics.

set values for the velocity field at the inflow edge, that is, they set Dirichlet boundary values. From a mathematical point of view, the flow problem could just as well be set up by assuming values for the force at that edge, that is, Neumann boundary values. This leads us to inquire how these two approaches differ if at all. Indeed, in [Gresho et al. 1981] the authors demonstrate the calculated flow vector field of an obstructed channel flow based on Neumann boundary values at the inflow edge. Interestingly, the authors state that the Neumann values are derived Dirichlet values. In particular they have postulated values for the velocity at the inflow, converted these velocities to forces and then proceeded with the Neumann boundary values.

In our investigation we determine a simple relationship between Neumann and Dirichlet boundary values for the steady state case. Carrying this forward we consider two common FDM techniques used for nonsteady or transient flows, *trapezoid* and *Adams-Bashford-2*. There are correspondences in the nonsteady case, but they are more complicated. In this case it is clear that an initial setting of forces or velocities result in very different outcomes. Indeed, by setting a startup function for force and then calculating the corresponding startup function for velocity results in a different startup velocity function at each applied node.

Although it is always mathematically possible to set Neumann boundary values for a node at the flow, this is not the case for Dirichlet boundary values. Indeed, the admissibility of Dirichlet boundary values lies in the physics not the mathematics. As our investigation is mathematical or linear algebraic, we decided to define a term to identify linear systems which admit Dirichlet boundary values.

In Section 2 we state the notation for the linear system arising from the Galerkin FEM applied to an incompressible Stokes flow. We also use this section to introduce an example. Later we use this example to demonstrate the results of Sections 3 and 4. In Section 3 we consider the steady state problem. Here we state results in a manner which is applicable to the nonsteady case. Finally the nonsteady case is handled in Section 4. Here we derive formulae relating Neumann boundary values to Dirichlet values and vice-versa. In both sections we provide point plots which demonstrate the formula for the example case.

We have included a note at the end to delineate the details of the example case.

2. Preliminaries

We begin by stating the governing equations for an incompressible Stokes flow. As we are primarily concerned with laminar flow we state the equations is two spatial dimensions.

Let $\vec{u} = \vec{u}(t, x, y) = (u(t, x, y), v(t, x, y))$ be a time dependent vector field in $\mathbb{R}^2$, and $P = P(t, x, y)$ be a real valued function. In addition $\vec{G} = (g_1, g_2)$ is a

time dependent vector field. In these equations $\vec{u} = (u, v)$ denotes the velocity field, P is the pressure and $\vec{G}$ represents external body forces such as gravity. Further suppose that $\vec{u}$ and P are sufficiently differentiable to support the following.

$$\frac{\partial \vec{u}}{\partial t} + \nabla P - \nu \nabla \cdot (\nabla \vec{u} + (\nabla \vec{u})^T) - \vec{G} = 0, \tag{2-1}$$

$$\nabla \cdot \vec{u} = 0, \tag{2-2}$$

Equation (2-1) is the *Stokes equation*. It is the Navier–Stokes equation with the inertial or convection term removed. It applies to viscous fluid flows with small Reynolds number. Equation (2-2) is referred to as the *continuity equation*. It arises from the incompressibility assumption. Alternatively (2-1) and (2-2) may be referred to as the *Stokes equations* governing an incompressible flow at low Reynolds number. There are equivalent formulations for these equations [see 1] that are derived from the given pair. Additionally, a fourth-order equation may be derived from these. This equation states a relationship for velocity without reference to pressure. The formulation given here is convenient for our purposes.

If Ω is the domain of the flow, then Ω is a connected compact set in $\mathbb{R}^2$. Take $[0, T]$ as the time interval. When t is fixed, then P lies in $L^2[\Omega] = L^2$ and both u and v are elements of $H^1 = \{u : \Omega \to \mathbb{R} : \int_\Omega \nabla u \cdot \nabla u < \infty\}$. Finally $\vec{G}$ represents external body forces such as gravity. Equation (2-1) restated in terms of the coordinate functions yields

$$\frac{\partial u}{\partial t} + \frac{\partial P}{\partial x} - \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y \partial x}\right) - g_1 = 0, \tag{2-3}$$

$$\frac{\partial v}{\partial t} + \frac{\partial P}{\partial x} - \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 v}{\partial y^2}\right) - g_2 = 0. \tag{2-4}$$

Below we suppose that the external body forces do not play a significant role in the flow and will ignore this term.

A discrete form of the Stokes equation is derived from FEM techniques applied to the spatial variables. If the flow is transient ($\partial \vec{u}/\partial t \neq 0$) then the resulting discrete equations yield approximate solutions via finite difference techniques.

For the purposes of the theory and examples developed below, we base the FEM on the (Q_1^4, Q_0^1) model. This model supposes the decomposition of the flow domain into the union of rectangles. The vertices of the rectangles are velocity nodes and centroids of the rectangles are the pressure nodes. For the succeeding examples we use a channel flow obstructed by a square obstruction (see Figure 1).

Denoting the partition by $\Omega = \bigcup_{e=1}^{s} \Omega^e$, we define finite dimensional subspaces V of H^1 and W of L^2. V is defined as the linear space of first-order polynomials $\{\phi_i^e : i = 1, 2, 3, 4; e = 1, \ldots, s\}$, where each ϕ_i^e is supported by Ω^e and equal to the i-th Lagrange polynomial on Ω^e. In turn W is the span of constant functions,

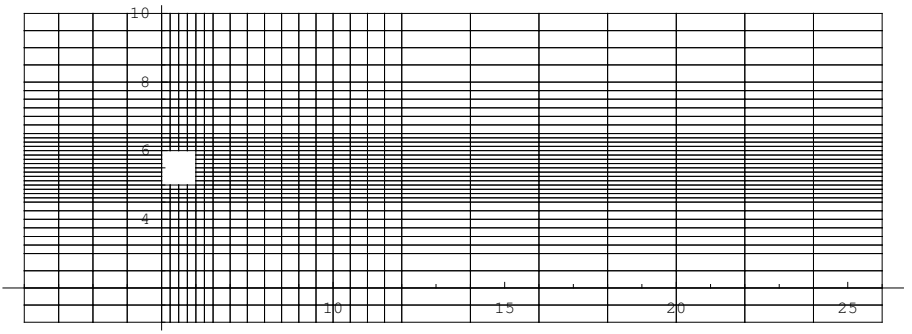


Figure 1. Decomposition of an obstructed channel into rectangular elements.

$\{P^e : e = 1, \ldots, m\}$ supported by the elements. The Galerkin FEM proceeds by seeking elements $\tilde{u}$ and $\tilde{v}$ in V and $\tilde{P}$ in W so that the residual (Equations (2-1) and (2-2) and evaluated at these functions) is L^2 orthogonal to V. In particular,

$$(R_1(\tilde{u}, \tilde{v}, \tilde{P}), \phi_i^e) = \int_\Omega R_1(\tilde{u}, \tilde{v}, \tilde{P})\phi_i^e = 0, \tag{2-5}$$

$$(R_2(\tilde{u}, \tilde{v}), \phi_i^e) = \Omega \int_\Omega R_2(\tilde{u}, \tilde{v}) P^e = 0. \tag{2-6}$$

Expanding these equations and using the divergence theorem to linearize the second-order term, we arrive at the following linear system for each element, e:

$$\begin{pmatrix} M_1^e & 0 & 0 \\ 0 & M_2^e & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \\ \dot{u}_4 \\ \dot{v}_1 \\ \dot{v}_2 \\ \dot{v}_3 \\ \dot{v}_4 \\ \dot{P} \end{pmatrix} + \begin{pmatrix} K_1^e & K_{12}^2 & L_1^e \\ K_{21}^e & K_2^e & L_2^e \\ (L_1^e)^T & (L_2^e)^T & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ P \end{pmatrix} = \begin{pmatrix} f_{11} \\ f_{12} \\ f_{13} \\ f_{14} \\ f_{21} \\ f_{22} \\ f_{23} \\ f_{24} \\ g \end{pmatrix},$$

where the dot represents differentiation with respect to t. The matrix entries are

$$M_1^e(i, j) = M_2^e(i, j) = \int_{\Omega^e} \phi_i^e \phi_j^e, \qquad K_{12}^e(i, j) = K_{21}^e(j, i) = \nu \int_{\Omega^2} \frac{\partial \phi_i^e}{\partial y} \frac{\partial \phi_j^e}{\partial x},$$

$$K_1^e(i, j) = \nu \int_{\Omega^e} 2\frac{\partial \phi_i^e}{\partial x} \frac{\partial \phi_j^e}{\partial x} + \frac{\partial \phi_i^e}{\partial y} \frac{\partial \phi_j^e}{\partial y}, \qquad L_1^e(i, 1) = -\int_{\Omega^e} \frac{\partial \phi_i^e}{\partial x},$$

$$K_2^e(i, j) = \nu \int_{\Omega^e} \frac{\partial \phi_i^e}{\partial x} \frac{\partial \phi_j^e}{\partial x} + 2\frac{\partial \phi_i^e}{\partial y} \frac{\partial \phi_j^e}{\partial y}, \qquad L_2^e(i, 1) = -\int_{\Omega^e} \frac{\partial \phi_i^e}{\partial y}.$$

On the right hand side we have

$$f_{1i}^e = \int_{\Gamma^e} \left(2\frac{\partial \tilde{u}}{\partial x}, \frac{\partial \tilde{u}}{\partial y} + \frac{\partial \tilde{v}}{\partial x}\right)\phi_i^e \cdot \vec{n}, \quad f_{2i}^e = \int_{\Gamma^e} \left(\frac{\partial \tilde{v}}{\partial x} + \frac{\partial \tilde{u}}{\partial y}, 2\frac{\partial \tilde{v}}{\partial y}\right)\phi_i^e \cdot \vec{n},$$

from the application of the divergence theorem to (2-5). Whereas (2-6) yields

$$g = \int_{\Gamma^e} (\tilde{u}, \tilde{v}) \cdot \vec{n}.$$

In both cases Γ^e denotes the boundary of Ω^e. Using standard processes [Huebner et al. 2001] we assemble these s linear systems in a single $(2m+s) \times (2m+s)$ system (where m is the number of nodes and s. is the number of elements). This is done by first identifying the corresponding node for each vertex of a single element. Then adding the linear equations which refer to a common node. The resulting system can be expressed in compact form as

$$\begin{pmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{P} \end{pmatrix} + \begin{pmatrix} K & L \\ L^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ P \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

Here M is $m \times m$ symmetric, K is $2m \times 2m$ symmetric and positive definite (from the underlying physics) and L is $2m \times s$.

3. Boundary values for the steady state problem

We begin our study of boundary values by considering the steady state problem. In this case we need only consider the equation

$$\begin{pmatrix} K & L \\ L^T & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ P \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{3-1}$$

as the discrete form of the steady state Stokes equation. We assert that the coefficient matrix of (3.1) is nonsingular. This assertion is equivalent to the statement that the flow has a unique solution in the discrete form stated in (3.1). In general this is not the case, but it may be achieved by imposition of boundary conditions at the channel edges and at the obstruction, as well as by the choice of model. The underlying physics assures us that the matrix K is symmetric and positive definite, as well as sparse and diagonally dominant. L is sparse.

Our primary concern is with the entries of f for the nodes along the inflow boundary. On the one hand we may designate a value for f_i. In this case the designated value implements driving forces applied along the inflow edge as is evident from the expression for f_i given in the previous section.These boundary values are then called *Neumann* or *natural*. Alternatively we may designate the

velocity components on this boundary. This alternative is implemented at the i-th node by replacing the i-th row of the coefficient matrix to the i-th row of the identity matrix, denoted $\vec{e}_i$ and then setting f_i to the desired velocity. These boundary values are referred to as *Dirichlet* or *essential*.

For simplicity of notation we write Equation (3-1) as $A\vec{u}=\vec{f}$. Now since A is nonsingular, then for any choice of $\vec{f}$ the system has a unique solution. Consider the process, just described, used to set Dirichlet boundary values. For this to be meaningful, the resulting coefficient matrix must be row-equivalent to A. Otherwise the resulting linear system, $B\vec{u}=\hat{f}$ would no longer represent the discrete form of the same differential equation. With this in mind we begin our analysis with the following definition, where for a matrix A, $A_{(i)}$ denotes the i-th row of A.

Definition 3.1. Let $A\vec{u} = \vec{f}$ be an n-by-n linear system of equations and take i, $1 \leq i \leq n$. Then we say that a Dirichlet boundary condition at f_i is *algebraically admissible* provided A is row-equivalent to B where the $A_{(j)} = B_{(j)}$ for each $j \neq i$ and $B_{(i)} = \vec{e}_i^T$.

From the definition it is apparent that a Dirichlet condition at f_i is algebraically admissible if there are elementary row matrices $E_0, E_1, \ldots, E_m$ with

$$B = \Big(\prod_{j\neq i} E_j\Big) E_0 A,$$

where E_0 is type 2, representing the multiplication of row i of A by a nonzero scalar, and for $j \neq 0$, E_j is type-3, representing the operation of adding to the i-th row a scalar multiple of some other row.

For the case at hand, a linear system arising from the FEM discrete form of the steady state Stokes equation, the matrix K is positive definite symmetric, sparse and diagonally dominant. Therefore for each $i \leq 2m$,

$$\vec{e}_i^T = \sum_j \alpha_j K_{(j)}.$$

Since K is diagonally dominant, α_i is not zero. If $E_{\beta s+t}$ denotes the row operation of adding β times the s-th row to the t-th row and $E_{\beta s}$ denotes the elementary operation of multiplying the s-th row by nonzero β, then

$$\vec{e}_i^T = \Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big) E_{\alpha_i i} A.$$

Therefore, in this case a Dirichlet condition at f_i is algebraically admissible for each $i \leq 2m$.

Theorem 3.1. *Let $A\vec{u} = \vec{f}$ be an n-by-n linear system of equations and take i, with $1 \leq i \leq n$. Suppose that a Dirichlet boundary condition at f_i is algebraically admissible. Then there exists a nonsingular linear transformation N such that*

the Dirichlet assumption for u_i is the i-th coordinate of $N\vec{f}$. Further all other coordinates of $N\vec{f}$ are unchanged.

Proof. From the comment following Definition 3.1, it suffices to set

$$N = \Big(\prod_{j\neq i} E_j\Big)E_0.$$

Now the remaining assertions are immediate. □

Next, supposing that A is nonsingular, we can get a specific representation for the elementary row operations. First we set up the notation. Set $E_0 = E_{\alpha_i i}$ and $E_j = E_{\alpha_k k+i}$. Now we may suppose that N has n factors by setting $E_j = E_{\alpha_j j+i}$ for each $j \neq i$ where $\alpha_j = 0$ if row j is not involved in reducing the i-th row of A. Finally define the column n-tuple $\vec{\alpha} = \alpha_{(j)}$.

Corollary 3.2. *If A is nonsingular, then $\vec{\alpha} = (A^T)^{-1}\vec{e}_i$, is the i-th column of $(A^T)^{-1}$.*

Proof. With the notation just introduced,

$$(E_{\alpha_i i}A)_{(i)} = \alpha_i A_{(i)} \quad \text{and} \quad (E_{\alpha_j j+i}A)_{(i)} = \alpha_j A_{(j)} + A_{(i)}.$$

Therefore,

$$(NA)_{(i)} = \sum_{j\neq i} \alpha_j A_{(j)} + \alpha_i A_{(i)} = \sum_j \alpha_j A_{(j)}.$$

Restating this as an expression for $(\vec{e}_i)^T$ we get

$$(\vec{e}_i)^T = \sum_j \alpha_j A_{(j)} = \Big(\sum_j (A^T)^{(j)}\alpha_j\Big)^T = (A^T\vec{\alpha})^T. \qquad \square$$

This yields the desired expression for $\vec{\alpha} = (A^T)^{-1}\vec{e}_i$, which is indeed the i-th column of $(A^T)^{-1}$.

In the case of (3-1), A is symmetric and we have:

Corollary 3.3. *If A is the coefficient matrix for the FEM discrete form of the Stokes equation, then $\vec{\alpha}$ is the i-th column of A^{-1}.*

Next we turn to the relationship between the Neumann boundary value $\vec{f}$ and the Dirichlet boundary value u_i at the i-th node.

Corollary 3.4. *Suppose that A is nonsingular and algebraically admits a Dirichlet condition at f_i then the Dirichlet value, u_i is related to $\vec{f}$ via $u_i = \vec{\alpha}\cdot\vec{f}$ where $\vec{\alpha} = (A^T)^{-1}\vec{e}_i$. (Here $\vec{\alpha}\cdot\vec{f}$ denotes the ordinary inner product in $\mathbb{R}^n$.)*

Proof. The relation

$$A\vec{u} = \vec{f}$$

yields,

$$u_i = \vec{e}_i\cdot\vec{u} = \vec{e}_i^T\vec{u} = (A^T\vec{\alpha})^T\vec{u} = (\vec{\alpha})^T A\vec{u} = (\vec{\alpha})^T\vec{f} = \vec{\alpha}\cdot\vec{f}. \qquad \square$$

We end this section by considering the following problem. Given a linear system with Dirichlet conditions applied, what is the corresponding linear system without Dirichlet boundary conditions, but rather Neumann boundary conditions.

Theorem 3.5. *Let A be an n-by-n matrix, which algebraically admits a Dirichlet boundary condition on the i-th row. Suppose that B is row-equivalent to A via the nonsingular matrix N as in Theorem 3.1 Consider a linear system $B\vec{u} = \hat{f}$, then the equivalent linear system $A\vec{u} = \vec{f}$ satisfies $f_j = \hat{f}_j$ for each $j \neq i$ and $f_i = \vec{\beta} \cdot \hat{f}$, where $\vec{\beta} = (B^T)^{-1}(A_{(i)})^T$.*

Proof. With the notation of Equation (3-1), Take N nonsingular so that $B = NA$ and $\hat{f} = N\vec{f}$. Now N is a product of elementary row matrices. Hence the same is true of N^{-1}. In particular,

$$N^{-1} = \Big(\Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big)E_{\alpha_i i}\Big)^{-1} = (E_{\alpha_i i})^{-1}\Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big)^{-1} = \Big(\prod_{j\neq i} E_{\beta_j j+i}\Big)E_{\beta_i i},$$

where $\beta_i = \alpha_i^{-1}$ and $\beta_j = -\alpha_j/\alpha_i$ otherwise. Setting $\vec{\beta} = (\beta_i)$, it now follows that $\vec{\beta} \cdot \hat{f} = f_i$. In turn

$$A_{(i)} = (N^{-1}B)_{(i)} = \sum_j \beta_j B_{(j)} = \sum_j (B^T)^{(j)}\beta_j = (B^T\vec{\beta})^T. \qquad \square$$

So $(A_{(i)})^T = B^T\vec{\beta}$ or $(B^T)^{-1}(A_{(i)})^T = \vec{\beta}$. The following point plots show first a set of given forces at points along the inflow edge of the example flow (Figure 2, left). We used B-splines to fit a continuous function to the given data. With this function we were able to infer forces at the inflow nodes and compute f via one point quadrature. Then we used Corollary 3.4 to compute the velocities shown in the second plot (Figure 2, right). As indicated by the mathematics, the calculated flow using either the Neumann or the Dirichlet boundary values produces identical velocity fields.

4. Boundary values for the transient flow

In this section we modify our results of the section to the case of a nonsteady Stokes flow. Our particular concern with Stokes flows is their application as the initial phase of a Navier–Stokes flow. In this setting it is natural to suppose that there is a velocity or force startup function, $v(t)$ with $t\epsilon(0, T]$, implemented at the inflow edge. In this section we will consider the discrete case of the nonsteady Stokes equation and determine the relationship between a velocity startup and a force startup.

For the nonsteady Stokes flow the spatial problem is realized via finite element techniques while the time dependent problem is developed via finite difference techniques. There are several competing finite difference techniques. For each, the

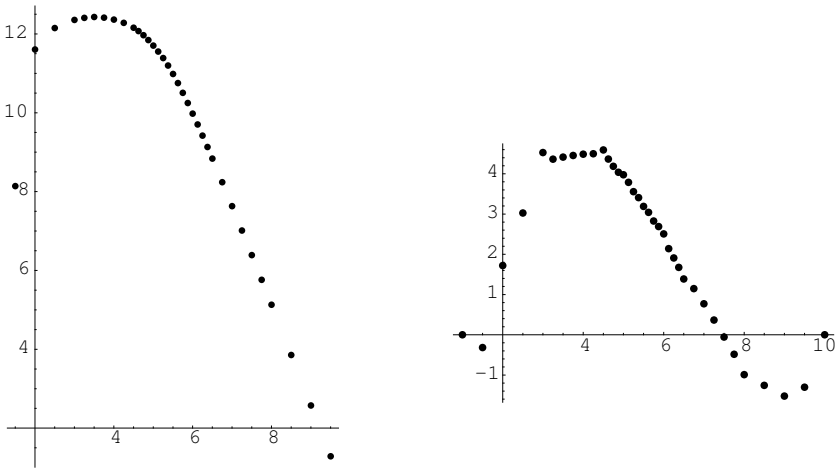


Figure 2. Left: derived forces on inflow edge. Right: computed velocities at the inflow edge.

function relating forces to velocities and pressures is distinct. We will develop two cases, the *trapezoid* (TR) method and the *Adam-Bashford-2* (AB-2) method. We begin with TR. Here we use superscripts to designate time steps.

$$\begin{pmatrix} M+\frac{1}{2}\Delta t\,K & \Delta t\,L \\ \frac{1}{2}\Delta t\,L^T & 0 \end{pmatrix}\begin{pmatrix} u^n \\ v^n \\ P^n \end{pmatrix} = \begin{pmatrix} M-\frac{1}{2}\Delta t\,K & -\Delta t\,L \\ -\frac{1}{2}\Delta t\,L^T & 0 \end{pmatrix}\begin{pmatrix} u^{n-1} \\ v^{n-1} \\ P^{n-1} \end{pmatrix} + \begin{pmatrix} \Delta t f^n \\ tg^{n-1} - \frac{1}{2}\Delta t\,L^T \begin{pmatrix} u^{n-1} \\ v^{n-1} \end{pmatrix} \end{pmatrix}, \quad (4\text{-}1)$$

where

$$g^n = L^T \begin{pmatrix} u^n \\ v^n \end{pmatrix}.$$

The term on the right, f, which is related to force is superscripted as we may suppose it varies with time. Further we suppose that the fluid starts at rest, so for $t = t_1$, $u^0 = v^0 = P^0 = 0$. Hence (4-1) becomes

$$\begin{pmatrix} M+\frac{1}{2}\Delta t\,K & \Delta t\,L \\ \frac{1}{2}\Delta t\,L^T & 0 \end{pmatrix}\begin{pmatrix} u^1 \\ v^1 \\ P^1 \end{pmatrix} = \begin{pmatrix} \Delta t f^n \\ 0 \end{pmatrix}. \quad (4\text{-}2)$$

As in Section 3, we use a notationally simplified version of these equations:

$$A\vec{u}^n = C\vec{u}^{n-1} + \Delta t \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix}. \quad (4\text{-}3)$$

For N nonsingular, we have

$$NA\vec{u}^n = NC\vec{u}^{n-1} + N\Delta t \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix}. \tag{4-4}$$

As in the steady state case, restriction of these two equations to the example flow assures us that A is nonsingular and that the Dirichlet boundary condition is algebraically admissible at each inflow edge node. Assuming that the fluid starts at rest implies that Equation (4-3) reduces to

$$\frac{1}{\Delta t} A\vec{u}^1 = \begin{pmatrix} f^1 \\ 0 \end{pmatrix} \quad \text{at } n = 1.$$

This equation is essentially the same as the one considered in Section 3 except that the coefficient matrix is not symmetric. Nevertheless, Corollary 3.4 and Theorem 3.5 apply to the present setting.

Theorem 4.1. *Consider the TR time step development of the nonsteady Stokes flow represented by (4-3). Suppose that A is nonsingular and that Neumann boundary values are set at $t = t_n$ via the coordinates of f^n. Then a boundary value f_i^n may be replaced by a Dirichlet boundary value*

$$u_i^n = (NC)_{(i)}\vec{u}^{n-1} + \vec{\alpha} \cdot \begin{pmatrix} f^n \\ 0 \end{pmatrix},$$

where $\vec{\alpha} = \Delta t (A^T)^{-1}\vec{e}_i$. Hence, $\vec{\alpha}$ is Δt times the i-th column of $(A^T)^{-1}$. In addition

$$N = \left(\prod_{j \neq i} E_{\alpha_j j+i} \right) E_{\alpha_i i}.$$

Proof. The assertion for $t = t_1$ follows immediately from Corollary 3.4. For $n > 1$, we need to first let

$$C\vec{u}^{n-1} + \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix}$$

take the role of $\vec{f}$ in Corollary 3.4 to get

$$\begin{aligned} u_i^n &= (NC)_{(i)}\vec{u}^{n-1} + N \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix} = NC_{(i)}\vec{u}^{n-1} + \vec{\alpha} \cdot \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix} \\ &= (NC)_{(i)}\vec{u}^{n-1} + \vec{\alpha} \cdot \begin{pmatrix} f^n \\ 0 \end{pmatrix}. \end{aligned} \tag{4-5}$$

The final equality holds since if the upper left hand block of A is k-by-k then $\alpha_j = 0$ for $j > k$. Indeed, the upper left block is itself nonsingular, so by Dirichlet admissibility, the i-th row is row-equivalent to the i-th row of the k-by-k identity

matrix. Finally, since the lower right block of A is zero, it now follows that

$$\vec{\alpha} \cdot \begin{pmatrix} f^n \\ g^{n-1} \end{pmatrix} = \vec{\alpha} \cdot \begin{pmatrix} f^n \\ 0 \end{pmatrix}.$$

The remaining assertions follow as in Section 3. □

Notice that for $n > 1$, the calculation of the Dirichlet boundary value requires the prior state. Therefore the results for the steady state problem do not carry over directly to the nonsteady flow. In particular even if f_i^n is fixed for each $n > 1$, u_i^n will vary with n.

We now particularize Theorem 4.1 to the case of a startup function for the force along the inflow edge. For this purpose we need to develop some notation. First (4-3) becomes

$$A\vec{u}^n = C\vec{u}^{n-1} + \Delta t \begin{pmatrix} \varphi(t_n) f \\ g^{n-1} \end{pmatrix}, \tag{4-6}$$

where $\varphi : (0, T] \to (0, 1]$ designates the startup function and $f = (f_i)$ with $f_i = 0$ for each node which is not on the inflow edge and $f_i = 1$ at the inflow edge. In turn (4-5) becomes

$$\begin{aligned} u_i^n &= (N_i C)_{(i)} \vec{u}^{n-1} + \Delta t \varphi(t_n) \overrightarrow{\alpha} \cdot f \\ &= (N_i C)_{(i)} \vec{u}^{n-1} + \Delta t \varphi(t_n) \left((A^T)^{-1}\right)_{(i)} f, \end{aligned} \tag{4-7}$$

where N is now subscripted to identify the row operations applied to the i-th row of A. Next we define $u = u_i$ and $u_i = \left((A^T)^{-1}\right)_{(i)} f$ as in Section 3. We can consider a corresponding startup function for Dirichlet boundary values at the inflow edge.

Corollary 4.2. *Suppose that $\varphi : (0, T] \to (0, 1]$ denotes a startup function for the force along the inflow edge of the transient Stokes flow. Let v be a second function defined on the time steps and taking values in $\mathbb{R}^\ell$, where ℓ designates the number of nodes on the inflow edge. If v is defined by*

$$v(t_n)_i = \frac{1}{u_i} (N_i C)_{(i)} \vec{u}^{n-1} + \Delta t \varphi(t_n),$$

then $v(t_n)_i u_i = u_i^n$.

Proof. The result follows immediately from (4-7). □

The startup force function results in separate velocity startup functions, one defined at each of designated nodes. We illustrate this result in the following plots. Figure 3 shows a burst startup function $\varphi(t) = 1 - e^{-t/0.1}$. The subsequent three point plots (Figure 4) show the corresponding velocity plots at selected nodes: 2, 6 and 12 along the inflow edge (see Figure 1). Note that the velocity startup functions though distinct are very similar. Indeed they appear linear. Lastly, Figure 5 shows the final value ($t = 0.5$) for the velocity startup function at each node. This plot is

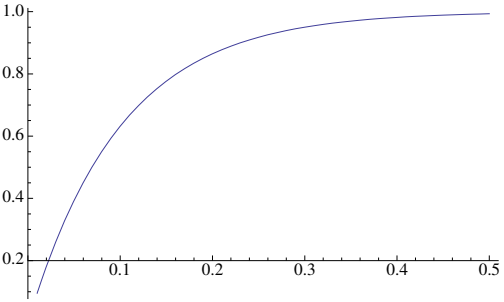


Figure 3. Burst startup function.

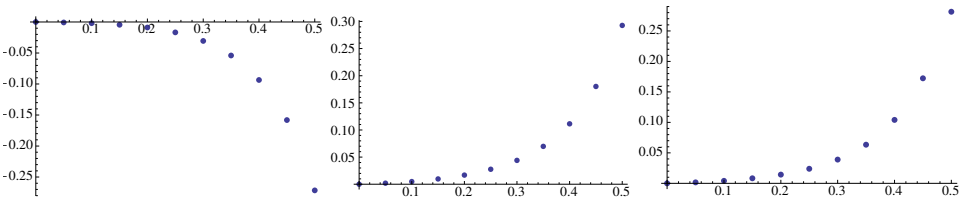


Figure 4. TR: Dirichlet values at nodes 2, 6, 12 (left to right).

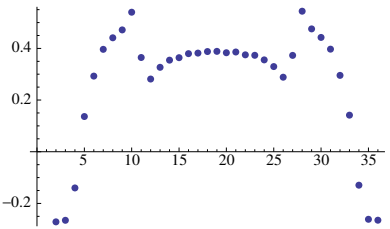


Figure 5. TR: inflow velocities at $t = 0.5$.

symmetric. The differences from node to node appear consistent with incompressibility as the flow reacts to the obstruction.

The next result considers the reverse setting where we begin with a Dirichlet boundary value and derive the corresponding Neumann value.

Theorem 4.3. *Consider the TR time step development of the nonsteady Stokes flow represented by (4-3). Suppose that A is nonsingular and algebraically admits Dirichlet boundary values along the inflow edge. Suppose for $t = t_n$ that Dirichlet boundary values are set on this edge via the coordinates of u^n to yield*

$$DA\vec{u}^n = \begin{pmatrix} u^n \\ \Delta t L P^{n-1} + g^{n-1} \end{pmatrix},$$

where D is a product of matrices N as described in Section 2. Fix a node i, then the corresponding Neumann boundary value at the i-th node is

$$f_i = \vec{\beta} \cdot \hat{f}, \quad \text{where } \vec{\beta} = (B^T)^{-1}(A_{(i)})^T \text{ and } B = NA.$$

Proof. We proceed as in Theorem 3.5. First we set

$$N = \Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big) E_{\alpha_i i},$$

as in Theorem 3.1 and

$$N^{-1} = \Big(\Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big) E_{\alpha_i i}\Big)^{-1} = (E_{\alpha_i i})^{-1}\Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big)^{-1} = \Big(\prod_{j\neq i} E_{\beta_j j+i}\Big) E_{\beta_i i},$$

where $\beta_i = \alpha_i^{-1}$ and $\beta_j = -\alpha_j/\alpha_i$ otherwise. Setting $\vec{\beta} = (\beta_i)$, it now follows that $\vec{\beta} \cdot u^n = f_i$ and

$$A_{(i)} = (N^{-1}B)_{(i)} = \sum_j \beta_j B_{(j)} = \sum_j (B^T)^{(j)} \beta_j = (B^T \vec{\beta})^T. \qquad \square$$

We turn next to the case of AB-2. This FDM procedure computes the current velocity field in terms of the weighted average of the prior two time steps via

$$\begin{pmatrix} M & \frac{3}{2}\Delta t\, L \\ L^T & 0 \end{pmatrix} \begin{pmatrix} u^n \\ v^n \\ P^{n-1} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{3}{2}\Delta t\, K + M & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u^{n-1} \\ v^{n-1} \\ P^{n-1} \end{pmatrix} + \begin{pmatrix} \frac{1}{2}\Delta t\, K & \frac{1}{2}\Delta t\, L \\ 0 & 0 \end{pmatrix} \begin{pmatrix} u^{n-2} \\ v^{n-2} \\ P^{n-2} \end{pmatrix} \begin{pmatrix} \Delta t f^n \\ g^{n-1} \end{pmatrix}, \quad (4\text{-}8)$$

where

$$g^n = L^T \begin{pmatrix} u^n \\ v^n \end{pmatrix} = P^n.$$

For a fluid starting at rest we have for $n = 1$

$$\begin{pmatrix} M & \frac{3}{2}\Delta t\, L \\ L^T & 0 \end{pmatrix} \begin{pmatrix} u^1 \\ v^1 \\ P^1 \end{pmatrix} = \begin{pmatrix} \Delta t f^1 \\ 0 \end{pmatrix}. \quad (4\text{-}9)$$

As before it is convenient to restate (4-8) in a simplified form:

$$B\vec{u}^n = D\vec{u}^{n-1} + E\vec{u}^{n-2} + \begin{pmatrix} \Delta t f^n \\ g^{n-1} \end{pmatrix}. \quad (4\text{-}10)$$

For N nonsingular,

$$NB\vec{u}^n = ND\vec{u}^{n-1} + NE\vec{u}^{n-2} + N\begin{pmatrix} \Delta t f^n \\ g^{n-1} \end{pmatrix}. \quad (4\text{-}11)$$

The next results are analogous to Theorem 4.1, Corollary 4.2 and Theorem 4.3.

Theorem 4.4. *Consider the AB-2 time step development of the nonsteady Stokes flow given by (4-8), (4-10). Suppose that Neumann boundary values are set at time step $t = t_n$ via the coordinates of f^n. Then the boundary value f_i^n may be replaced by a Dirichlet boundary value*

$$u_i^n = (ND)_{(i)}\vec{u}^{n-1} + (NE)_{(i)}\vec{u}^{n-2} + \vec{\alpha}\cdot\begin{pmatrix} f^n \\ 0 \end{pmatrix},$$

where

$$\vec{\alpha} = \Delta t (B^T)^{-1}\vec{e}_i \quad \textit{and} \quad N = \Big(\prod_{j\neq i} E_{\alpha_j j+i}\Big) E_{\alpha_i i}.$$

Hence, $\vec{\alpha}$ is Δt times the i-th column of $(B^T)^{-1}$.

Proof. The expression for u_i^n follows from (4-11) and the given decomposition of N as a product of elementary matrices. The given expression results from

$$\sum_i \alpha_i B_{(i)} = (NB)_{(i)} = \vec{e}_i^T.$$

The final statement is immediate. □

Turning to a startup function for force we have:

Corollary 4.5. *Suppose the setting of Theorem 4.4 and suppose that*

$$\gamma(0, T] :\to (0, 1]$$

denotes a startup function for the force along the inflow edge of the transient Stokes flow. Determine a second function, δ, defined at the time steps and taking values in $\mathbb{R}^\ell$, where ℓ designates the number of nodes in the inflow edge by

$$\delta(t_n)_i = \frac{1}{u_i}(N_i D)_{(i)}\vec{u}^{n-1} + \frac{1}{u_i}(N_i E)_{(i)}\vec{u}^{n-2} + \Delta t\gamma(t_n).$$

Then $\delta(t_n)_i u_i = u_i^n$, where $u_i = ((B^T)^{-1})_{(i)} f$.

Proof. As in the TR case we now particularize (4-11) for the startup function, γ, the product of elementary operations associated to the i-th inflow edge node, N_i, and then consider the i-th entry of the result to get

$$\vec{u}_i^n = (N_i D)_{(i)}\vec{u}^{n-1} + (N_i E)_{(i)}\vec{u}^{n-2} + \Delta t\gamma(t_n)((B^T)^{-1})_{(i)} f.$$

The result is now immediate. □

Finally we consider the reverse case.

Theorem 4.6. *Consider the AB-2 time step development of the nonsteady Stokes flow represented by (4-10). Suppose that for $t = t_n$ Dirichlet boundary values are set along the inflow edge via the corresponding coordinates of u_n to yield*

$$FB\vec{u}_n = \begin{pmatrix} u^n \\ g^{n-1} \end{pmatrix},$$

where F is a product of matrices N as described in Theorem 4.4. Fix a node i, then the corresponding Neumann boundary value at the i-th node is $f_i = \vec{\beta} \cdot \hat{f}$, where $\vec{\beta} = (G^T)^{-1}(B_{(i)})^T$ and $G = NB$.

Proof. We proceed as with Theorem 4.3. First we set $N = \left(\prod_{j \neq i} E_{\alpha_j j+i}\right) E_{\alpha_i i}$, as in Theorem 3.1, then resolve

$$N^{-1} = \left(\prod_{j \neq i} E_{\beta_j j+i}\right) E_{\beta_i i},$$

where $\beta_i = \alpha_i^{-1}$ and $\beta_j = -\alpha_j/\alpha_i$ otherwise. Setting $\vec{\beta} = (\beta_i)$, we have $f_i = \vec{\beta} \cdot \hat{f}$. Finally,

$$B_{(i)} = (N^{-1}G)_{(i)} = \sum_j \beta_j G_{(j)} = (G^T \vec{\beta})^T,$$

which yields the desired expression for $\vec{\beta}$. □

The plots in Figures 6 and 7 show output for AB-2. They are analogous to Figures 4 and 5.

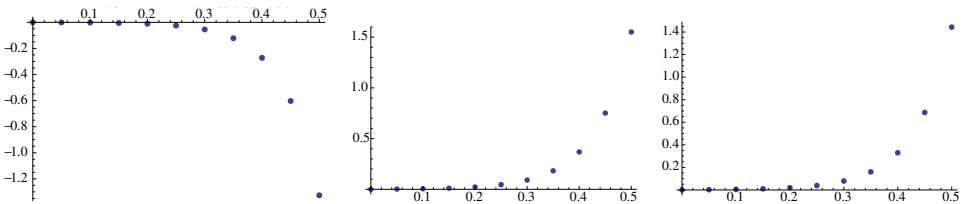


Figure 6. AB-2: Dirichlet values at nodes 2, 6 and 12 (left to right).

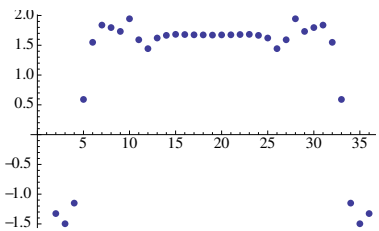


Figure 7. AB-2: inflow velocities at $t = 0.5$.

A note on the illustrations

All programming was done in Mathematica.

Geometry. Channel: lower left vertex at (1, 1); upper right vertex at (26, 10). Obstruction: lower left vertex at (5, 5); upper right vertex at (6, 6).

FEM. 976 elements: 1052 velocity nodes; 976 pressure nodes; 2104+976 degrees of freedom; max x increment = 1.0; min x increment = 0.25; max y increment = 0.5; min y increment = 0.125.

FDM. $\Delta t = 0.05$.

Fluid. water, $\nu = 0.89$.

Boundary values. All surfaces are nonslip and nonpenetrating. Dirichlet boundary values are set to zero. The outflow edge is included in the Neumann boundary with values set to zero. Transient flows are started at rest.

References

[Gresho and Sani 2000] P. Gresho and R. L. Sani, *Incompressible flow and the finite element method* (2 vol.), Wiley, Chichester, UK, 2000. Zbl 0988.76005

[Gresho et al. 1981] P. M. Gresho, R. L. Lee, and C. D. Upson, "FEM solution of the Navier–Stokes equations for vortex shedding behind a cylinder: experiments with the four-node element", *Adv. Water Resources* **4** (1981), 175–184.

[Huebner et al. 2001] K. H. Huebner, D. L. Dewhirst, D. E. Smith, and T. G. Byrom, *The finite element method for engineers*, Wiley, New York, 2001. Zbl 0575.73087

Received: 2007-08-10 Accepted: 2010-10-23

jloustau@msn.com *Department of Mathematics, Hunter College (CUNY), New York, NY 11374, United States*

bbobegbe@gmail.com *Hunter College (CUNY), New York, NY 11374, United States*

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@mathscipub.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve
2010 vol. 3 no. 4